INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Data Science Applied to Public Fund Management

Tiago Afonso Frade Martins

Master in Data Science

Supervisor:
PhD Ana Maria de Almeida, Associate Professor,
Iscte - University Institute of Lisbon

PhD Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor,
Iscte - University Institute of Lisbon

September, 2023

BUSINESS SCHOOL

TECHNOLOGY AND ARCHITECTURE

Department of Quantitative Methods for Management and Economics
Department of Information Science and Technology

Data Science Applied to Public Fund Management

Tiago Afonso Frade Martins

Master in Data Science

Supervisor:
PhD Ana Maria de Almeida, Associate Professor,
Iscte - University Institute of Lisbon

PhD Elsa Alexandra Cabral da Rocha Cardoso, Assistant Professor,
Iscte - University Institute of Lisbon

September, 2023

Acknowledgment

I would like to thank my Supervisors, Professor Ana de Almeida and Professor Elsa Cardoso, for all their support for this work. The technical and knowledge expertise of Professor Ana helped me understand the domain of XAI in such a way that, without it, I would find it much harder to know. For Professor Elsa, the help regarding the organization, preparation, and writing of my thesis is invaluable, and the quality of this work, but also my discipline, is positively impacted by these Professors. I would also like to thank Professor Luís Nunes who brought invaluable business knowledge and helped steer me in the right direction by suggesting improvements to the work done here. I am extremely grateful for the help Professor Ricardo Paes Mamede gave in the extraction of interpretations for obtained results, and by helping me find the most suited literature for the problem at hand.

Thank you to my all friends for being there when the times got difficult. Thank you to João Paulo as we are in the same journey, and we managed to pull through. Thank you to Tomás for making me understand things in a new perspective. Lastly, thank you to my parents, for giving me the motivation I needed not only to keep working on this dissertation but to keep improving it.

Resumo

O crescimento na utilização de modelos de *Machine Learning* (ML) tem vindo a acentuar-se ao longo da última década, com a introdução prática de modelos de alto desempenho como redes neuronais profundas, apenas conseguido com a melhoria dos recursos computacionais para a sua utilização. Este tipo de modelos, apesar de obterem um alto desempenho, costumam por norma ser de difícil interpretação, ganhando o termo "caixa-negra", ou *black-box*, por não se saber ao certo como o modelo funciona e opera. Para endereçar este problema, está a ser desenvolvida uma área de estudo, Explainable Artificial Intelligence (XAI), especificamente para introduzir um grau de explicação nas previsões realizadas por estes modelos *black-box*. O objetivo principal desta tese foi utilizar XAI sobre estes modelos, de forma a procurar uma uniformização na forma como os dados são introduzidos, previstos e posteriormente, explicados. Para o efeito, foi realizada uma revisão sistemática de literatura, para definir XAI em si, uma taxonomia de categorização de métodos explicativos, bem como determinar aplicações práticas de XAI. Estas aplicações práticas serviram de base para a implementação de diversos métodos XAI, terminando com uma experiência principal, com dados reais. Para as explicações obtidas com os modelos XAI, foi possível criar um conjunto diversificado de explicações para todas as experiências e validá-las com sucesso, uma vez que os métodos XAI estão geralmente de acordo quanto às características consideradas mais importantes para o processo de previsão.

Palavras-chave: XAI, Explicabilidade, Machine Learning, Previsão, Ciência de Dados, Aplicações Financeiras

Classificação JEL:
L86 - Information and Internet Services • Computer Software
M15 - IT Management

Funding

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [ISTAR Projects: UIDB/04466/2020 and UIDP/04466/2020], as well as by project MAIPro POAT-01-6177-FEDER-000059.

Abstract

The growth in usage of Machine Learning (ML) models has been increasing over the past decade, with the practical introduction of high-performance models such as deep neural networks, only obtained by the improvement of computational resources for its usage. These types of models are generally high in performance, though this comes with the cost of being difficult to interpret, gaining the reputation of being *black-box* models. To tackle this issue, a specific area of Artificial Intelligence, eXplainable Artificial Intelligence (XAI) was formed, to introduce a degree of explanations on predictions made by the black-box models. The end goal of this dissertation was to make use of XAI methods on these models, to search for a standardization on how data is inserted, predicted, and finally, explained. For this purpose, a systematic literature review was made, to define XAI as well as a taxonomy of XAI methods and practical applications of such methods. These practical applications serve as a baseline in the implementation of several XAI methods, concluding with a main experiment on a real dataset. For the explanations obtained with the XAI models, it was possible to create a diverse set of explanations for all experiments and successfully validate them, as the XAI methods are typically in agreement over what features were deemed most important for the predictive process.

Keywords: XAI, Explainability, Machine Learning, Prediction, Data Science, Financial Applications

JEL classification:
L86 - Information and Internet Services • Computer Software
M15 - IT Management

Contents

List of Figures

List of Tables

Acronyms

AI: Artificial Intelligence.
ALTAI: Assessment List for Trustworthy AI.
DT: Decision Tree.
GNB: Gaussian Naive-Bayes.
LR: Logistic Regression.
ML: Machine Learning.
MLP: Multi-Layer Perceptron.
RF: Random Forest.
XAI: Explainable Artificial Intelligence.
XGB: eXtreme Gradient Boosting.

CHAPTER 1

Introduction

1.1. Machine Learning and XAI

The rising trend of Machine Learning (ML) models led to wider adoption of such methods in many use cases. These ML models can be white-box approaches, where the internals of the models are observable and self-explanatory or easily interpretable, or black-box models. These models are generally seen as one object, with an input and an output, and not much more in terms of either interpretability or explainability.

The differences between both approaches also encompass performance. One of the key reasons black-box approaches are used in most real-world applications is due to their huge potential for excellent predictive performance, while white-box approaches tend to be directed toward pedagogical matters or where performance is not a key issue. Thus, the investigator needs to be able to discern the advantages and respective disadvantages of each type of approach: Do we need the best predictive performance or do we need to emphasize possible explanations for the model, even if the performance is at best, satisfactory?

The same questions were raised in [**4**], where the author predicted an increased usage of ML models with higher degrees of interpretability through the extension and modification of such models. With the addition of contributions by economists and other social scientists toward the formal definition of problems, and proposing solutions to them, it would result in the implementation of more suited ML models in the area.

This brings us to the area of Explainable Artificial Intelligence (XAI), whose main goal is that of introducing explainability to ML models. The area did not have much traction up until 2018 [**1**], where it could be argued that the area of XAI gained more visibility. Now the argument goes beyond the question of performance versus explainability, there is also the possibility of expanding highly performant methods by introducing a degree of explainability. However, XAI is not exactly what [**4**] had envisioned months prior, and while it largely responds to the questions posed by the author, it is necessary to formally define what it truly represents.

According to the Assessment List for Trustworthy AI (ALTAI), explainability is defined as a "feature of an AI system that is intelligible to non-experts"[1]. For a system to be intelligible it must be explained without recurring to a technical description. However, the definition proposed by ALTAI is not directed toward the investigator who works with black-box models. Another paper refers to explanations as a means for humans to

[1] `https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment`

trust black-box methods by either summarizing its results or by providing insights about decisions that have been made, and to also be auditable [**22**].

The question of trust in implemented ML systems is of extreme importance as seen in [**38**], where the author further details that it is not trust that is enabled, but rather the decision to trust in an ML system.

1.2. The necessity for XAI

The necessity for this research area comes not only from experts who wish to better understand the predictive models that they are working with but also from legal implications regarding the use of Artificial Intelligence (AI) tools.

Firstly, experts want to understand better black-box models to scrutinize their models for validation, as pointed out in [**22**] where an analysis of several papers in the area of image classification revealed that changes were made to a target class through imperceptible transformations to pixels. Or in the same paper where a criminal risk assessment tool was determined to have racially biased predictions. These examples, with unexpected behavior in the case of image classification or with unintentional biases, demonstrate the necessity for this research area.

Not all ML needs an explanation in place. As [**16**] stated, simpler systems where unsatisfactory results are inconsequential or systems that have already been well-documented have their decisions met with intrinsic trust, "...even if the system is not perfect." Rather, the author proposes a need for the interpretability of ML systems due to the "...incompleteness in the problem formalization." The existence of explanations makes the improvement of problem formalization possible.

Secondly, the legal issues, mostly covered by GDPR, raise the need for this research area as a legal obligation. Based on the resulting literature review presented in Appendix D it was possible to find several explicit mentions in GDPR. Focusing on the connection between XAI and GDPR is Article 5, where personal data should be processed in a "...transparent manner in relation to the data subject...," meaning the user who allowed access to its data should be informed of what exactly is being done with its personal data. Another example is seen in Article 14, paragraph 2.g), where it is defined that the data subject has the right to have the information on the "... the existence of automated decision-making...," as well as the process and consequences of such decision-making.

1.3. Context: MAIPro Project of non-compliance monitoring and alert

This dissertation mainly addresses a real-world data study by the POAT/2021 project named MAIPro "Project of non-compliance monitoring and alert". The main purpose of MAIPro is the timely prediction of possible non-compliance by publicly funded projects of either project-planned timings or previously fixed financial goals. The challenge and data were provided by the project's official partner IAPMEI - *Agência para a Competitividade e Inovação, I.P.*. This project performed extensive data and domain understating over IAPMEI's data in order to extract relevant features resulting in the implementation of

ML models to predict several targets, such as ineligibility, cancellation, budget, and time deviation (when compared to the original planning). MAIPro aims to test the potential of an automated system to generate an alert for eventual non-compliance based on data known at the time of the project's application or at other key moments in the project's monitoring. In order to support properly informed decisions, the alert should be substantiated according to the variables directly involved in this result, thus, the explainability of the model's results is fundamental for successful outcomes for this particular project.

The data from IAPMEI concerns the application of funding assigned by the European Union for competitiveness enhancement of Portuguese small and medium companies. With the purpose of reaching an economic and monetary convergence, the EU has been emphasizing the importance of executing its multiannual financial framework in an effective, and efficient manner. While the EU is responsible for financing projects over several areas, it is the EU member states that manage how these amounts are applied. Each member has to manage its own finance support board for its country's regions as well as supervise the funds' application via national institutions. This supervision happens for the totality of the process of application, from the moment the application for funding is made up until an approved and funded project is finished. In the case of Portugal, the public institute that supervises and manages strategic competitive calls for funding, and monitors approved projects' progress is IAPMEI.

IAPMEI's mission is to promote and develop companies' competitiveness and growth, advancing innovation, entrepreneurship, and investment in small and medium companies whose activities are under the tutelage of the Ministry of Economy and Seas.

The institute was created in 1975, with the main purpose of providing assistance to businesses in the industrial sector (Decree-Law No. 57/75). The main responsibilities of IAPMEI (Art. 3.º) included but were not limited to the reorganization and conversion of businesses that might be likely to become competitive, socially, and economically viable, and to promote voluntary cooperation between businesses. For these purposes, the Institute has autonomy for financing or subsidizing studies targeted towards market analysis or economic viability, as well as their respective necessary actions (Art. 4 and 7.º).

Given the financial nature of the institute and its important role in assisting small and medium enterprises through subsidies and incentives, it is not surprising that improving its own efficiency in regard to whom subsidies are given, and in which amount, is one of its main concerns. These incentives are given after an analysis of the applications for the project's funding.

1.4. Problem statement

The project mentioned in the previous section needs to have justifications for its decisions, especially for those that are made based on ML models. These justifications come from the necessity for understanding the predictive models in place. Arguments could be made for the need to comply with GDPR but that regulation was made with EU citizens in mind and not companies. This ultimately means that IAPMEI cannot blindly trust

the results obtained in the predictive process. Therefore, the explanations necessary to justify IAPMEI's decisions come from the implementation of XAI methods which will aid in understanding what features are deemed important by the predictive models.

However, this area of research suffers from a lack of generally accepted taxonomy and definitions, with several authors providing their own suggested taxonomy based on particular perspectives of usage. This is to be expected within a recent area of study, but there is a need to reach (i) a proper taxonomy of XAI methods, and (ii) define the categories for already implemented XAI methods to be classified on. In fact, the different perspectives must be addressed from a more holistic point of view. If the previous two necessities are met, the research area would benefit from a general agreement of opinion on the fundamentals of what XAI is and how XAI methods are classified, leading to more effective dissemination of knowledge among peers, and facilitating the creation of XAI techniques through the exact requirements that are needed.

Another problem that is still prevalent is how is it possible to evaluate XAI techniques, as supported by the authors of [**19**]. [**33**] exemplifies such evaluation as descriptive accuracy, meaning the "...ability of the interpretations to properly describe what the model has learned." For this dissertation, and to tackle this issue, it was decided to evaluate the results obtained through XAI methods without using this descriptive accuracy, at least not directly. Since various XAI techniques are used, the legitimacy of the obtained results is given by the degree of agreement on what features are deemed more important and what features are changed for the case of counterfactual methods.

The main goal is to introduce explainability towards the results of models built over a real-world dataset: IAPMEI data on European-funded projects. The sought explainability will come from the implementation and analysis of XAI methods that have been found with a systematic literature review. However, as previously pointed out there is no concise definition of how XAI models are categorized. This led to the necessity of performing not one but two systematic literature reviews (SLR): the first one to determine how can a taxonomy for these methods be defined, and the other to look for adequate candidates for XAI methods to use in the case study of IAPMEI data. The XAI methods implemented were also tested using two open datasets for their value in demonstrating examples of explainability in somewhat more accessible data.

In summary, this dissertation will answer the following research questions:

(1) What is XAI, and what is its relevancy?
(2) How should XAI techniques be classified?
(3) Are existent XAI methods relevant for real-world applications?

1.5. Contributions

A relevant contribution comes in the form of an accompanying systematic literature review. This survey serves the purpose of assisting in the definition of XAI, the categorization of XAI methods through the realization of a taxonomy, and an analysis of XAI

methods used in finance. This survey has been described in a paper already submitted to IEEE Access[2], and is currently in a second round for editing, upon suggestion by the editor and reviewers. One other contribution of this literature review was the collection of various datasets, some of which were used in this dissertation. In addition, implementing newly researched XAI models on datasets obtained in the literature review, as well as on the dataset provided by IAPMEI, will offer three different applications for these methods, contributing to the expansion of XAI as a research area. The main contribution of this dissertation comes in the practical application of existent XAI methods and how these can be used to enhance trust in the outcomes of black-box models, as seen in Chapter 4.

1.6. Structure

This dissertation is structured into five chapters: Introduction, Literature Review, Experiments, Explaining project's cancellation prediction, and Conclusions.

The first chapter introduces XAI as a research area, stating its purpose and relevance in the increasingly regulated area of AI implementations. Some of the problems surrounding XAI are also described, as well as this work's contributions. Finally, a contextual description of the project where this work is framed is given, to facilitate the comprehension of the nature of the real-world dataset used in the case study.

The second chapter presents the systematic literature review, where a brief summary of the accompanying already submitted paper in which a taxonomy of retrieved XAI methods is proposed. These methods are analyzed for possible further implementation, resulting in the description of how these models operate, along with a final paragraph detailing if and why they were implemented.

In the third chapter, we find two specific experiments made with open data to evaluate how the different XAI methods behave in more easily accessible and understandable data and context. The creation of several experiments serves the purpose of demonstrating these models' capabilities in how we interpret the behaviour of models and their respective explanations, and the implementation of several models helps validate whether the explanations provided by such methods are conclusive or not.

The fourth chapter describes the methodological process of the main case study: domain understanding, data understanding - what it represents, and how it should be processed - data preparation, modelling, evaluation and analysis. This chapter is the focus of this dissertation in terms of experiments, detailing the main bulk of work done in order to understand, process, and model the data, along with the presentation of corresponding results.

The final chapter gathers all key points from each individual section, along with a critical overview of the work done in this dissertation, its limitations, and some recommendations for future work.

[2] `https://ieeeaccess.ieee.org/`

CHAPTER 2

Literature Review

The literature review was made in the form of an accompanying article (Appendix D) presenting two systematic literature reviews made using the SCOPUS citation database[1]. The justification for the usage of this engine instead of Web of Science[2] or Google Scholar[3] comes from the fact that the former has more restrictions in place regarding what papers are indexed, and the latter has no restrictions in what articles are indexed. It was decided that it was necessary to take into account the recency of XAI, and SCOPUS has less restrictions than Web of Science but more than Google Scholar, making it more suitable for the process of gathering studies. Nonetheless, some papers on XAI were also obtained manually through arXiv[4] as they were not indexed in SCOPUS, but had great value for the purposes of the literature review.

The first of the reviews had the purpose of providing context of XAI as an evolving research area and to help understand how to properly categorize XAI techniques. This search was done without adding many restrictions on how papers were initially chosen. However, several eligibility filters were applied for the eligibility of the paper for the subsequent analysis, resulting in a reduction in the number of analyzed papers from 70 to 17 surveys. From these articles, it was possible to propose a concise and rather simple taxonomy of XAI methods that can be found in Appendix D. This taxonomy proposes the categorization of XAI methods in three major categories:

Stage Does the XAI method generate explanations in a "post-hoc" manner, that is, after the predictions are made, or is the model intrinsically explainable, "ante-hoc?"

Model Is the XAI method specific to a model, "model-specific," or can it be applied generally without restrictions, "model-agnostic?"

Scope Are the explanations provided by the XAI method presented in a "global" scope, where it is possible to observe the general behaviour of the predictive model, or are the explanations provided on an instance basis, that is, in a "local" scope?

The second review aimed at understanding what XAI methods were being used in finance, resulting in a thorough review of 33 practical applications. The vast majority of these papers either made use of an XAI method, or implemented a novel technique and, by categorizing each method with the taxonomy proposed in the first review, it was

[1] https://www.scopus.com/search/form.uri?display=basic
[2] http://webofscience.com/
[3] https://scholar.google.com/
[4] https://arxiv.org/

possible to understand, at a glance, how each XAI method operated. Another relevant result of this second review was the collection of the open datasets that have been used in each reviewed paper.

From the literature review, sixteen XAI methods were obtained, all of which are post-hoc in nature. Throughout this chapter, and based on this proposed taxonomy, a summary of how they work is provided. Finally, a justification for the implementation, or not, of each XAI model is provided. Based on the literature review performed in the accompanying paper, and the taxonomy defined there, Table 2.1 presents all XAI methods that were possible candidates for implementation, along with their respective values for Stage, Model, and Scope categories.

TABLE 2.1. Overview of candidate models for the dissertation

XAI Method	**Stage**	**Model**	**Scope**
Anchors	Post-hoc	Agnostic	Local
BELLATREX	Post-hoc	Specific	Local
CASTLE	Post-hoc	Agnostic	Local
CERTIFAI	Post-hoc	Agnostic	Local
DALE	Post-hoc	Agnostic	Global
DiCE	Post-hoc	Agnostic	Local
inTrees	Post-hoc	Specific	Global
LIME	Post-hoc	Agnostic	Local
LTreeX	Post-hoc	Specific	Local
MANE	Post-hoc	Agnostic	Local
PASTLE	Post-hoc	Agnostic	Local
PDP	Post-hoc	Agnostic	Global
PermuteAttack	Post-hoc	Agnostic	Local
Rational Shapley Values	Post-hoc	Agnostic	Global/Local
SHAP	Post-hoc	Agnostic	Global/Local
TREPAN/Hidden-layer-clustering	Post-hoc	Agnostic	Local

Anchors[**37**]

This local method provides explanations based on sets of rules that are most relevant for the predictive outcome. It follows the philosophy of interpretable explanations, those that are more easily understandable to humans.

The algorithm first starts with an empty rule, which can be applied to every instance. During each iteration of the algorithm an optimization is made by creating more restrictive rules which are then applied. These rules extend the initial set of rules by one additional predicate. This choice is based on the candidate with the highest estimated precision, where this new candidate is verified as being the desired Anchor. If it is, then the algorithm terminates and presents the explanation, while if otherwise then the process of candidates repeats.

Anchors can be applied to any given predictive model (model-agnostic), and is local in nature, as its explanations come in the generation of sets of rules for a singular instance

of data that help the user understand more clearly why the predictive outcome is what it is.

BELLATREX - Building Explanations through a LocaLly AccuraTe Rule EXtractor[**13**]

The proposed method is model-specific, as it is used to explain predictions made by Random Forests and it is a local method, as it explains a singular instance.

The algorithm extracts trees that generate the most similar predictions to the instance being observed. These trees are then represented as a vector, either being a function of the tree or the path used by the instance, further solidifying the local nature of the method.

With these vector representations, a projection is made through Principal Component Analysis (PCA), in an effort to "... remove the noise, to improve computational efficiency and to enable a better visualization of the subsequent clustering.". Afterward, clustering is made to these projections through K-Means++ [**3**], with the final step being the construction of a surrogate model prediction based on these clusters, which will serve as the explanation.

BELLATREX is a method that can only be applied to Random Forests, and thus is model-specific. This algorithm is local in nature, aiming to explain a singular instance. Its explanations are derived from a surrogate model, where rules are extracted to justify the predictive outcome.

CASTLE - Cluster-aided space transformation for local explanations[**28**]

CASTLE, or Cluster-aided space transformation for local explanations, is a XAI method that extracts rule-based explanations from black-box classification models, via global knowledge (for all instances). Firstly, clusters of instances are identified based on their common behavior and classification by the predictive model, with these clusters representing global knowledge. These clusters should be homogenous, by satisfying properties such as high purity, high coverage, and low overlap.

CASTLE is a model-agnostic approach, and although it combines both knowledge at the instance and global level, it aims to explain a target instance (local). For the explanations, a space transformation approach is done with the purpose of explaining the instances closest to the instance of interest, with the resulting, and transformed data being fitted on a transparent model such as a decision tree or a linear model.

CERTIFAI[**39**]

This XAI technique introduces the explainability of black-box models through the generation of counterfactual examples. These examples are generated from a target instance of data using a genetic algorithm, resulting in outcomes different from the original instance. An evolutionary process is then applied to approximate these counterfactual examples to the original data instance.

This method is applicable to any predictive model (model-agnostic), and operates locally to explain individual instances through the generation of counterfactual examples.

DALE - Differential Accumulated Local Effects[23]

Differential Accumulated Local Effects (DALE) is an XAI method which approximates ALE, as it is typically infeasible to compute, with issues in high-dimensional datasets, and with vulnerability to out-of-distribution sampling. Therefore, this method addresses these drawbacks by exploiting partial derivatives without altering data points. DALE is presented in two formats: a first-order DALE approximates the local effects of individual features, while the second-order DALE approximates the combined effects of pairs of attributes. Either of these formats protects from out-of-distribution sampling and is faster than the exact calculations when using ALE.

DALE is a model-agnostic approach, and tries to explain the behaviour of the predictive model as a whole through the generation plots, where it is possible to visualize the effect a feature has on the target variable.

DiCE - Diverse Counterfactual Explanations[32]

This local method introduces explainability by creating counterfactual explanations of an instance of data. These explanations come in the form of transformed instances where the predictive outcome is different from the original sample. This method tries to mitigate the problem of creating samples where feature values are radically different from the original sample, so the authors propose an approach that tries to combine explanations generated that are close to the original instance, that have fewer changes to features, but also by adding user constraints, to introduce some level of feasibility to the counterfactual instances.

This counterfactual method, being model-agnostic, can be applied to any predictive model. It explains a target instance through the creation of examples based on this target instance but with different feature values, and distinct outcomes.

inTrees[15]

This XAI method introduces explainability through the extraction of interpretable information from tree ensembles such as Random Forests or boosted trees. The authors illustrate the framework by having three major components: rule extraction; rule processing; and rule summarizing into a learner.

For rule extraction, the method extracts rule conditions as well as their respective rules, and forms rules by assigning outcomes to the conditions. In the rule processing section, inTrees ranks rules, pruning irrelevant variable pairs of a rule, proceeding with a selection of "relevant and non-redundant rules", and finally, the discovery of recurring feature interactions. The last step sees the summarized rules placed into a learner and then used to predict new data.

The usage of this model-specific technique is limited to predictive models such as Random Forest or boosted trees. It has a global scope as it summarizes the behavior of the predictive model by presenting a summary of the rules used in the predictive process.

LIME - Local Interpretable Model-agnostic Explanations[36]
A method whose explanations derive from the identification of an "interpretable model over the interpretable representation" of the classifier. This means that the feature representation must be understandable by humans for the explanation to be interpretable.

For this purpose, LIME calculates the locality of the instance that is observed, as well as the complexity of the explanation. This complexity is exemplified as being the depth of the tree, in the case where a decision tree is the predictive model. Finally, the method calculates the unfaithfulness of the method that is the explanation by LIME, in relation to the model which is observed. The goal for these variables is to minimize unfaithfulness while having the complexity as low as possible in order to be interpretable by humans.

To minimize this unfaithfulness, and for the method to be model-agnostic, a sampling of data, weighed on the proximity to the instance at hand is made. This sampling is made on nonzero elements at random. Afterward, the original representation, the instance at hand, is used as a label for the explanation model, where LIME goes through an optimization process to create the explanation.

LIME is a model-agnostic approach that attempts to explain a singular instance through the visual representation of feature importance in the predictive process.

LTreeX[14]
LTreeX is described by the authors as a local method, meaning it tries to provide insights for specific instances of a dataset. The method summarizes Random Forests, which creates a surrogate model directly from this predictive model. This method selects a subset of trees that are closer to the original ensemble, where a vector representation of these trees is created. Afterwards, clustering is applied to reduce the dimensions of these trees, in an effort to reduce the complexity of interpretability of such trees. The application of this clustering generates rules which are then presented to the user as an explanation for the random forest's prediction. While the initial proposal only dealt with binary classification problems, the authors extended the XAI method to include multi-label classification problems.

LTreeX is a model-specific algorithm, and can only be used with Random Forests. The explanations generated by this model come in the form of sets of rules, and try to explain one target instance.

MANE - Model-Agnostic Non-linear Explanations[41]
MANE is a model-agnostic approach, designed specifically for providing explanations to deep learning models. Fundamentally, the technique works by treating the target classifier as a black-box, where its features are processed by Gradient Boosted Decision Trees in order to extract cross features, to resolve the problem of nonlinear decision boundaries, where linear regression is then applied to approximate such boundary of the target classifier. This enables the method to understand the behavior patterns of the instance of data, and by comparing with other methods such as LIME, the authors conclude that MANE's resulting explanation is simpler and more effective than the previously mentioned XAI

methods, requiring fewer features to explain a prediction with retaining lower error than these methods.

This model-agnostic technique can be employed by any deep learning model, and whose explanations are presented in the form of a sequence of the most important features for the predictive process.

PASTLE - Pivot-Aided Space Transformation for Local Explanations[**29**]
This local algorithm focuses on reducing the sample space of the data into pivots, regions in space where features behave similarly among different instances of data. This involves a projection of data in a space where each dimension represents the proximity to a pivot. Then, a transparent model is fitted on this new data and will assign a weight to each dimension, indicating how the proximity to the pivot influences the outcome. Transparent models are fitted on the original and the transformed data, with the goal of approximating the decision function of the black-box model in the closest range of the target instance, even if their output is different

PASTLE is a model-agnostic method that aims to explain one instance of data through not only feature importance, but also by indicating the necessary changes for the target feature to have a different outcome.

PDP - Partial Dependence Plots[**20**]
This global method applies the Monte Carlo method to demonstrate the marginal effect of one or two features on the predicted outcome of an ML model. Such effect is calculated by selecting the features of interest, where they are separated from the rest of the features. From the latter feature space the marginalization of ML predictions over such features is made, showing the relationship between the two features in the former set. Then, by applying the Monte Carlo method leads to the calculation of the average marginal effect on the prediction.

Partial Dependence Plots is a model-agnostic technique, and with a global scope as the explanations are created to help understand how the target feature is affected by another feature.

PermuteAttack[**25**]
This is a counterfactual method, meaning the XAI model generates synthetic examples based on a real instance where the target feature has an opposite value. PermuteAttack is a genetic algorithm, starting with a random set of synthetic samples and for each iteration, these samples go through selection by randomly selecting new samples based on their fitness, crossover, which is the combination of features of two parents in a random order and finally, mutation, which is the process of perturbing some randomly chosen features. The final goal is to obtain an instance with the least number of permuted features and with a minimal change to the value of such features, resulting in a counterfactual explanation.

The algorithm perturbs data by changing the values of randomly selected features. These changes are random but feasible to exist within the real data. Then, the selection process begins, where fitness is calculated for all selected samples. The samples with

higher fitness are more likely to be part of the next iteration, while the opposite is true for those with lower fitness. From the samples that go through to the next iteration, they are grouped by two with this new pair called parents. In the mutation process, the features of parents are swapped, leading to the selection of some of the features according to their importance in changing the outcome, with their respective values being randomly replaced by a possible value present in training data. This is the mutation step, where it leads to less likelihood of changes in features with a lower effect on the outcome.

PermuteAttack is a model-agnostic approach as it can be applied to any predictive model, and generates explanations through the creation of counterfactual examples for one specific instance. Thus, it operates in a local scope.

Rational Shapley Values[**44**]

The method developed by [**44**] synthesizes both Shapley Values - a game theory concept where the algorithm places the features in a competition to determine the winner (the target value), and counterfactual explanations in a single method. The XAI method searches for a sub-group of instances similar to the sample of interest, and calculates the resulting explanations. Rational Shapley Values also introduces user input to assist the XAI model in terms of what features should not be changed, and their order of perceived importance.

This method is model-agnostic, and is local in scope due to its synthesis of both Shapley Values as well as counterfactual explanations. The counterfactual examples are based on a target instance of data, where feature values are changed according to the feature importance given through the application of Shapley Values.

SHAP[**31**]

This method is an approximation of the game theory concept of Shapley Values. SHAP is often used instead of Shapley Values due to the original concept being complex in terms of computational resources and therefore unfeasible for practical applications. As stated by the authors, the method attributes "to each feature the change in the expected model prediction when conditioning on that feature". It can also be used to understand the general behavior of the model through the generation of explanations that indicate the relative importance of features for the predictive process. This means that each feature will have a positive or negative coefficient if it impacted positively or negatively the respective prediction.

SHAP can be applied to any predictive model and its explanations help understand a target instance or the behaviour of the model as a whole. This is achieved through the calculation of feature importance.

TREPAN/Hidden-layer[**12**]

The authors of this approach combine TREPAN trees and neural networks to provide localized explanations. The method starts by clustering the data from a hidden layer representation of a neural network, where the TREPAN methodology is applied in order to build a decision tree at a cluster level. This approximates the neural network at each

cluster, resulting in a set of rules defined by the TREPAN tree that are used for each cluster, in order to explain the target feature, specifically, the majority class at each leaf node. The explanations from this model stem from each leaf node of the TREPAN tree, which provides a set of rules, defined as reason codes in the paper, that explain the majority class.

This last method is model-agnostic and generates explanations for a local instance, presenting them as sets of rules to justify the prediction.

Several XAI methods were found in the literature review but it was not possible to implement the majority of these methods. As presented in Table 2.2, six of the XAI methods did not have a code repository, and one did not have an implementation in *Python*, making it impossible to adapt to the code structure. One method was found to be model-specific, and since this dissertation tries to provide explanations for any predictive model, it had to be excluded from usage in the Experiments. Finally, an attempt was made to use three more XAI methods but, due to constraints in their adaptation, had to be excluded from the Experiments.

In short, the explanatory methods used provide an insight on feature importance (PDP, SHAP, and LIME) or, by selecting a singular project, creating synthetic samples that have the opposite target value (DiCE and PermuteAttack).

Table 2.2. Summary of viewed methods and reasons for their implementation or exclusion for this dissertation

XAI Method	**Implemented?**	**Reason for no implementation**
Anchors	No	(i)
BELLATREX	No	(ii)
CASTLE	No	(ii)
CERTIFAI	No	(ii)
DALE	No	(i)
DiCE	Yes	-
inTrees	No	(iii)
LIME	Yes	-
LTreeX	No	(ii)
MANE	No	(ii)
PASTLE	No	(i)
PDP	Yes	-
PermuteAttack	Yes	-
Rational Shapley Values	No	(iv)
SHAP	Yes	-
TREPAN/Hidden-layer-clustering	No	(ii)

Reason codes are as follows: (i) - Constraints when implementing the model; (ii) - No repository or code found; (iii) - Model-specific approach; (iv) - Repository is not available in Python

CHAPTER 3

Applications of XAI in Public Datasets

As stated in the introduction, experiments were made on three datasets, of which two were obtained through the reviewed literature and one was accessed in the context of the MAIPro project, courtesy of our partner (IAPMEI). The chosen two initial datasets have the purpose of demonstrating how the XAI methods operate in a more transparent manner since the fact that they are public means the whole pipeline of data preparation up to explanation is available indeterminately.

3.1. Methodology and Predictive Model Selection

The methodology used for the experiments is the Cross Industry Standard Process for Data Mining (CRISP-DM)[**40**]. It was created in late 1996 from the need for a more standardized process model for data mining, and this process was later expanded in 2000. It is divided into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Business understanding is perhaps the most important stage of this methodology as it is here where the objectives of the study are formed. Not only that, it is in this phase where the overall plan for achieving these goals is made. For the second phase, data understanding, an initial gathering of the data is made along with an overview of its description and possible problems regarding data quality. Data preparation is the third stage, where the processing of the data is made to construct the dataset to be used, whether that means cleaning nulls, filling empty values, or creating features from attributes present in the untreated data. For the fourth stage, the implementation of models is made, with the purpose of satisfying the goals outlined in the first stage. The fifth phase of this methodology is the evaluation of obtained results in the previous stage, and its respective analysis. It is expected to have a thorough review of previous steps to determine whether some factors had been overlooked.

The final stage of CRISP-DM is the Deployment of the tools built. This phase was not within the project's scope, although we hope this dissertation can help guide the application of this methodology to the case at hand and similar cases.

For the purposes of this dissertation, the programming language chosen is *Python*. This choice comes from its immense relevance, not only in the implementation of ML models, but also with its maturity in terms of available libraries for all processes related to Data Science, from data profiling libraries to black-box models' libraries.

Since the main purpose of this dissertation is the explanation of black-box models on different experiments, it requires a good understanding of how these methods are implemented, and how they interact with the data as well as the explanatory methods. Among the publicly available libraries that include predictive models, are PyTorch[1], Keras[2], TensorFlow[3], and Scikit-learn[4]. While all are widely used in data science, only Scikit-learn has the traditional ML models, whereas PyTorch, Keras, and TensorFlow all focus on deep learning.

For the predictive models, a general approach was made, and without any specific criteria for the choice of predictive models. This resulted in the final list of models used: Logistic Regression (LR); Gaussian Naive-Bayes (GNB); Random Forest (RF); Multi-Layer Perceptron (MLP); eXtreme Gradient Boosting (XGB); Decision Tree (DT). The implementation of several predictive methods will assist in determining the most suitable model for each differing dataset and also validate the obtained results.

The main contribution of the usage of these publicly available datasets was to build a code structure that was adaptive to each dataset and required little change in order to function, analyze incoming data, predict, and explain the predictions made. One of the keynotes taken from the literature review was that there was a lack of experiments. Since each XAI method experimented on one or two datasets, and as each paper presented a singular method, there is a lack of comparative experiments. This dissertation also aims to address that research gap.

Finally, the initial experiments with these two datasets were fundamental to the creation of this code structure, and after the experiment with IAPMEI's dataset, this was extremely clear, with little to no adaptation needed in order to complete the process of analysis, prediction, and explanation of these predictions.

3.2. Experiment 1: German Credit dataset

Business Understanding and Data Profiling The German Credit dataset [**27**] classifies people as having either good or bad credit risk. It contains 1000 instances and 20 features. Donated in 1994 to the UCI repository[5], it still is amply used in machine learning tasks including practical applications for XAI, as numerous studies experimented with this specific dataset [**7**],[**10**],[**11**],[**15**],[**24**],[**44**],[**34**].

In reality, the German Credit dataset presents two versions: the original dataset, with categorical features, and another version, created by Strathclyde University, where the original dataset was altered to make it more suitable for using algorithms that do not support categorical variables, resulting in such features being coded as numeric. To understand what each of the existent features means Table 3.1 presents a dataset's dictionary.

[1] `https://pytorch.org/`
[2] `https://keras.io/`
[3] `https://www.tensorflow.org/`
[4] `https://scikit-learn.org/stable/`
[5] `http://archive.ics.uci.edu/`

TABLE 3.1. Feature descriptions

New attribute name	Description
Status of existing checking account	Status of existing checking account
Duration in month	Duration in month of the credit
Credit history	Credit history
Purpose	Purpose of the credit
Credit amount	Credit amount
Savings account or bonds	Savings account/bonds
Present employment since	Present employment since
Install. rate (%) of disposable income	Installment rate in percentage of disposable income
Personal status and sex	Personal status and sex
Other debtors or guarantors	Other debtors/guarantors
Present residence since	Present residence since
Property	Property
Age in years	Age in years
Other installment plans	Other installment plans
Housing	Housing
No. of existing credits at this bank	Number of existing credits at this bank
Job	Job
No. people being liable for	Number of people being liable to provide maintenance for
Telephone	Telephone
Foreign worker	Foreign worker
Risk	Good (=1) or Bad (=2) Risk

Data Preparation This dataset was carefully prepared, and as such, it was not necessary to remove features or to detect outliers. Although the feature names are rather intuitive, their corresponding values are not. Based on the dictionary present in the dataset's repository, we proceed with two changes of the current feature values: for any feature where the order does matter, old feature values were replaced with sequential numerical values. This first transformation resulted in the replacement of feature values present in Table 3.2.

For all categorical features where order does not matter, the corresponding values were mapped to their natural language counterpart using an adequate procedure. Then, a One-Hot Encoding (OHE) has been applied to these newly-altered values. The second transformation to feature values is shown in Table 3.3 (which will then be used for proceeding with an OHE).

Finally, we needed to scale the numerical features. Since any features that were transformed with OHE are in the range of [0,1], we proceeded to apply scaling to the features: 'Duration in months', 'Credit Amount', 'Age in years', 'No. of existing credits at this bank', 'No. people being liable for'.

Modelling The target feature for this dataset is whether a client is described as having good or bad credit risk. In total, the distribution of instances by the target feature - Good

Table 3.2. New feature values for ordinal features

Feature	Old value	New value	Value meaning
Status of existing checking account	A11	1	... < 0 DM
	A12	2	0 <= ... < 200 DM
	A13	3	... < 0 DM
	A14	4	... < 0 DM
Savings account or bonds	A61	1	... < 100 DM
	A62	2	100 <= ... < 500 DM
	A63	3	500 <= ... < 1000 DM
	A64	4	... >= 1000 DM
	A65	0	unknown / no savings account
Present employment since	A71	1	unemployed
	A72	2	... < 1 year
	A73	3	1 <= ... < 4 years
	A74	4	4 <= ... < 7 years
	A75	5	... >= 7 years
Job	A171	1	unemployed/ unskilled - non-resident
	A172	2	unskilled - resident
	A173	3	skilled employee / official
	A174	4	management / self-employed / highly qualified employee / officer

(0) or Bad (1) risk is unbalanced, presenting 635 Good (0) risk instances and 228 Bad (1) ones, that is, the instances representing the class "Bad (1) risk" are only 26.42% of all instances. The complete distribution of features for this dataset is detailed in Appendix A, where the information of mean, standard deviation, and quartiles is also presented.

Starting with a search for the best set of hyper-parameters for the predictive models, the best ones were selected for the prediction process. To train the predictive models a split on train/test data is made, representing 70%/30% of the total dataset respectively. Finally, explanations were generated with the chosen XAI methods. For the predictive process, to improve the legitimacy of obtained results, a set of 5 different seeds was used. These seeds are used both in the split of train and test data, and when initialising the predictive models.

Evaluation An issue that occurred with the fitting of this dataset has been the existence of over-fitting for the models Random Forest and XGBoost. As shown in Table A.2, both Random Forest and XGBoost indicate the presence of over-fitting in the training data.

One of the justifications for this might be the complex range of hyper-parameters available. This can be observed in the running times displayed in Table 3.4, showings that these models spent much more time in hyper-parameter optimization when compared to the other models. Another possible justification for the presence of over-fitting in these models is the fact that the training dataset contains a low number of samples.

Overall, results as presented in Table 3.5 were satisfactory , with all models except for Decision Tree and Gaussian Naive-Bayes reaching similar results on the metric F1-score.

Table 3.3. New feature intermediate values for OHE application.

Feature	Old value	New value
Credit history	A30	none_paid_duly
	A31	all_paid_duly
	A32	existing_duly_until_now
	A33	delay
	A34	critical
Purpose	A40	car_new
	A41	car_used
	A42	furniture_equipment
	A43	radio_television
	A44	domestic_appliances
	A45	repairs
	A46	education
	A47	vacation
	A48	retraining
	A49	business
	A410	others
Personal status and sex	A91	male_divorced_separated
	A92	female_divorced_separated_married
	A93	male_single
	A94	male_married_widowed
	A95	female_single
Other debtors or guarantors	A101	none
	A102	coapplicant
	A103	guarantor
Property	A121	real estate
	A122	soc_savings_life_insurance
	A123	car_other
	A124	unknown
Other installment plans	A141	bank
	A142	stores
	A143	none
Housing	A151	rent
	A152	own
	A153	free
Telephone	A191	none
	A192	yes
Foreign worker	A201	yes
	A202	no

Notably, while the precision values of the better models are above average, the recall values are rather low. When comparing each model's results there is no clear winner but, after excluding Random Forest and XGB not only due to their running times (Table 3.4) as well as the problem of over-fitting, Logistic Regression becames the best model.

With the choice of Logistic Regression as the predictive model comes explanations to help understand the model. As briefly stated in the last paragraph of the literature review,

TABLE 3.4. Running time for hyper-parameter optimization and prediction process (in seconds)

Model Name	Hyper-parameter Optimization	Prediction
DT	0.81	0.02
GNB	0.04	0.02
LR	4.24	0.02
MLP	31.49	0.09
RF	200.59	0.20
XGB	254.84	0.26

TABLE 3.5. Results on test data

Model	ROC AUC	Accuracy	F1-score	F1-score weighted	Precision	Recall
DT	54.42%	73.75%	20.93%	67.39%	52.94%	13.04%
GNB	57.17%	40.54%	45.39%	37.58%	30.05%	92.75%
LR	68.45%	78.76%	53.78%	77.57%	64.00%	46.38%
MLP	47.23%	26.64%	39.87%	14.98%	25.51%	91.30%
RF	65.02%	79.15%	47.06%	76.37%	72.73%	34.78%
XGB	66.01%	77.22%	49.57%	75.77%	60.42%	42.03%

the majority of obtained explanations come from Feature Importance - calculations made on a predictive model to determine the relative importance of the features present in the dataset - obtained through the usage of PDP, SHAP, and LIME. Additional explanations are provided in the form of artificial samples, generated by counterfactual XAI methods (DiCE and PermuteAttack). Logistic Regression is not an inherently black-box model and deriving explanations from it makes it easier to understand other XAI methods that can be used.

Proceeding with the evaluation of XAI methods applied on this dataset we found out that PDP, SHAP, and LIME agree on which is the most important feature for prediction: 'Status of existing checking account'. Table 3.6 shows features and their importance for each one of the three XAI methods. The remaining features, ranked by their respective importance to the target, all have a similar impact, which is relatively low when compared to the one displayed by 'Status of existing checking account'. Table 3.6 also shows that there is some disagreement over what features are most important. It is possible to observe that some features appear in one XAI method but not in the others. For instance, the feature 'Other installment plans_none' displayed some relevance with SHAP but not with the remaining methods. Appendix A contains all the results for each XAI method, when applicable.

Regarding the use of counterfactual methods we found out that DiCE does not reach the same conclusions of the previous three methods. The usage of this technique resulted in the change in values for two different features, not present in the list of most important

TABLE 3.6. Most important features

Feature	PDP	SHAP	LIME
Other installment plans_none	-	0.03	-
Install. Rate (%) of disposable income	-	0.03	0.04
Credit history_all_paid_duly	0.14	-	-
Personal status and sex_male_single	-	0.02	-
Purpose_retraining	0.1	-	-
Other installment plans_none	-	-	0.03
Credit history_none_paid_duly	0.2	-	0.03
Duration in month	0.14	0.03	0.04
Present employment since	0.12	0.03	0.04
Purpose_car_used	0.16	0.02	0.05
Purpose_car_new	0.12	0.05	0.06
Status of existing checking account	0.25	0.12	0.13

features presented in Table 3.6. The features DICE found important are presented in Table 3.7, and are the following: 'Age in years' and 'Credit amount'. As the features found important by DiCE do not belong the list of most important features in any of the other XAI models, but the other XAI models are in agreement regarding the most important features, a decision was made to not use the explanations obtained with DiCE.

TABLE 3.7. Changes made to feature values when applying DiCE as an explanatory technique.

Age in years	Credit amount
0.27	**0.55**
0.27	0.02
0.20	0.09
0.17	0.06

The first row represents the original instance, and the three other rows represent the generated counterfactual examples.

On the contrary, PermuteAttack changes two feature values in its generated counterfactual instances that are both present in the previously obtained results with the XAI methods PDP, SHAP, and LIME. The two features that had their values changed, and their new values are presented in Table 3.8.

Now it became necessary to reflect on the validity of the results obtained. The explanations generated by XAI methods should not be biased by personal or sensitive data, whether that be age, gender, or ethnic background. This is due to the fact that these features represent intrinsic characteristics of individuals, rather than objectively important factors for the predictive process. An analysis of obtained results point to the possibility of some bias, especially toward gender, with the feature 'Personal status' and 'sex_male_single' being considered important by SHAP. However, this feature is not considered in PDP nor LIME and has not been used in counterfactual examples generated

TABLE 3.8. Changes made to feature values when applying PermuteAttack as an explanatory technique.

Status of existing checking account	**Present employment since**
0.67	**1.00**
0.67	0.50
0.00	1.00

The first row represents the original instance, and the two other rows represent the generated counterfactual examples.

either by DiCE or by PermuteAttack. Another feature raising the question of bias is 'Age in years' since it is used in counter-examples created by DiCE, with two of the three generated samples presenting this feature with lower values. Since none of the other features raise this problem and the features mentioned above are not being generally agreed upon as important to the prediction of the target, the validity of results is assured.

3.3. Experiment 2: Default Credit Card Clients

The second dataset for this chapter classifies the probability of default in credit card payments of Taiwanese customers. The dataset contains 30,000 samples and 24 attributes. The original purpose of this dataset was to determine and compare the predictive accuracy of the probability of default among six different methods [**47**], in which the predictive outcome would be the probability of default instead of a binary answer of yes or no. The model with the highest performance was the Artificial Neural Network.

This dataset was donated in 2016 and is freely available at the UCI repository[6]. The dataset is relevant in both the classic tasks of prediction in ML as well as in the practical applications of XAI methods [**12**], [**45**], [**9**]. Table 3.9 shows the need to rename each feature, representing the old feature name, as well as the new name for each feature, and its definition.

Business Understanding and Data Profiling As stated in the introduction of this experiment, this dataset contains the information on payment of credit cards. These payments can be from the usage of credit for several purposes, but after an initial analysis of the conversion rate of NT$ to EUR, it seems more likely that the credit usage comes from personal credits (the exchange of NT$ to EUR stands at 1 NT$ = 0.030789 EUR - on the day of the query, Feb 28, 2023). Unfortunately, there is a lack of contextual information regarding the nature of the dataset but also in what conditions it was created, and as the feature list in Table 3.9 shows, there is no way to know exactly what these credits were used for, as there is no feature regarding the purpose of the credit. This makes it more difficult to understand the dataset and what records should be kept or removed.

The dataset contains categorical features, detailed in numeric coding. The translation of the meaning of such features is shown below (Table 3.10):

[6] `https://archive.ics.uci.edu/`

Table 3.9. Feature descriptions.

Original Attribute	New Attribute name	Description
X1	Given credit (NT$)	Amount of given credit in NT dollars. Both individual and supplementary credit
X2	Gender	Gender
X3	Education	Level of education
X4	Marital status	Marital status
X5	Age	Age
X6-X11	Past, monthly payment (-1) - Past, monthly payment (-6)	History of past, monthly payments
X12-X17	Past, monthly bill (-1) - Past, monthly bill (-6)	Amount of past, monthly bill statements
X18-X23	Prev. payment in NT$ (-1) - Prev. payment in NT$ (-6)	Amount of previous, monthly payment in NT dollars

Table 3.10. Categorical features meaning

Feature	Value	Meaning
Gender	1	Male
Gender	2	Female
Education	1	Grad. School
Education	2	University
Education	3	High School
Education	4	Others
Marital Status	1	Married
Marital Status	2	Single
Marital Status	3	Others
Past, monthly payment	-1	Paid duly
Past, monthly payment	1	Payment delay for one month
Past, monthly payment	2	Payment delay for two months
Past, monthly payment	3	Payment delay for three months
Past, monthly payment	4	Payment delay for four months
Past, monthly payment	5	Payment delay for five months
Past, monthly payment	6	Payment delay for six months
Past, monthly payment	7	Payment delay for seven months
Past, monthly payment	8	Payment delay for eight months
Past, monthly payment	9	Payment delay for nine months

Feature names are also somewhat unintelligible, but when connecting to the information given regarding the features' dictionary, their respective meaning is clearer. Regarding inconsistencies of values present in the dataset, the feature 'Education' has several values that are not defined in the dictionary, and the same is said for the feature 'Marital status'. Finally, the features 'History of past, monthly payments' (X6-X11) have instances in an undocumented state (-2), the majority of observations have an undocumented value

(0), and there is no instance with the value of 9, even though this value is documented in the feature dictionary. However, in [**9**] the dictionary is described, albeit with some differences to the original repository. Namely, the features 'History of past, monthly payments' have the interval of [-2, 9] as being possible values. Another difference is that undocumented feature values are placed in the category "Others".

Data Preparation Even if there is a lack of contextual information on the conditions by which the dataset was created, there is still a need to perform some treatment to ensure the best quality of data so as to provide valid results of both predictive and explanatory methods.

The first change made to the dataset was the feature names, to improve readability and prepare for the presentation required by XAI methods. For this purpose, feature names were transformed based on their respective description. The corresponding table with the association of the original feature name and changed feature can be found in the previous section (Table 3.9), containing the new feature names in the column "New attribute name".

The second change was the removal of observations based on undocumented feature values. The only features that presented such values were 'Education', and 'Marital status'. In summary, the 399 observations under the conditions reported in Table 3.11 were removed from the dataset.

TABLE 3.11. Rows removed based on undocumented values

Feature	**Value**	**Observations removed**
Education	0	14
Education	5	280
Education	6	51
Marital status	0	54

The features 'Past, monthly payment (-1)' and 'Past, monthly payment (-2)' also presented values whose documentation was unclear. It was concluded that it was necessary to search for supporting literature that made use of this dataset and that we were missing crucial information that would be essential to propose the correct approach. So, after analyzing work done on this dataset, more precisely the work of [**9**], it was decided to define the undocumented values for these features. As such, the undocumented values of -2 were given the meaning of "No consumption", and the value of 0 was given the meaning of "Use of revolving credit".

Next, the correlation matrix for the features related to past payments and bills was analyzed in order to simplify the dataset's structure. This resulted in the correlation matrix below which details those with high correlation with one another. Based on the correlation matrix present in Fig. 3.1, it was decided to remove the following features: 'Past, monthly payment (-3)'; 'Past, monthly payment (-4)'; 'Past, monthly payment (-5)'; 'Past, monthly payment (-6)'; 'Past, monthly bill (-3)'; 'Past, monthly bill (-4)'; 'Past,

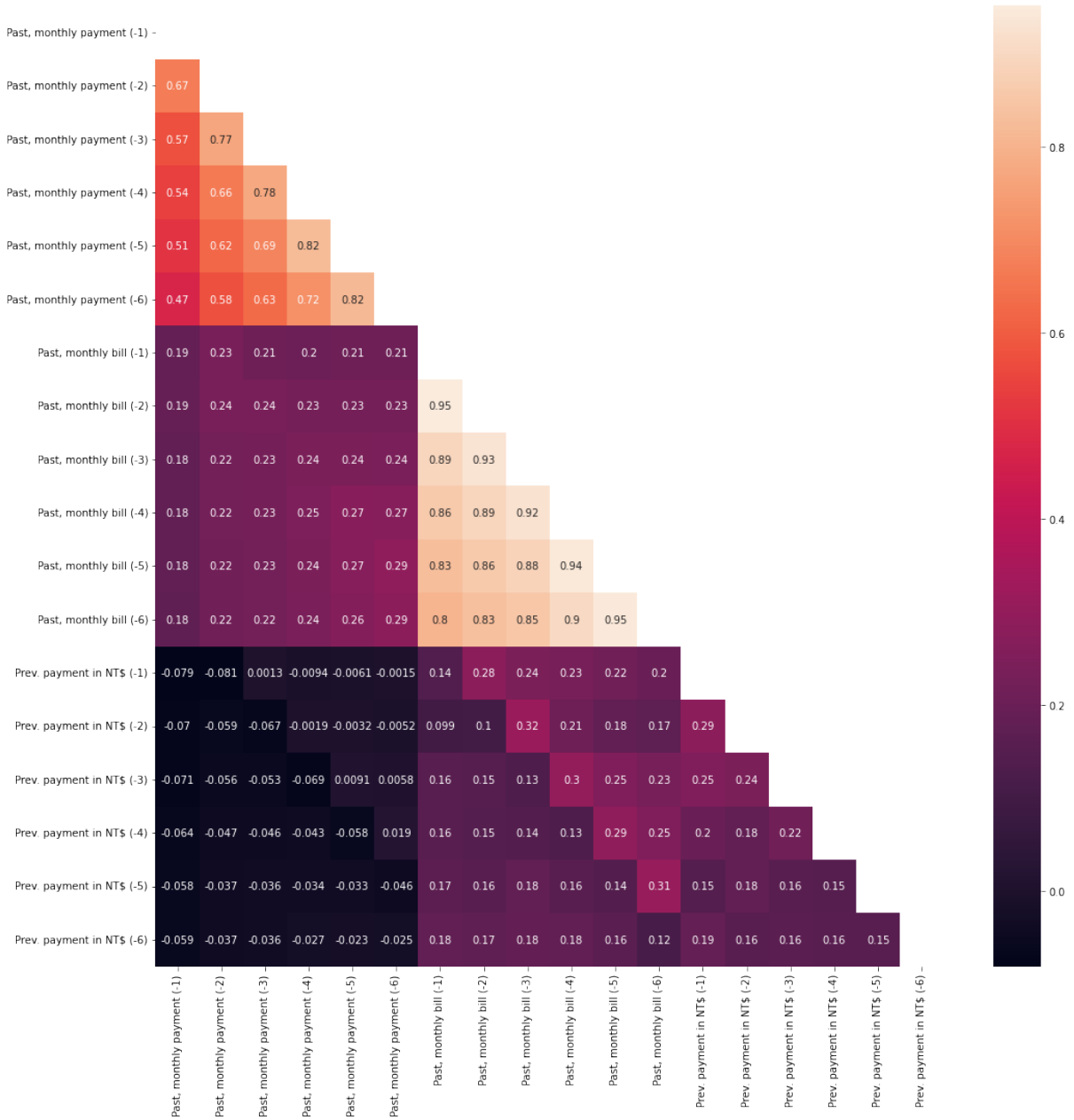


FIGURE 3.1. Correlation matrix for the features related to past payments and bills

monthly bill (-5)'; 'Past, monthly bill (-6)'; 'Prev. payment in NT (-3)'; 'Prev. payment in NT (-4)'; 'Prev. payment in NT (-5)'; 'Prev. payment in NT (-6)'. The decision for the removal of these features came from the fact that while they had a high correlation with features within the same group, this was not verified for the rest of the features. As such the features containing more recent information were considered to be more relevant.

Another step in data processing was the transformation of categorical features by applying One Hot Encoder (OHE). This strategy is largely used when the order given by such type of features is not important, but it is necessary to explicitly define this, as predictive models do not know this beforehand. As such, OHE is only applied to the

features 'Gender' and 'Marital status'. The numerical encoding for these features is given by Table 3.12.

TABLE 3.12. Value change for the features Gender and Marital status for OHE

Feature	Old value	New value
Gender	1	male
	2	female
Marital status	1	married
	2	single
	3	others

One other transformation was the outlier identification and treatment. Two differing approaches were made, one specifically for the features 'Past, monthly payment', and another using Inter-Quartile Range (IQR) for the numerical features.

The first step involved the direct removal of values of the 'Past, monthly payment' features that were above 2, as an initial experiment with Inter-Quartile Range was made to these features, but this resulted in all values above 1 being removed. It was decided the best course of action was to keep the values whose drop-off in frequency was small, as the frequency of observation whose values are more than 2 are negligible. In total, for these two features, 751 instances were removed from the 29601 remaining values.

For the remaining numerical features, the measure IQR was utilized to identify possible samples for removal and was applied to all numerical, and continuous features: 'Age'; 'Given Credit (NT$)'; 'Past, monthly bill (-1)'; 'Past, monthly bill (-2)'; 'Prev. payment in NT$ (-1)'; 'Prev. payment in NT$ (-2)'. IQR was applied with a range of 1.5 and resulted in the removal of 6034 instances, with the reduced dataset now totaling 22816 observations.

Finally, the last transformation for this experiment is the scaling of all features to ensure all features are weighed equally. With all these steps applied to the dataset, we kept the following features: 'Given credit (NT$)'; 'Education'; 'Age'; 'Past, monthly payment (-1)'; 'Past, monthly payment (-2)'; 'Past, monthly bill (-1)'; 'Past, monthly bill (-2)'; 'Prev. payment in NT$ (-1)'; 'Prev. payment in NT$ (-2)'; 'Gender_female'; 'Gender_male'; 'Marital status_married'; 'Marital status_others'; 'Marital status_single'.

With the preparation made in this section, the total number of instances considered was reduced by 7184, from 30000 to 22816 instances. This dataset is unbalanced, with 17560 (0) no default instances, and with 5256 (1) default instances. The instances representing default (1) sum to 23.04% of all instances. A more detailed statistical description is presented in Appendix A.

Modelling The target feature is whether a client defaults (1) on the payment of his credit card or not (0). The pipeline used for this process is exactly the same as the one used for Experiment 1: German Credit.

Evaluation The results on training data (Table B.2) indicate that no overfitting occurred. This was surprising due to its existence in Experiment 1. However, the key difference in this experiment is the large number of instances that are available for the training of predictive models. This strengthens the argument that the German Credit dataset contains too few samples for predictive models such as Random Forest and XGBoost to properly train with.

While the accuracy in training is respectable, recall is below satisfactory with all models presenting results below 40% in this metric. This means that the models are not successful at finding clients who are expected to default (1) on their credit card payments.

Furthermore, the results obtained by Random Forest and XGBoost, while overall better when compared to the other four models, have their results overshadowed by the time taken in the process of hyper-parameter optimization. As shown in Table 3.13, Random Forest took 24 minutes to complete this process, and XGBoost took almost 50 minutes to find the best set of hyper-parameters.

TABLE 3.13. Running time in seconds

Model	**Hyper-parameter Optimization**	**Prediction**
DT	1.945	0.089
GNB	0.237	0.079
LR	9.551	0.128
MLP	261.696	1.127
RF	1447.536	2.251
XGB	2930.750	39.183

The results for test data are objectively worse than those obtained with training data. However, a dropoff in performance is expected, and such a decrease in obtained results is not large enough to consider the possibility of overfitting.

The argument made previously that the predictive models employed do not correctly predict cases where the user will default on their next credit payment still holds true, with extremely low results in the recall metric. These are unsatisfactory results, but as stated in this dissertation, the main contribution is to better understand black-box models through the application of XAI methods.

The six employed models provided similar results, though RF, and XGBoost all are slightly better when compared to the other four models. By taking into consideration these results, the time needed to execute both the hyper-parameter optimization as well as the time for prediction, it was decided to choose RF as the best predictive model out of the six and to provide explanations for this method instead of any other.

Unlike what was observed in Experiment 1, here the three most important features are ranked equally by PDP, SHAP, and LIME as shown in Table 3.15. 'Education' is deemed to be the most important feature, by far, to determine the target feature.

TABLE 3.14. Results on test data

Model	ROC AUC	Accuracy	F1-score	F1-score weighted	Precision	Recall
DT	62.24%	79.96%	40.31%	76.98%	64.28%	29.37%
GNB	65.43%	78.76%	46.90%	77.55%	55.30%	40.71%
LR	58.78%	78.95%	31.87%	74.72%	62.64%	21.37%
MLP	63.42%	80.53%	42.71%	77.77%	66.74%	31.69%
RF	63.39%	80.88%	42.72%	77.97%	68.95%	30.94%
XGB	62.96%	80.88%	41.74%	77.77%	70.00%	29.74%

For the remaining features, Table 3.15 indicates that all of them are related to the history of the client, whether that be bills to be paid or payments done in the last one to two months.

TABLE 3.15. Most important features ordered by ascending importance

Feature	PDP	SHAP	LIME
Past, monthly payment (-1)	0.035	-	-
Prev. payment in NT$ (-1)	0.04	-	-
Prev. payment in NT$ (-2)	-	0.02	0.01
Prev. payment in NT$ (-1)	-	0.01	0.01
Past, monthly bill (-2)	0.07	0.02	0.02
Given credit (NT$)	0.1	0.03	0.03
Age	0.175	0.04	0.04
Education	0.4	0.07	0.08

For counterfactual generated examples, the original instance had its target value set to 1, and the XAI models generated synthetic instances with the target value of 0. DiCE had more conclusive results, using the two most important features for the previously analysed XAI methods, and a few more. It had two to three feature values changed for each generated instance, and had several features that were transformed, from 'Given Credit' to features such as 'Past, monthly bill (-1)', and 'Prev. payment in NT$ (-1)'. Table 3.16 summarizes the results obtained through the implementation of DiCE.

TABLE 3.16. Changes made to feature values when applying DiCE as an explanatory technique.

Given credit (NT$)	Age	Past, monthly bill (-1)	Past, monthly bill (-2)	Prev. payment in NT$ (-1)	Prev. payment in NT$ (-2)
0.0	**1**	**0.2**	**0.1**	**0.2**	**2**
0.0	1	0.0	0.1	0.2	2.9
0.8	1	0.2	0.1	0.7	2
0.0	0.1	0.2	0.9	0.2	2

The first row represents the original instance

PermuteAttack provided a more direct approach to the problem by generating only a singular instance. This instance only had a change in feature value for 'Education', from the original value of 1 to 0.5, and with 'Education' being a categorical feature where the order matters, and with the highest value meaning lower education, meaning a change in the education level of the client leads to a worse prediction in regard to its risk. Overall, PermuteAttack's instance goes in line with what was previously seen with the feature importance given by PDP, SHAP, and LIME. Not only that, this is a more direct approach, and more easily comprehensible.

CHAPTER 4

Explaining Project's Cancellation Prediction

One of the main goals of this dissertation is to test XAI methods on models trained on real-world data provided by IAPMEI. While the previous examples use public and well-known data, IAPMEI's dataset is large and spans over 30 different tables and over 2300 attributes. Given its complexity, it is not expected to use all attributes as features, unlike the previous experiments. As such, the difference between attributes, and features needs to be clarified. In this dissertation, the term attribute is used when referring to the initial data. Feature is used when referring to an attribute that underwent treatment or any type of processing, and is expected to be used in the final dataset. This distinction helps us understand the scope of the dataset in relation to what was determined to be relevant to extract and use for this dissertation.

4.1. Business Understanding

IAPMEI is a public institute with the purpose of providing assistance to micro, small, and medium companies through incentives, and subsidies. The institute then has the task of supervising the whole process of applications to then approve them as projects. As seen in Fig. 4.1, this process begins with the analysis of applications, the decision to allocate financial support as well as verifying incentives, and the identification of irregularities. These irregular situations can lead to the cancellation of a project, with the restitution of amounts spent in the project, or can lead to budget or temporal detours. These detours negatively affect the efficiency and efficacy of how IAPMEI conducts its operations, either through the poor utilization of EU funds or by negatively affecting the growth of the economy. As stated in the introduction, the dataset used was provided by IAPMEI in the context of MAIPro Project, which has the challenge of predicting, among other objectives that are not worked on in this dissertation, possible project non-compliance.

It is necessary to clarify what each step represents, not only to understand the competencies of the Institute but to analyze which of these are important for the later sections. Table 4.1 describes what each step means while Figure 4.1 provides a visual illustration of the project life cycle:

For the purpose of this dissertation, only the projects that were successfully concluded or were canceled will be utilized. Specifically for the canceled projects, only those in which the type of cancellation was "Cancelled after contract" were considered. This decision was made due to business rules, where only those projects where IAPMEI is expected to take action are selected. There are two other types of cancellations that do not verify this rule: "Cancellation due to expiration" (147 projects) and "Waived by the company" (94

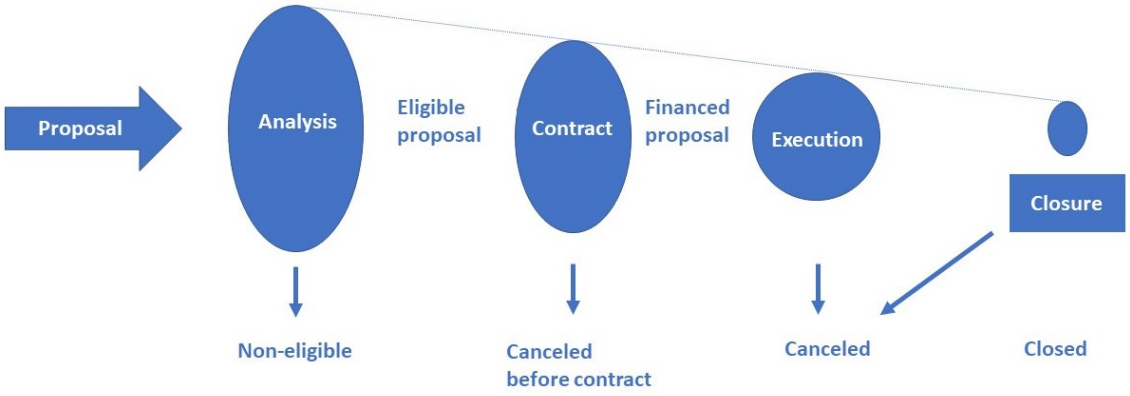


FIGURE 4.1. Project life cycle

TABLE 4.1. Definition of each step in a projects life cycle

Main step	Sub-step	Explanation
Analysis	-	Analysis of the application with the decision of attributing financial support
	Eligible proposal	The analysis resulted in a positive outcome, and the application can proceed to contract signature
	Non-eligible	The analysis concluded that the application should be terminated
Contract	-	Contract signature with the first payment
	Financed proposal	No issues detected with the project, resulting in its successful financing
	Cancelled before contract	IAPMEI or the applying company determined that there were issues regarding the project, resulting in its cancellation
Execution	-	Supervision of the project up until its closure
	Project financed	Successful financing of the project
	Cancelled	IAPMEI or the applying company determined that there were issues regarding the project, resulting in its cancellation
Closure	-	Closure of the project
	Closed with / without detours	The project closed successfully, regardless of whether budget detours were made
	Cancelled after contract	The project was canceled after the completion of the contract

projects). The same business rule was applied for the motive of cancellation, resulting in the exclusion of projects which has its motivation for cancellation as "Waived by the company". The causes that led to project cancellation with such motivation are probably diverse in nature, making it more difficult for an ML model to predict the target feature. In total, 272 canceled projects were excluded.

The filtering of the records on such criteria means that we are selecting the worst offenders when it comes to the problem of effective and efficient allocation of resources,

as the cancellation of a project that has concluded means that resources were held during the entirety of the project's life cycle. By only selecting the records which IAPMEI could analyse, this dissertation can help determine which factors are deemed most important for the cancellation of a project.

4.2. Data Understanding

In order to determine how predictive models behave, a careful decision on what data to use is necessary. The dataset provided by the institute has been worked on for over two years, in two different projects: the first where the main focus was extracting the data from eXtensible Markup Language (xml) or Excel files (xlsx), resulting in a more cohesive data structure and with the final goal of predicting project cancellation, and an attempt at predicting a project's ineligibility regarding expenses; the second project was a continuation of the previous, and this project aimed to predict several target features such as the ineligibility of a proposal, the cancellation of a project, and two target features for budget, and temporal deviation from the initially proposed values.

During this second project datasets were updated, which resulted in the extraction of more attributes from the original data. This dissertation is associated with the final stage of the second project and will use only the most recent extraction, starting from the tables already processed into Comma Separated Values (csv). This dataset contains information on the proposal and its management from 2014 to 2021.

There is a broad range of information made available by this dataset, where it is possible to discern data about the proposal itself, the analysis of the project when in execution, and after closure, as well as data related to the financial status of the company. All non-public data was provided by IAPMEI under a strict non-disclosure agreement, whose protocols were respected throughout the development although they are transparent to this description.

Data was also obtained from public sources in order to establish social-economic and global attributes. This data was mainly obtained through the Portuguese National Statistics Institute[1] (*Instituto Nacional da Estatística*, INE). Additional data was obtained through IES (Informação Empresarial Simplificada), and this data provides information on the financial status of the company. Table 4.2 provides a summary that also allows a glimpse at the scale of the dataset where the data is grouped by the step in the project life cycle, as illustrated in Figure 4.1, or where it was obtained from, for the features obtained from IES and INE. A summary description of what type of information it relays, the number of tables present in this group, and finally, the number of attributes present in all of these files.

Although the number of available characteristics in this dataset is large, only a few attributes were selected for this work. Several characteristics that contained gathered information at different times were excluded from this dissertation as these posed a risk

[1] `https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_main`

Table 4.2. Structure of the dataset

Main group	Description	Number of tables	Number of attributes
Proposals	Demographic data on the company, its consultants, the proposal, and the project	27	453
Analysis of the proposal	Data regarding the analysis of the proposal	1	514
Analysis of the closure	Data regarding the analysis of the closure of the proposal	2	430
Request for payments	Data regarding the request of payments	1	134
Expenditure	Expected expenditure at the time of the proposal	3	342
IES	Financial data on the company, consultant or supplier	2	303
INE	Social, and economic data for each location at level 2 in NUTS nomenclature	8	157
Status of the project	Data on whether the project is closed or if it has been canceled after closure	2	23
	Total	**46**	**2356**

to the results obtained, as there could be inconsistencies between different parts of these features but also data leakage. Another reason was the focus on the problem at hand: in an effort to reduce noise in the selection of features a selection of a smaller set of features that are deemed important was made. Features were also excluded due to technical reasons. One other reason led to a decision being made to exclude textual characteristics of an application for this dissertation, attributes that were empty in more than 5% of records were also excluded. Overall, the selection was based on the acquired experience of previous work in the projects, resulting in the selection of a small set of features. Still, the fact that this dataset has been previously curated does not mean that the optimization of what variables to use is strictly defined. Revisiting these choices is a continuous process. There is also an issue with the quantity of data, which will be described below.

For this work, the number of attributes used is drastically inferior to what was previously shown in Table 4.2. One of the reasons for this is to guarantee the applicability of results when delivering a final proposal, as the inclusion of too many features leads to an increase in the time necessary to train models, and there is a large tendency for overfit to happen due to the relatively small number of projects present in this dataset (Fig. 4.2). Another reason for this was the aforementioned understanding of the dataset and the fact that most of the information necessary for the prediction of a project's cancellation comes from the first proposal, without much input from the steps afterward. To better illustrate this is Table 4.3 which shows only the set of files/attributes used in this dissertation, and

where is possible to observe the lack of input from attributes related to IAPMEI's analysis of the proposal, the analysis of the project's closure, and the data related to requests for payments:

TABLE 4.3. Attributes used for the dissertation

Main group	**Number of tables**	**Attributes used to build the dataset**
Proposals	4	47
Analysis of the proposal	0	0
Analysis of the closure	0	0
Request for payments	0	0
Expenditure	1	12
IES	2	24
INE	8	146
After merge of the dataset	2	3
Total	**17**	**232**

The number of attributes/features represented above includes the total of attributes used even if they are not present in the final dataset. For instance, 24 attributes were used from the IES group (financial information on the company) but only 8 features were carried over to the final dataset. Another example is the usage of 146 attributes from INE. The data extracted from this source was organized into several attributes, one of which was NUTS II, and the rest of the attributes were either the year in question or the month. In reality, only a quarter of the number of features presented in the table above were used for the predictive and explanatory process. The final number of features present in the dataset after the data preparation is 39.

It is not only necessary to define what features were used in this dissertation, but also the quantity of data used. The majority of projects were removed as stated previously in this chapter, with only the projects that were closed successfully, and those that were canceled after closure with a motive other than "Waived by the company" being considered. This resulted in the filtering of the initial 6795 projects, resulting in a final list of projects containing only 1100 records. Figure 4.2 represents the breakdown of the initial records, from the universe of 6795 projects and the filters of interest that were applied in order to reduce the number of projects to only those of interest.

As for the number of projects used in this dissertation, the justification for the removal of the vast majority of projects comes from previously mentioned business rules. While arguments could be made for the validation of this experiment if the original number of instances were used, the significant reduction of the number of instances helps fixate the task at hand, that of increasing IAPMEI's efficiency, and efficacy, in regards to how to best manage its projects. The staggering reduction in the number of attributes used also illustrates this and helps demonstrate the fact that a researcher must be selective with the information, know what data is indeed useful, and differentiate from that which is

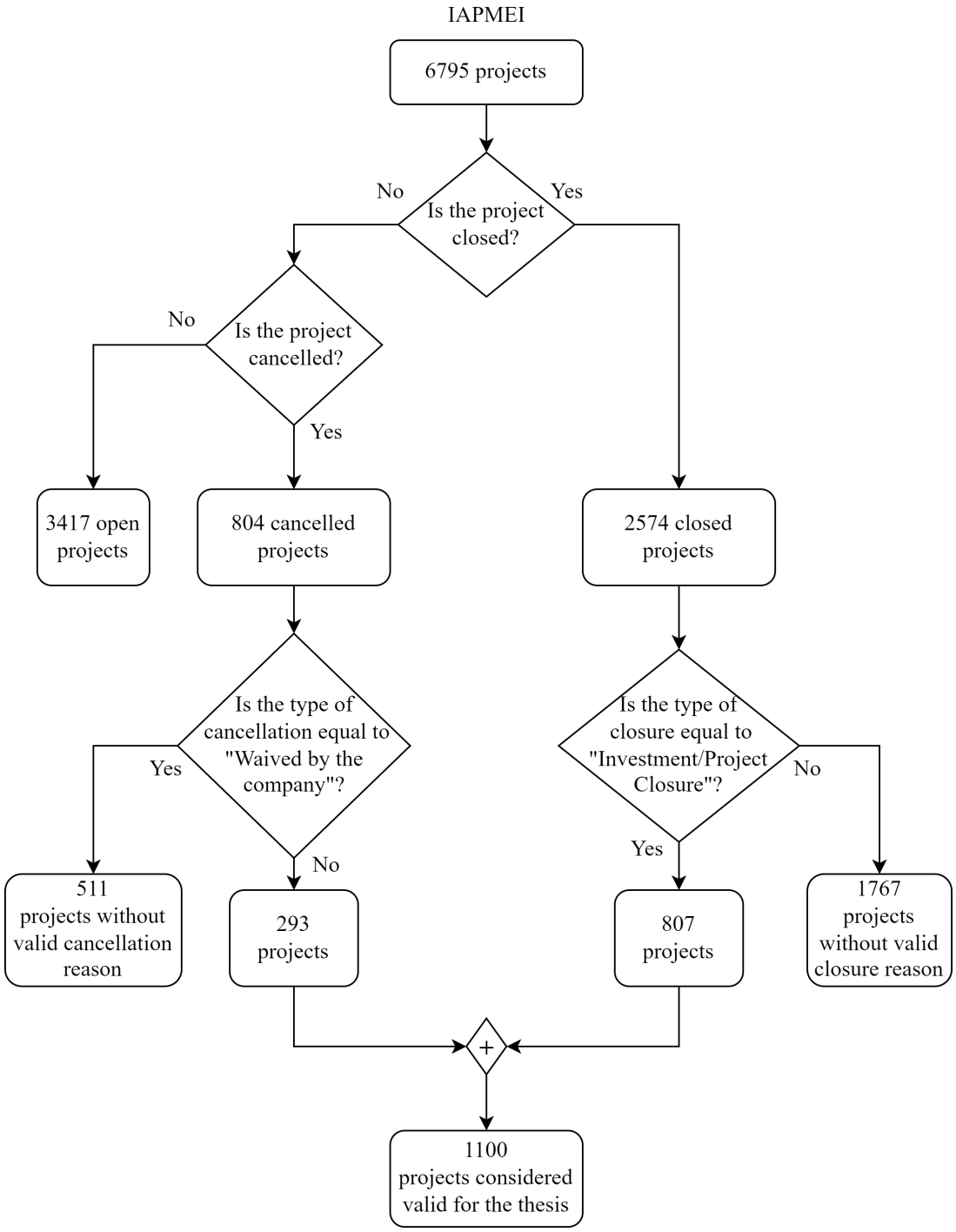


FIGURE 4.2. Filtering of the initial projects

not by applying knowledge of both what the data communicates as well as business logic and rules.

From the 1100 projects considered for this dissertation, 24 of them were made by companies whose size is neither micro, small, or medium. While these projects were valid given the business rules in place for filtering data (Fig. 4.2), IAPMEI is an entity whose primary focus is on smaller-sized companies, and these 24 projects are an exception to

the normal operations of the institute. As such, it was decided to exclude these projects from the final dataset.

Before the section on data preparation, a general analysis of the dataset was made, not only to understand the data that is being used and experimented with but also to help shape experiments in the modeling section.

When analyzing the distribution of projects by their companies' respective code of economic activity (CAE) through Fig. 4.3 and Table C.13, the vast majority are within manufacturing industries, representing 91.91% of the 1076 projects considered for this dissertation. Projects that do not fall in this category are either the other four categories, with the remaining 25 being placed in the "Others" category as their companies were in much less frequent activities.

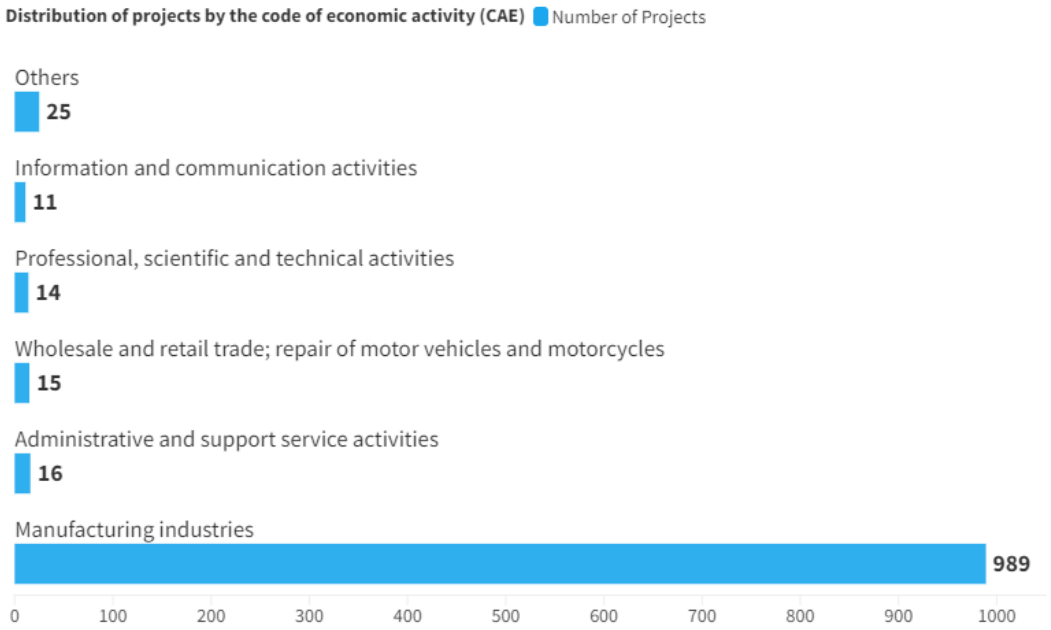


FIGURE 4.3. Distribution of projects by the code of economic activity

As shown in Fig. 4.4, for the year of application of a project, most projects that were created between 2014 and 2017 were successfully closed, with a peak in closed projects in 2015. From 2015 onwards a steady decline in closed projects was observed, with no projects created from 2018 to 2021. As for canceled projects, the same trend was seen, though all projects from 2018 to 2021 were canceled.

Regarding the total investment (Fig. 4.5) made to a project by the geographical location of the applying company, the region "Centro" is closely followed by "Norte", and is the most invested NUTS II regions. The autonomous regions of Azores and Madeira are not contemplated by IAPMEI's activity, with each region having its own entity for the management of projects of this nature.

It is also important to analyse the size of applying companies, the respective number of projects by the size of the company as well as the total investment made by their size. This analysis is present in Fig. 4.6, and by also identifying projects that were closed or

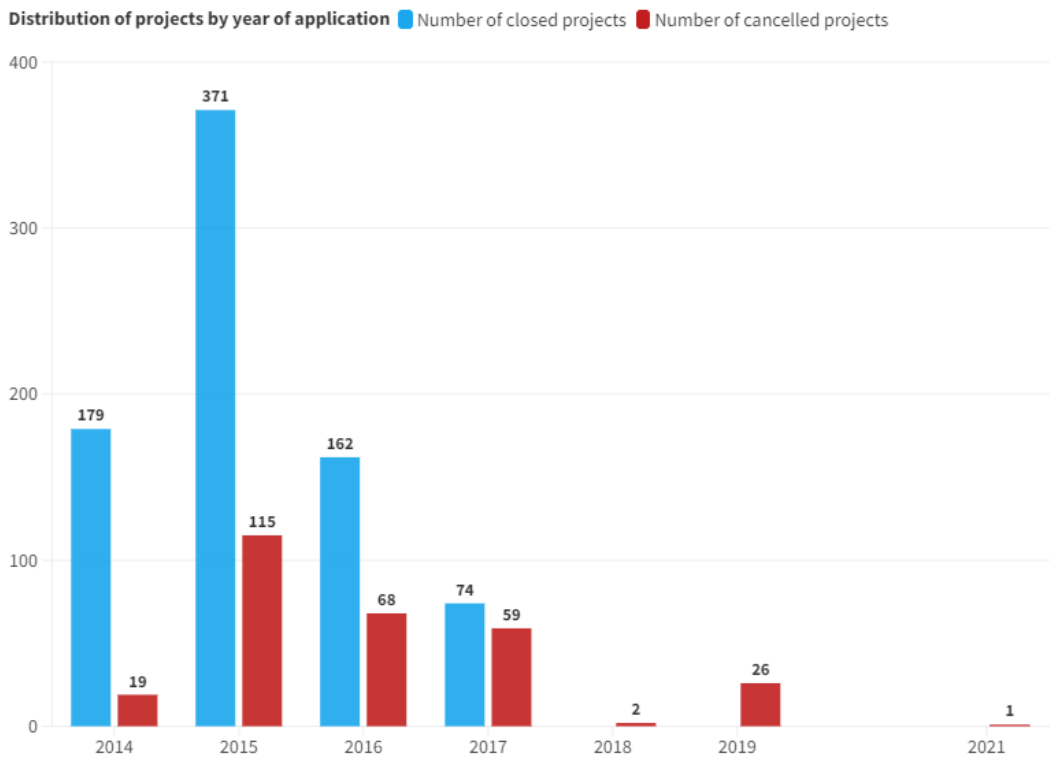


FIGURE 4.4. Distribution of projects by year of the proposal

canceled it was possible to identify a problem within Micro companies: companies of this size see their projects being canceled much more often than small and medium companies, with a rate of project cancellation of 51.53%. Even for small companies the rate of project cancellation is much higher than for medium-sized companies, with the first being 24.38% and the second being 14.99% respectively.

An analysis of the total investment for a project given the applying company's size is presented in Fig. 4.7 which further supports the argument related to micro-companies: Not only do they have a higher rate of project cancellations when compared to the other companies, but they also have a higher rate of investment for projects that were canceled, with 65.11% of the total investment for micro companies being canceled. This figure is relatively lower for small companies, with 30.81% of the total investment being canceled, and even lower for medium companies, where this rate represents just 15.68% of investment in canceled projects.

The correlation matrix for all features was also made and is present in Appendix C though only the features that had stronger relations are presented here. These features are mainly those related to financial indicators but also include those that provide information on the company size. Intuitively, smaller companies such as micro-enterprises and small companies tend to have lower financial indicators, and this is well represented with the correlation matrix shown in Fig. 4.8, where micro and small companies are negatively correlated to such features, and with medium companies being positively correlated to the financial indicators, and number of workers.

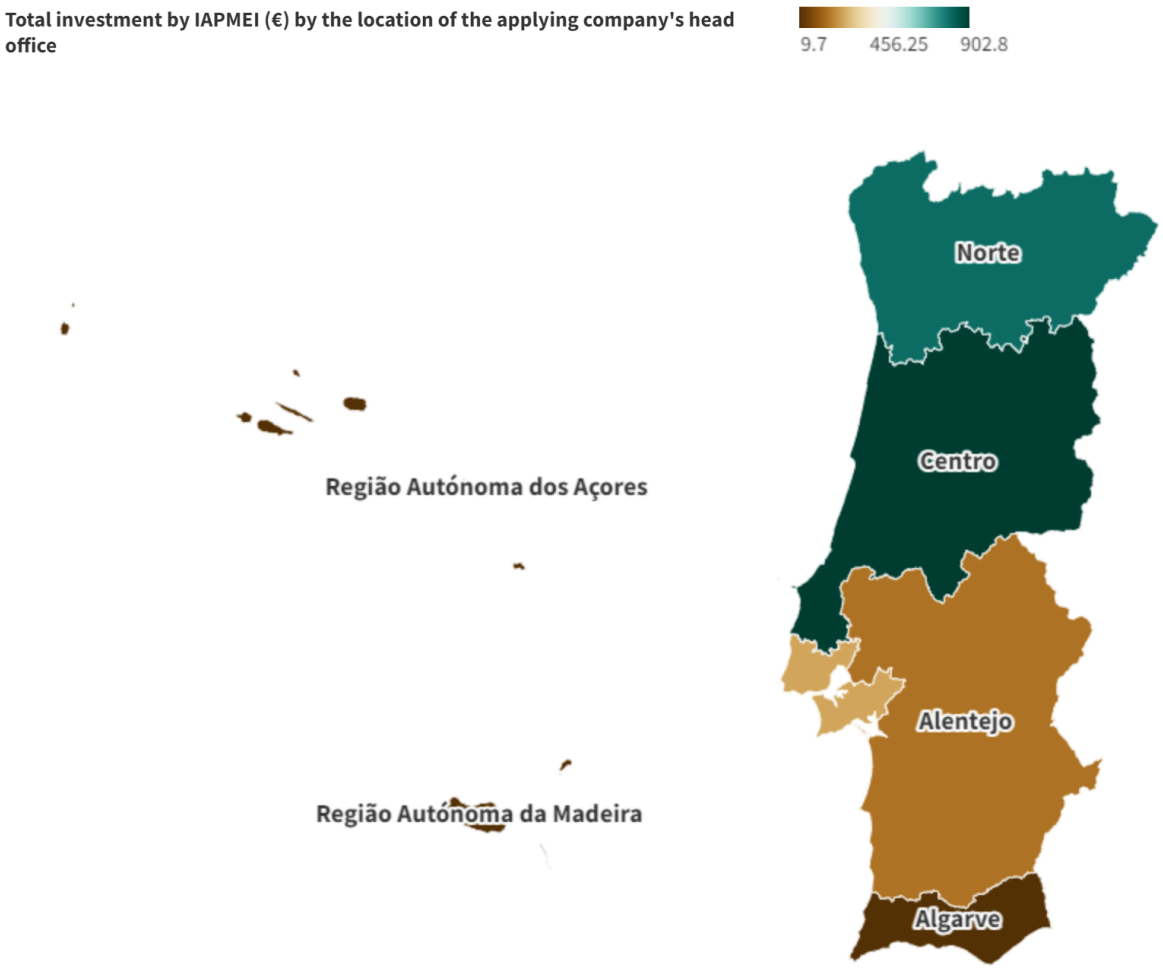


FIGURE 4.5. Total investment by IAPMEI (€) by location of the applying company's head office

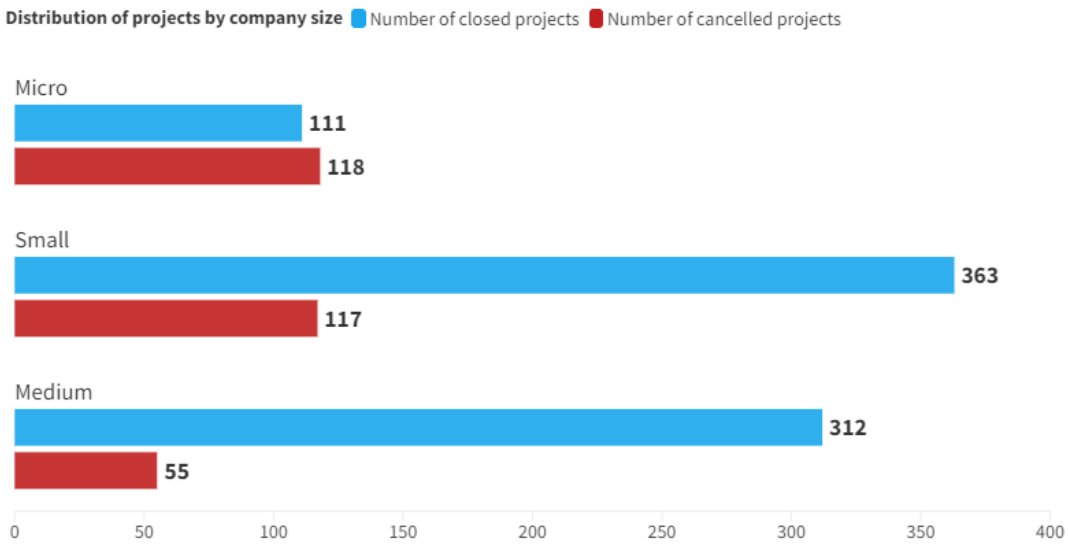


FIGURE 4.6. Distribution of projects by the size of the applying company

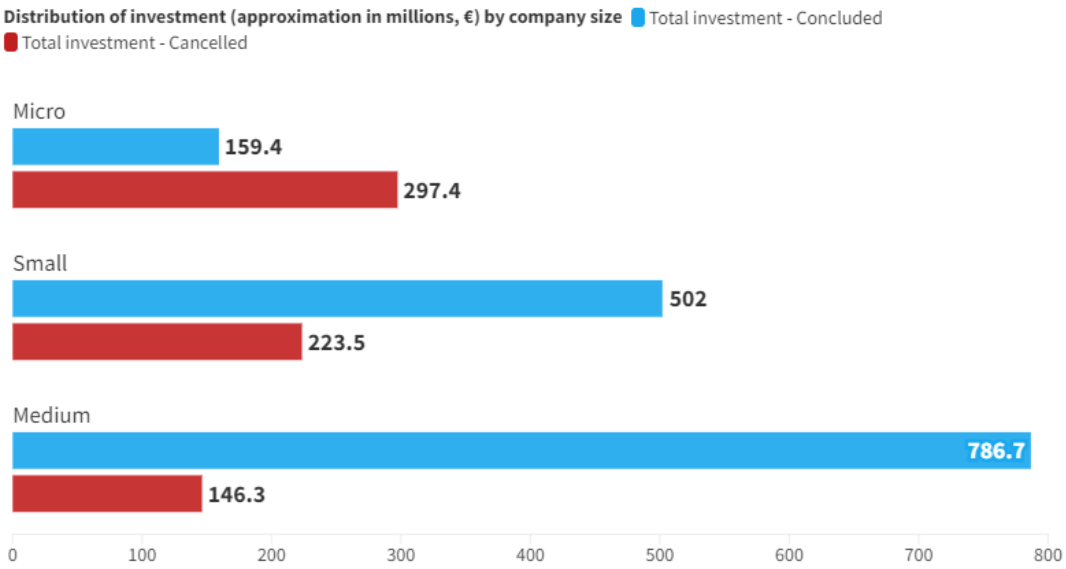


FIGURE 4.7. Distribution of investment, approximated in €, by the size of the applying company

Another affirmation is that smaller companies are generally younger, in comparison to larger companies, and this holds through the analysis of the correlation between the size of the company with the feature Young company, which is a binary feature that is 1 if the company is younger than four years, and 0 otherwise. The number of workers also reflects, indirectly, the size of the company, being negatively correlated to the feature that represents micro-enterprises, and with such correlation increasing the larger the size of the company.

Finally, from the correlation of these features with the target feature it is possible to conclude that for the most part, the features regarding financial indicators are negatively correlated to the target feature. With the target feature representing 1 for canceled projects and 0 for projects that were not canceled, this indirectly means that companies with larger turnover, assets, and net profit or loss have lower rates of project cancelation. With the analysis made previously on the size of the company, this is verified, with micro-companies being positively correlated with the target feature, and with such correlation decreasing the larger the company in question becomes.

Regarding more informative correlation matrixes, one is presented in Appendix C for the features related to INE. These features largely represent socio-economic information for each NUTS II region, and with the exception of the feature 'Poverty rate NUTS II', every feature from this group is highly correlated with each other. The section below describes in more detail how each feature was made, and its purpose.

4.3. Data Preparation

The step of data preparation is crucial to guarantee not only the high quality of the data but also that the predictive models are working with the correct data. Thus, it was necessary to build a working dataset with 228 relevant attributes. Since the list of

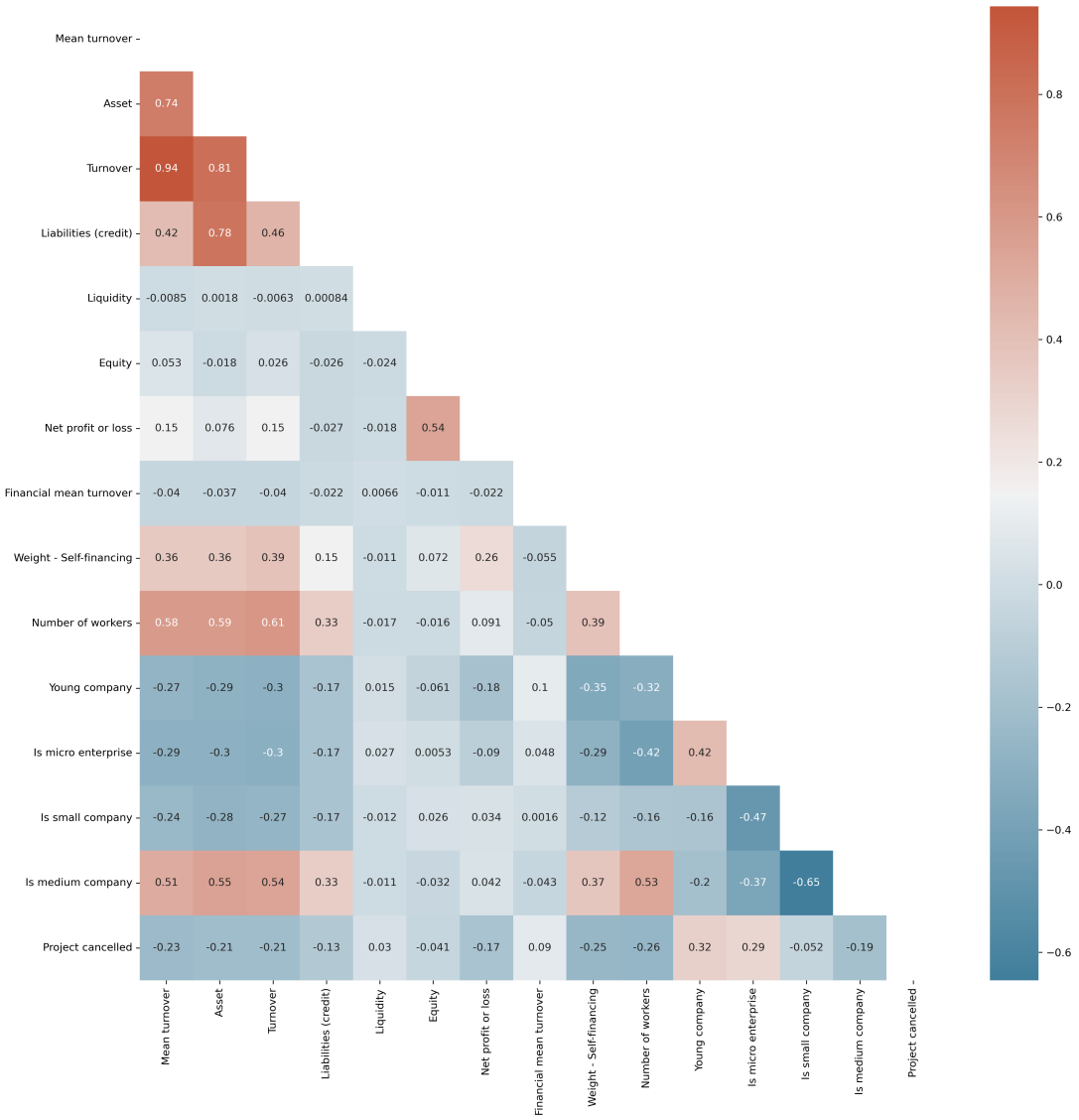


FIGURE 4.8. Correlation matrix for the features related to IES

total attributes used is long, and to avoid cluttering, only the features used in the final dataset will be described in this section, while the complete list of attributes used, as well as features created in order to construct the dataset, is available in Appendix C. The order of presentation follows the order presented in Table 4.3. For each grouping the explanation of how the feature was implemented and respective treatment performed is detailed.

4.3.1. Proposals

Being the core of the application towards the project, this grouping represents all the information related to the project's proposal submitted to IAPMEI. The 47 attributes contain demographic information (e.g., when and where the company was formed) and also a summary of the company's economic information, such as the Economic Code of Activity (CAE). However, other important information can be found in other data

tables that was unfeasible to be grouped within the main table of proposals (from the expenditure predicted in the proposal to information on the company).

Table 4.4 describes the set of features used to buid the working dataset. By order of appearance, the first feature is 'Young Company', with the purpose of determining whether the company was created at most four years before the proposal. For this feature, three attributes were used: the first two are related to the commencement of the activity, and the third represents the date when the company was started. As there were cases raising inconsistencies regarding which date was the correct one for this feature, the earliest date was chosen. Afterwards, this date was compared with the date of the submission of the proposal to verify whether the company had been created until four years before its project proposal.

The other four features represent the dimension of the company ('Is micro enterprise', 'Is small company', 'Is medium company', and 'Is non-SME'), and were created from a categorical attribute that represented the dimension of the company. The last feature of the group of proposals is a binary feature which determines whether the investment is made in the same NUTS II as the head office ('NUTS II of Project = NUTS II of head office'). For this, it was necessary to map the district of the head office and the project with its corresponding NUTS II. This correspondence resulted in two attributes: one for the NUTS II of the head office and another for the NUTS II of the proposal, and the final feature was created by verifying whether these two attributes are equal or not.

Regarding the expenditure of the proposal, several features were created. The first seven ('Weight - Equity', 'Weight - Self-financing', 'Weight - Foreign capital', 'Weight - Partners', 'Weight - Total incentive', 'Weight - Reimbursable incentive', and 'Eligibility(%)') were created by applying a filter to each row. The filter varies from feature to feature and has the purpose of selecting the rows of interest for the feature creation. Finally, each feature is then created by adding nine original attributes representing a parcelled expenditure. Table C.1 defines what headings are used for each one of these seven features. Additionally, for these features, a validation feature was added ('Errors in weights') which has the value of 1 if any of the weights of the expenditure exceeds the value of 1.

For the final two features, the first ('Number of workers') was extracted as-is, and the second ('Value of training') was created by filtering the table of expenses on external services by a specific rule, where each row was grouped by the project identifier, and summed.

4.3.2. Expenditure

All of the five features present in Table 4.5 were created by filtering for a specific class of expense and having its respective value added. The resulting auxiliary feature represents the total expense per project and per type of expense and it was then divided by the total investment of the project in order to create the final features present in Table 4.5.

TABLE 4.4. Profiling of the features created from the proposal information.

Feature	Number of projects filled	Mean	Standard Deviation
Investment (€)	1100	2025738	2786257
Young company	1098	0.21	0.41
NUTS II of Project = NUTS II of head office	1100	0.99	0.11
Value of training	1100	1101.13	7770.67
Number of workers	1100	11.06	5.81
Weight - Equity	1100	0.15	0.64
Weight - Self-financing	1100	0.18	0.20
Weight - Foreign capital	1100	0.11	0.16
Weight - Partners	1100	0.13	0.15
Weight - Total incentive	1100	0.59	0.13
Weight - Reimbursable incentive	1100	0.05	0.14
Eligibility (%)	1100	0.96	0.10
Errors in weights	1100	0.01	0.09
Is micro enterprise	1100	0.21	0.40
Is small company	1100	0.44	0.50
Is medium company	1100	0.33	0.47
Is non-SME	1100	0.02	0.15

TABLE 4.5. Profiling of the features created from the expenditure.

Feature	Number of projects filled	Mean	Standard Deviation
Expenses - Civil construction	1028	0.14	0.14
Expenses - Engineering services	1028	0.02	0.03
Expenses - Equipment	1028	0.73	0.19
Expenses - IT	1028	0.04	0.09
Expenses - Other	1028	0.07	0.11

4.3.3. IES

The financial indicators were created from the information provided by IES. All of the features present in Table 4.6 were constructed from several different attributes. These features follow the correct business rules regarding financial indicators, but as IES is a different system than the one used by IAPMEI, some data processing was necessary. For instance, all of the relevant attributes had their null values filled with 0, and attributes with negative values were replaced with 0 in order to maximize the filling of records, and the validity of the information present in this table. Furthermore, these indicators were mostly created using the year prior to the proposal, but there might have been the case that, for some external reason, did not have information for this year. This was corrected with a broader approach, where the information of the closest year was chosen (e.g. if the

financial information for the year before does not exist but if it exists for two years prior, then that information was selected).

As shown in Table 4.6, there are some projects without information regarding IES. These values were kept and treated with the processing mentioned above. There is also an issue regarding three features: 'Liquidity'; 'Equity'; and 'Financial mean turnover'. These features have infinite (inf) as their respective mean. This does not signify that all values are filled with "inf" but that some projects that have such values skew the distribution of the remaining projects. A decision was made to include such projects but to replace "inf" with 0, as these problematic values were due to the fact that, for one reason or another, it was not possible to properly calculate the financial indicators for the company.

TABLE 4.6. Profiling of the features created from IES.

Feature	**Number of projects filled**	**Mean**	**Standard Deviation**
Mean turnover	1093	5345291	8116707
Asset	1093	4286591	6758712
Turnover	1093	3685638	6009630
Liabilities (credit)	1093	607762.9	1628532
Liquidity	1062	inf	
Equity	1021	inf	
Net profit or loss	1090	0.02	0.04
Financial mean turnover	1093	inf	

4.3.4. INE

The National Institute of Statistics, or INE, is a public institute with the main purpose of providing official statistical information in an effective and efficient manner (Decree-Law No. 136/2012, 2nd of July). The data provided by INE is guaranteed to be anonymous and publicly accessible.

For this dissertation, eight features were used as shown in Table 4.7, mostly comprising social and economic information per NUTS II. This information is obtained externally from INE's website and requires some treatment as these indicators were created from yearly median but some indicators did not have the information available on a yearly basis, but on a monthly basis. This makes the number of attributes much higher than the number of features actually used as seen when comparing the number of attributes used in 4.3 and the resulting features present in Table 4.7.

In total, a common time interval was considered with the purpose of ensuring cohesiveness between the periods of these indicators and the projects, and also between each indicator. The best fit for this period spans from 2014 up to 2021, but as shown in Table 4.7 there are some in which not all years were available. Generally speaking, on INE's website the user can extract exactly what indicators are needed, and there is also the possibility to somewhat organize how the final output is displayed, but one indicator in

particular posed the issue of not having information available on a yearly basis, as mentioned in the beginning of this group. This feature is 'Company closure NUTS', where the time period is correct but 96 attributes were used in order to extract the yearly median for this time period. In order to build this feature it was necessary to have the yearly sum of company closures, which was done through the sum of each pair of year, and month. Afterward, and with the calculated sum, it was possible to extract the yearly mean.

Table 4.7. Profiling of the features created from INE.

Feature	**Number of projects filled**	**Mean**	**Standard Deviation**
Gini Index NUTS	1100	30.99	0.71
Company closure NUTS	1100	5265.50	1961.13
Population density NUTS	1100	770.29	435.18
College Network NUTS	1100	73.44	26.016
Number of SME NUTS	1100	338655.70	103553.80
Mean salary NUTS	1100	896.98	93.85
Unemployment rate NUTS	1100	8.43	1.37
Poverty rate NUTS	1100	17.35	1.73

4.3.5. After merge of the dataset

There are some features that had to be created only after the merge of all features present in the dataset. These features, along with a summary profiling are presented in Table 4.8. The first is the target feature ('Project cancelled'), where the project closure or cancellation is present in two different files, representing projects that have closed or cancelled, respectively. Projects are considered closed (0) if present in the first file and canceled (1) if they are present in the latter. The two other features were created through attributes present in the proposal, the first which has the goal of generalizing how many cancellations there are for a specific code of economic activity in relation to the total of projects with such code, and the second with the purpose to quantify the cancellations in relation to the total number of projects with a given NUTS II.

Table 4.8. Profiling of the features created after the merge.

Feature	**Number of projects filled**	**Mean**	**Standard Deviation**
Project cancelled	1100	0.27	0.44
Historical frequency of cancellation CAE	1089	0.27	0.13
Historical frequency of cancellation NUTS	1100	0.27	0.11

4.3.6. Outliers and further filtering of the data

Outlier detection is a concern in any data science task, and is extremely useful in ensuring data quality. While this step was trialed for the dissertation, the final dataset saw no implementation of outlier detection, and removal, to any of its features. In the first iteration, the removal of outliers drastically reduced the number of projects, making it unfeasible to apply for this dataset. Another iteration saw its detection only on the features 'Investment (€)', and the 'Value of training', where outliers were detected but ultimately not removed, as not only the mean for both these features was relatively similar, but it saw the removal of false outliers, as all outliers detected for Value of training were due to the fact that this feature is filled mostly with 0.

As stated previously in Section 4.2, there are 24 projects that were removed even if they were considered to be valid samples. This exclusion is summarized in Table 4.9. Further, some projects were not listed in the table of the expenditure of the proposal. This means that the features representing the weights of the expenses cannot be calculated. This was solved through the identification of these records, which totalled 57 projects, and by replacing their respective lack of value with 0. If all of these projects were to be removed then the final dataset would contain 1019 projects as shown in Table 4.9, representing a reduction in the number of projects by 5.29%. In summary, only the projects of large companies were excluded, resulting in a final dataset with 1076 projects, representing a reduction of just 2.2% projects when compared to the initial 1100 projects.

TABLE 4.9. Total number of projects - breakdown by the size of the company.

Company size	Number of projects	Final number of projects
Microenterprise	229	229
Small company	480	480
Medium company	367	367
Non-SME	24	0
Total	**1100**	**1076**

Finally, the correct scaling of each feature is necessary to ensure that the predictive models do not give more importance to one feature over another. All features were scaled to the interval of [0, 1], though some that were the result of calculations such as those related to Weights of financial indicators were already scaled to this interval.

4.4. Modelling

Each step in the CRISP-DM methodology is iterative, and here is necessary to ensure the quality of tests performed and the correct implementation of the code structure. The code structure used for IAPMEI largely follows what was done in the previous experiments in Chapter 3, albeit with some differences due to the review of the code structure.

One of the key differences from the experiences done in Chapter 3 is the inclusion of several oversampling and undersampling methods. The dataset contains 786 closed (0)

projects and 290 cancelled (1) projects. The key reason for including the usage of such samplers was due to the class imbalance of the target feature, where the majority class represents 72.95% of the total projects. Oversampling methods create synthetic data of the minority class so that both classes are evenly distributed, while undersampling methods remove records from the majority class to reach this balanced distribution. In total, nine different samplers were chosen without any criteria and were used on the dataset, five of which are undersamplers, three being oversamplers, and two that employ a combination of both under and oversampling to reach an even distribution. There are other samplers available but they were not used here due to time constraints. An important note for the usage of such samplers is that they were only applied to the training data. This means that the predictive models are trained on both real and synthetic data, but validation is made solely on real data. This was made to evaluate the performance of models only on real data and increase the trust in obtained results.

The undersampling methods used are the following:

- Generic undersampling[2]
- Tomek[**42**]
- Cluster Centroids[3]
- Neighbourhood Clean[**30**]
- Nearest Neighbours[**46**]

The oversampling methods used are the following:

- Generic oversampling[4]
- SMOTE[**8**]
- ADASYN[**26**]

The methods which employ both under and oversampling are the following:

- SMOTETomek[**5**]
- SMOTEENN[**6**]

The code starts with the definition of the sampler method to use. Afterward, a search for the best set of hyper-parameters for each individual predictive model is made, and these are then used to create the proper model, which is fitted and used to predict test data. For the predictive process, five different seeds are used to improve the validity of the obtained results. These seeds were randomly selected and are the same for each differing model, and sampler. The definition of these interchangeable seeds makes the results reproducible and helps mitigate the inherent problem that comes with only using a single seed, that of a test split where the evaluation of the predictive models yields misleading results.

[2]Available and used from: `https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html`

[3]Available and used from: `https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.ClusterCentroids.html`

[4]Available and used from: `https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.RandomOverSampler.html`

To help organize how experiments were made, an experiment is considered as a run containing the results of one sampler, for the five different seeds and for the six predictive models, resulting in one experiment containing 30 different results. In total, 33 experiments were made, though experiments from 1 to 3, and from 6 to 9 were invalidated due to the usage of SMOTE in the test data. For this reason, these invalid results are not present in Table 4.10. Experiments from 17 to 27 were considered to be final. Experiments 28 to 33 were to validate results with those obtained in [**43**], albeit with a level of scrutiny as the conditions for A. Vila's experiments are different than those made here. Major differences include the usage of different features than those used here, and the usage of a different filtering process for the dataset than those used for this dissertation.

TABLE 4.10. Description of each experiment made on IAPMEI's dataset

Experiment	Description
0	Initial experiment with overfitting in training data
4-5	Experiments without over or undersampling methods tried to separate projects by their respective companies' size. Too few instances of data led to the exclusion of this experiment
10-12	Experiments with only SMOTE where instances are only created with training data. Equal proportion of projects by company size in train/test split. Results were not accepted due to their low value in the performance metrics
13-16	Manual hyperparameter tuning. Results were inconclusive
-	Issues were first detected when comparing obtained results with those obtained by A. Vilas' [**43**]. This experiment involved the prediction process using A. Vilas' pipeline.
17-27	Final set of experiments with the correct business logic in place. These experiments contain the results of all samplers used.
28-29	Further differences were found in this dissertation' and A. Vilas' [**43**] approach. These experiments contain A. Vilas' data and the random sampling used in previous experiments.
30-31	A. Vilas' [**43**] sampling split data in test and train based on the index of the project, not a random split like in this dissertation. For this experiment, A. Vilas' data and sampling were used.
32-33	Data was re-obtained to include the project index number. Afterward, A. Vilas' [**43**] sampling was used to split the data.

4.5. Evaluation

For the evaluation of the performance of employed predictive models, a general overview of obtained results is made. This overview contains the results of the experiments made with all samplers, and for the analysis of obtained results, a calculation of the mean of the results from the five different seeds is made. With this overview, it is possible to discern what experiment was most successful in predicting project cancellation. Afterward, further analysis is made on this specific experiment along with the explanations yielded by XAI methods.

In total, 33 experiments were made. Experiments 0-17 demonstrate the iterative process of building a pipeline for target prediction, where numerous problems were faced such as how to implement samplers, whether the size of the company mattered for the prediction, and experiments where manual optimization was tried. Another problem that makes the comparison of obtained results from these experiments with the results shown below is the fact that these experiments contained a prior version of the dataset, which had fewer projects due to a more restrictive filter, which excluded projects that had any feature with an empty value. These experiments are excluded from this dissertation as they were deemed not valid for analysis. The set of valid experiments, from experiments 17 to 27, are present in Appendix C.

While it is more important to predict whether a project was correctly classified as canceled, it is also important to acknowledge whether it was correctly classified as closed. For this purpose, the F1-score metric was chosen as the main comparator between experiments. An initial experiment was also done which had no implementation of any sampler, which serves as a baseline for comparing each experiment. All experiments had test and train data split in 30% and 70% respectively, resulting in Table 4.11:

Table 4.11. F1-score for test data for all samplers

Experiment	**DT**	**GNB**	**LR**	**MLP**	**RF**	**XGBoost**
None	42.3%	57.7%	41.6%	46.5%	61.9%	59.0%
Undersampling	56.9%	57.6%	61.6%	48.4%	66.2%	63.0%
Oversampling	49.1%	53.6%	62.4%	60.3%	62.2%	61.9%
Tomek	49.6%	57.4%	42.3%	59.8%	62.3%	62.3%
Smote	51.2%	52.6%	61.4%	60.5%	67.3%	66.1%
Adasyn	46.5%	51.0%	59.7%	59.8%	66.1%	63.8%
SmoteTomek	**53.1%**	**60.3%**	**66.9%**	**65.5%**	**68.3%**	**65.7%**
SmoteTeenn	52.3%	55.6%	58.5%	59.2%	63.9%	62.1%
Cluster	51.2%	56.4%	60.9%	59.1%	58.8%	57.9%
NeighbourhoodClean	55.5%	61.6%	62.0%	63.7%	67.2%	64.2%
NearestNeighbours	57.4%	51.8%	62.3%	62.1%	64.8%	64.2%

Generally, obtained results are satisfactory, at least when compared to [**43**]. One of the possible reasons for such a discrepancy regarding obtained results is the differences in how data was filtered and sampled. A. Vilas' approach contains more canceled projects, but fewer closed projects, than those present in this dataset. While A. Vilas [**43**] considered projects from 2015 up to 2019, a less restrictive timeline was made in this dissertation, which resulted in the inclusion of projects from 2014 up to 2021. Another difference is the features used, which impacts how predictive models perform to predict the target feature. Finally, the filter of projects also used the motive of cancellation, which was not used in the filtering process present in [**43**]. As such, while the obtained results are inferior to those obtained by Alberto, valid justifications were made to demonstrate key differences in both approaches.

Models such as Decision Trees and Gaussian Naive-Bayes had relatively worse predictive performance when compared to Random Forests and XGBoost, with Logistic Regression and Multi-Layer Perceptron falling in-between in terms of their respective performances. Generally, when comparing to the baseline the usage of any sampler saw an improvement on evaluation metrics, with the exception of the sampler Cluster with the predictive model RF, or the samplers Cluster, SMOTEENN, NearestNeighbours, ADASYN, and Oversampling, in which GNB came short when compared to the results obtained using the baseline. To further analyze obtained results, the experiment with the sampler SMOTETomek was chosen as it has, generally, the best results out of the other experiments.

With the best experiment chosen, it is necessary to detail the obtained results before proceeding with the explanations. This experiment, with the sampler SMOTETomek, saw both the addition of projects to the minority class, and also the removal of projects from the majority class. Before the implementation of the sampler there were 550 closures and 202 cancellations. After the implementation of SMOTETomek the number of projects changed to 543 closures and 543 cancellations. The test data did not see any transformation by the sampler, and it contains 236 closed projects, and 88 canceled projects.

While the addition of projects to the minority class is significant, with 341 synthetic samples being generated for the train data, only seven projects were removed from the majority class. This can be modified in the samplers parameters but for this dissertation, a decision was made to retain as many projects as possible.

An initial analysis of results from train data shows a problem with this experiment, for the models Random Forests and XGBoost: Overfitting. As Table 4.12 demonstrates, for the methods Random Forest and XGBoost all metrics are 100%, and with the dropoff in performance present in Table 4.11, the fact that overfitting exists is a certainty.

Table 4.12. Results for train data

Model name	**ROC AUC**	**Accuracy**	**F1-score**	**Precision**	**Recall**
DT	95.9%	95.9%	95.9%	95.7%	96.0%
GNB	73.8%	73.8%	73.2%	76.3%	72.1%
LR	82.1%	82.1%	81.7%	83.4%	80.1%
MLP	80.5%	80.5%	79.9%	82.2%	78.0%
RF	100.0%	100.0%	100.0%	100.0%	100.0%
XGB	100.0%	100.0%	100.0%	100.0%	100.0%

This is a recurring problem in the process of building predictive models, and this is due to, in this case, a complex set of hyper-parameters alongside the low number of samples present in training data. This justification comes from the analysis of the running time needed, represented by Table 4.13, where the time needed for hyper-parameter optimization suggests too much focus on Random Forest and XGBoost. However, it is

also important to note that these models have a wide array of possible hyper-parameters, which contribute directly to more time spent in the optimization process.

TABLE 4.13. Running time in seconds

Model name	Hyper-parameter tuning	Prediction
DT	0.57	0.02
GNB	0.07	0.02
LR	9.07	0.89
MLP	50.49	0.42
RF	383.00	0.37
XGB	644.18	0.18

The running time for both Random Forest and XGBoost far surpasses the time needed for the remaining models to predict and for the generation of explanations. This is mostly due to the usage of a wide array of possible hyper-parameters, which led to the existence of overfitting. Ultimately, it was decided to maintain this experiment as the best out of the others, as this issue is also present in the experiments with other samplers.

While the metric F1-score was used to find the most suited experiment for analysis, and the overall best predictive model, it does not take into account class imbalance. As seen in the previous section, there is a presence of a large degree of imbalance in the two classes. As such, it is necessary to use a metric that encompasses all metrics such as F1-score but also takes into consideration this imbalance. F1-score (weighted) helps mitigate this issue, but it needs to be analyzed critically as it ends up not being between precision and recall. Overall, it is possible to see that the performance of all models was better than previously shown, though their issues largely lie in the recall and precision, indicating issues in the prediction of a true positive (the project is canceled), and false negatives (the project that is canceled was predicted as not canceled).

TABLE 4.14. Results on test data

Model	ROC AUC	Accuracy	F1-score	F1-score weighted	Precision	Recall
DT	67.78%	71.85%	53.09%	72.59%	48.55%	58.86%
GNB	72.84%	74.14%	60.28%	74.69%	55.12%	70.00%
LR	78.15%	80.00%	66.86%	80.56%	61.05%	74.09%
MLP	77.07%	78.95%	65.51%	79.55%	60.07%	72.95%
RF	77.75%	83.89%	68.35%	83.53%	73.24%	64.32%
XGB	76.05%	82.65%	65.69%	82.21%	70.91%	61.59%

Analyzing the confusion matrix for XGBoost, Fig. 4.9 helps visualize the imbalance in the test data, with 88 samples being for the positive class and the remaining 236 being for the negative class.

Moving toward the explanations provided by XAI methods, the legitimacy of an explanation comes from whether differing XAI agree on how explanations are created. For

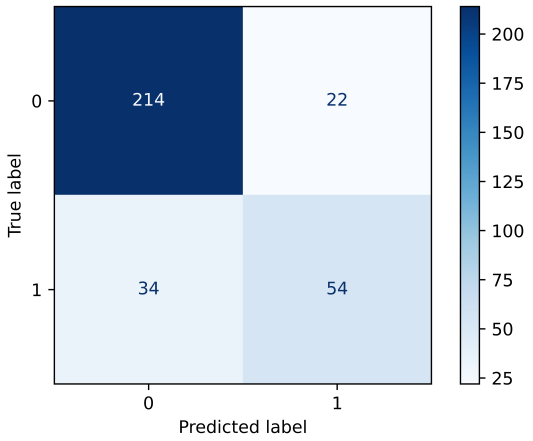


FIGURE 4.9. Confusion matrix of obtained results for XGBoost

feature importance techniques such as SHAP, this would mean that a feature is at the top of most important features, and for counterfactual models, this would be represented as a feature that was changed so that its prediction would be altered.

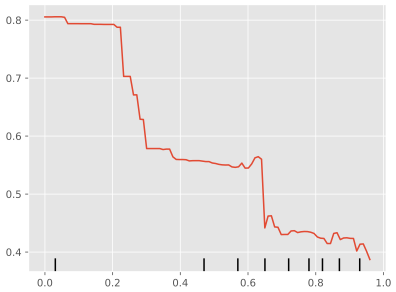
For the analysis of XAI methods, it was decided to emphasize obtained results with XGBoost, as although the performance of the model is slightly worse when compared to thos obtained with Random Forests, the running time for XGBoost is objectively better than those obtained with Random Forest, as shown with Table 4.15:

TABLE 4.15. Running time in seconds

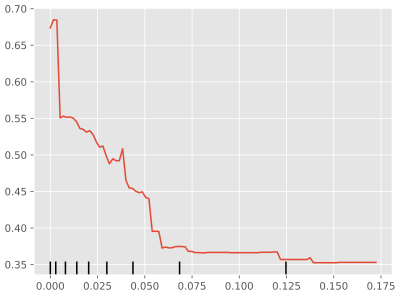
Model name	**DiCE**	**SHAP**	**LIME**	**PDP**	**PermuteAttack**
DT	0.58	31.71	0.08	9.87	3.67
GNB	0.51	62.09	0.04	11.19	15.27
LR	0.44	29.60	0.03	10.07	6.66
MLP	0.47	31.34	0.03	10.45	7.23
RF	1.04	384.98	0.09	104.67	315.52
XGB	0.62	71.96	0.04	18.10	15.90

Starting with an analysis of PDP, obtained results indicate six features that are most responsible for determining whether a project is canceled or not. Ordering from highest to lowest by their respective approximated difference, between minimum and maximum values, these are: 'Expenses - Equipment'; 'Mean turnover'; 'Historical frequency of cancellation CAE'; 'Asset'; 'Expenses - Civil construction'; and finally, 'Financial mean turnover'. In Fig. 4.10 the three most important features are summarized in the various generated partial dependence plots. In these plots it is possible to observe the effect a given feature (x axis) has on the value of the target feature (y axis).

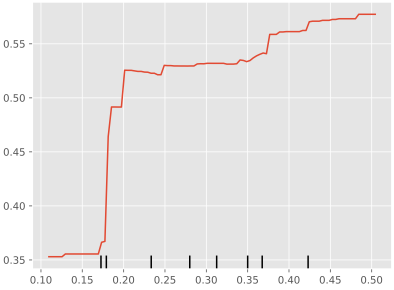
For the explanations provided by PDP, the feature 'Expenses - Equipment' is the one which has the most effect on the outcome of the target feature, followed by 'Mean turnover' and 'Historical frequency of cancellation CAE'. The remaining features also have a considerable impact on the target feature and will be used to verify their presence in the other XAI methods.



(A) PDP for Expenses - Equipment



(B) PDP for Mean turnover



(C) PDP for Historical frequency of cancellation CAE

FIGURE 4.10. Three features with the most impact for PDP after ordering by differences presented in the target feature.

For the method SHAP the same features as PDP were observed as having similar impact, though there is a slight change to the order of importance. Namely, features such as 'Asset' were deemed to have slightly more importance in PDP than SHAP, but its importance is negligible when compared to lower-ranking features, and the top three most important features. Fig. 4.11 shows obtained results for SHAP.

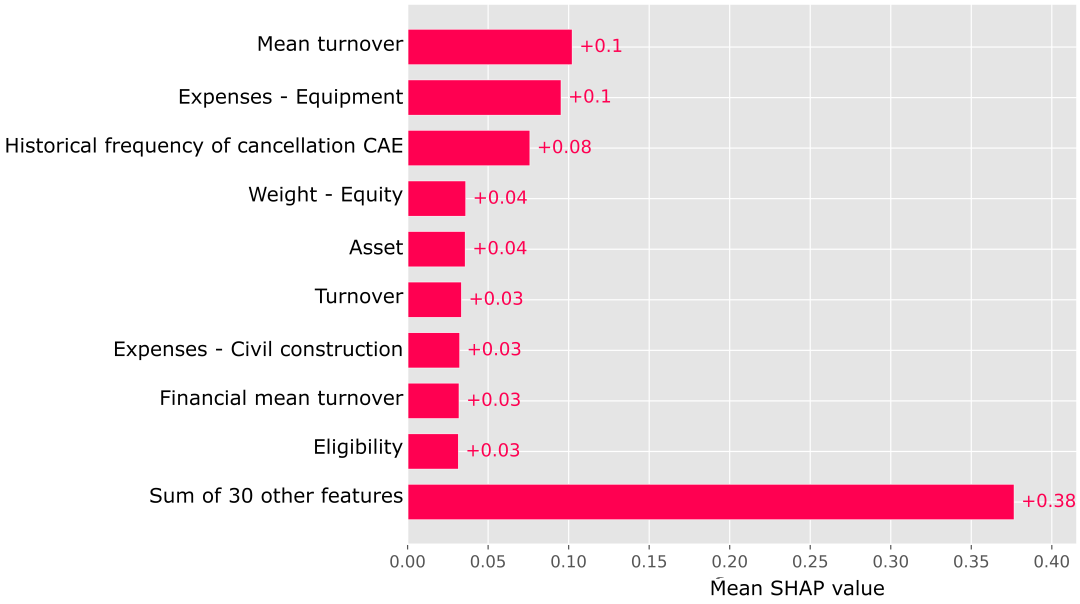


FIGURE 4.11. Summary of SHAP for the most important features.

For the analysis of LIME, DiCE, and PermuteAttack, a singular sample is used. There is an impediment to showing complete data from IAPMEI's dataset, and as such, only portions of obtained results are presented for the analysis of LIME and affected features along with their new values for the counterfactual methods (DiCE and PermuteAttack).

Overall, LIME found the same features as PDP and SHAP to be important for the predictive process, but this time, 'Net profit or loss' is one such feature instead of 'Expenses - Civil Construction'. For this sample, the third feature, 'Historical frequency of cancellation CAE', was calculated to have similar importance to the remaining features. A summary of the results presented by LIME is shown in Fig. 4.12

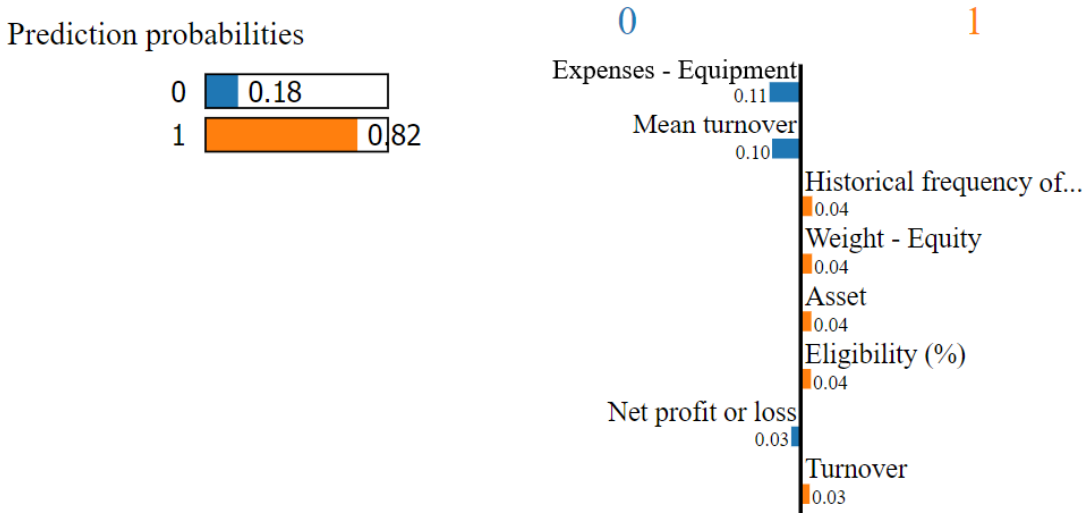


FIGURE 4.12. Summary of LIME for the most important features

It was possible to determine common factors to determine if a project would end up canceled. However, it is also necessary to add explanations on a local basis. This was done not only with LIME but also DiCE and PermuteAttack, where three and ten counterfactual instances originated, respectively. From these generated instances it was possible to determine what features were changed to alter the outcome. For DiCE, four different features were used to alter the outcome, with only one of them ('Expenses - Equipment') being present in the previous analysis of XAI methods such as PDP, SHAP, and LIME. However, in the generated examples, the feature 'Weight - Reimbursable incentive' had its value changed in two of the counterfactual examples, with a significant difference in values when compared to the original instance.

For PermuteAttack the generation of samples resulted in the transformation of just two features, one of which ('Expenses - Equipment') is present in the three previously analyzed methods, and the other being present only in LIME. The original instance had relatively low values in 'Expenses - Equipment', and almost all generated samples saw a drastic increase in the value of such features. This means that XAI methods considered that a company whose expenses in these services tend to end with its project closed rather than canceled, but might indicate

Table 4.16. Resulting samples generated by DiCE

Investment (€)	Expenses - Equipment	Weight - Total incentive	Weight - Reimbursable incentive
0.117	**0.200**	**0.608**	**0.132**
0.212	0.800	0.608	0.132
0.117	0.200	1.000	0.800
0.117	0.900	0.608	0.300

The first row in the table represents the original instance's features values.

Table 4.17. Resulting samples generated by PermuteAttack

Net profit or loss	Expenses - Equipment
0.018	**0.200**
0.505	0.823
0.018	0.830
0.018	0.930
0.018	0.700
0.018	0.960
0.018	0.920
0.018	0.910
0.018	0.836
0.018	0.870

The first row in the table represents the original instance's features values.

Before the analysis of the explanations obtained with the different XAI methods, it is necessary to clarify that the ratio of project cancelation increased over the years (Fig. 4.4), with all projects that started in 2018 and 2019 being canceled. This might be due to the COVID-19 pandemic which has had a tremendous impact on almost all sorts of economic activities. However, the percentage of projects where this might have happened is minimal at best (projects with the application year of 2018 or 2019 represent 2.6% of the total projects), with projects with the year of application running between 2014 and 2017 making up the vast majority of projects in the dataset. Unfortunately, it is not possible to provide the exact number of projects closed during the pandemic as the date for project closure/cancelation has not been provided.

The analysis made here is mostly speculative since no literature was found that reflects on the exact problem of project cancelation in the context of the usage of structural funds. As seen previously with the explanations given by PDP, for the feature 'Expenses - Equipment', the higher the expense in this area the lesser the probability of cancelation. This might be due to the fact that a higher expense in this type of investment makes the project more risky from the viewpoint of IAPMEI. A project with higher risk will, in turn, be much more carefully analyzed and monitored, resulting in projects of this magnitude only being approved for execution if the company that is undertaking it strongly

demonstrates the necessary capability to complete it. Generally, companies of larger size tend to run with these types of higher, more expensive projects, which might help explain the correlation seen in Fig. 4.6, where the larger companies have less canceled projects. Another feature deemed important is 'Mean turnover' which can be grouped with the previous feature as the size of the company is directly correlated to the financial indicators of that company, as observed in Fig. 4.8 where micro companies are negatively correlated to such indicators and with this correlation increasing the bigger the size of the company.

Given the correlation of financial indicators with the company size, in the majority of cases, companies of smaller stature tend to have lesser economic freedom than larger companies. As discussed in [**2**], smaller companies raise the hazard rate - the probability of failure conditional on survival to the age - in only the first four years of the companies' life. Even more crucial, it was observed that small companies showed a consistently higher hazard rate than larger companies. In another work, the authors observed that diversifying companies that are entering a new industry or market tend to be larger[**18**] and thus these show a higher rate of survival when compared to other types of companies that are entering such an industry. Another interesting observation was that, on average, the diversifying companies that survive are much larger in size than completely new companies that are entering the industry, in the long term. In relation to IAPMEI's dataset, Tables C.10-C.12 present the distribution of project cancelations by whether the company is young or not, and show that smaller companies pose a greater risk of project cancelation than larger companies. In connection with the previously mentioned studies, the smaller the company the higher the risk, in general. Moreover, young companies are generally riskier than larger ones. Finally, the authors of [**17**] analysed the manufacturing industry plants concluding that the company's experience at the time of entry is an important factor for determining a subsequent exit, with larger companies being more easily able to exit the industry by shifting the production line of a plant instead of closing it outright.

The last feature considered more important is 'Historical frequency of cancelation CAE'. Possible explanations for the importance of this feature include the existence of different survival rates for companies with different CAE, as different economic sectors have different barriers to entry. For example, the sector that includes restaurants has a lesser barrier of entry when compared to the economic sector of an oil rig. The former has comparatively low costs for the start of a company, while the latter has a much higher barrier to entry. For the economic sector of a restaurant, this results in a high rotation of companies, or in other words, company closures. For the sector of the oil rig, this ultimately means that fewer companies will operate in this activity, but the number of company closures or drop outs will also be lower. This rationale is supported by the author of [**21**] that state that the first entry of small-scale companies in an industry or market is a common occurrence, but also is their exit, In fact, it was observed that these companies, while able to enter the market, tend to have a relatively short life expectancy.

'Historical frequency of cancelation CAE' as a feature needs to be analysed thoroughly. There are many CAEs where a low number of projects exists, and the vast majority of companies whose projects IAPMEI has invested in seem to be associated to only one CAE, as shown in Table C.13. Nevertheless, this is due to the fact that a company may register and operate with many CAES but there is always a main one and that is the one used here. Given the data present in Table C.13, it is possible to observe the most crucial issue with this feature: more weight is given to cancelations of projects having CAEs with fewer projects than those displaying a higher number of projects. By the analysis of group C, manufacturing industries (Table C.13), the ratio of cancelation is 20%, while for group Q, human health and social support activities, the ratio is 60%. This observation makes the extraction of useful information from this feature somewhat dubious as CAEs with fewer projects have their cancelations have more importance than CAEs with more projects, which may not necessarily indicate that the sector is riskier, but rather that there simply aren't sufficient projects for that CAE to determine accurately its corresponding risk.

It is necessary to verify whether the statements made on the related literature, since they refer to other geographies, are applicable to Portugal. One specific study that worked on Portuguese companies' data in the context of the analysis of company performance after their entry into an industry [**35**]. The authors observed that smaller companies show the highest probability of exit. However, they found that, for the survivability of a company, the initial size of the company matters less than the current size of the company and the latter is what helps to determine the company's survival. In short, companies that start small but face fast growth after entry have a greater probability of survival.

In summary, the XAI methods applied for explainability of the models indicate as the two most important features to be 'Expenses - Equipment' and 'Mean turnover'. These features indirectly represent the monetary freedom a company has, and through the analysis of the partial dependence plots (in Fig. 4.10a and 4.10b), this is clearer for companies with either low expenses in this category or low mean turnover. These companies have a much higher probability of having their projects canceled than those with higher expenses and turnover. As seen for the counterfactual methods, almost all counterfactual examples generated saw a drastic increase in the value of the feature 'Expenses - Equipment'. This should be viewed critically, as it might not outright indicate that a company should have more expenses, but rather that it should have the possibility to do so, being of bigger size. It was possible to find support for both arguments, though there is a need of more studies in the specific context of this dissertation, that of usage of structural funds.

CHAPTER 5

Conclusions

The body of knowledge of XAI as a research area is still under consolidation, mostly due to the fact that it is still a recent research area and as such, suffers from a lack of generalized and formal definitions. While the definition of XAI is agreed upon, the same cannot be said for the classification of its methods. Furthermore, the implementation of XAI techniques does not follow a systematic process either, with some being readily available for implementation (SHAP) while others only present in their repository (PermuteAttack). These facts help to prevent a wider adoption of explainability in different applications. This dissertation intends to tackle both of the aforementioned points through the accompanying systematic literature review as well as with the experiments here described, performed using public and private (IAPMEI) datasets.

The biggest challenge faced was to learn what XAI stands for at the moment. In order to understand what are the relevant definitions and state of the art within this research area, a literature review was deemed necessary. However, given the requisites of this dissertation, it was quickly found that this search had to address two different points: firstly, a theoretical approach to XAI with the sole purpose of not only defining XAI but categorizing XAI methods as well. Secondly, a practical survey, through the search for practical implementations of XAI techniques related to the financial sector. This division led to the investigation of XAI by performing two systematic literature reviews, which have been synthesized into a singular scientific article. This work helped to understand not only XAI as a newly-formed research area but also gave an insight into what to expect from XAI techniques, as well as helping to find potential methodsfor usage in this dissertation.

Therefore, a selection of candidate model techniques has been made. This selection focused mainly on whether it was possible to use the method in a *Python* environment. In total, five XAI techniques were chosen and later used for the experiments. Unfortunately, although the implementation of the method Anchors was planned, it had to be discarded due to several technical difficulties. This dissertation differs slightly from a regular DS project by the fact that its architecture is designed to employ XAI methods. These methods are applied after the predictions made by ML models, that is, post-hoc, and provide a broad range of possible explanations for the behavior of the predictive models.

We started experimenting with two public datasets (German Credit and Default credit card clients). Experiment 1 had the prediction goal of determining good or bad credit risk, and indicated the presence of overfitting for some predictive models due to the large difference in results for the training and test steps. Regarding the explanations of the

models, the general objective of having a consensus of the most important features was achieved, albeit the fact that DiCE did not use any of the important features for the generation of counterfactual instances. On the other hand, PermuteAttack did manage to use two of the most important features. In short, for the German credit dataset, the generation of explanations was successful.

In what concerns Experiment 2, the results show more promise than the ones from the previous experiment. In this case, overfitting was not an issue, mainly because of the larger number of instances in this dataset. As for the explanatory models, the outlook is also brighter. Specifically, for the DiCE method, the features deemed important by PDP, SHAP, and LIME have also been used for the generation of counterfactual examples. The model PermuteAttack generated only a singular instance, with a sole feature having its value changed. Notably, the value change was made on the feature considered most important by PDP, SHAP, and LIME.

The most important experiment, however, was the one performed using the IAPMEI dataset, where the main purpose is to predict the cancelation of publicly funded projects. Careful consideration was taken in the process of business understanding. We have analyzed the life-cycle of a funded project, with a detailed description of its sub-steps. Similarly, for data understanding, since there was a large number of available information to work with, it required an in-depth analysis of what attributes were useful for the task at hand and whether they were usable (e.g. given the percentage of null values). It was observed that micro-enterprises tend to pose a greater risk than bigger companies (Fig. 4.6), but an experiment in which the considered projects were grouped by company size led to inconclusive results, probably because of the low number of projects in the dataset.

In regard to data preparation, the usage of correct business rules enabled the construction of informative and valid features, along with the treatment of null values on the few features that required it, that enabled the usage of projects that would otherwise be discarded. Regarding the modeling stage, the pipeline built for the experiments described in Chapter 3 was used, requiring some fine-tuning due to the introduction of several sampling methods. By using such samplers it was possible to generate synthetic projects in the case of over-samplers, or to reduce the relative difference in the majority class of the dataset, in the case of under-sampling methods. By also using five different seeds for the initialization of the predictive models, and for the split train/test data, enables the reproduction of the experiments made, as well as providing more consistent results, rather than relying on a singular, and random seed which might have given misdirected results. The number of experiments also illustrates the iterative process that is inherent to a Data Science project.

In terms of analysis of results, there are the results given by predictive models and the results of XAI models. While an initial analysis indicated a more negative outlook over the results, by taking into consideration the balance of classes in the target feature, it was possible to determine that these results were satisfactory. More emphasis was given to

the generation of explanations and verifying their respective legitimacy. It was found that the XAI models generated explanations were, generally, in accordance with what features were deemed most important, and this means that it is possible to confirm a positive implementation and usage of XAI techniques.

However, the explanatory process does not end here: it is necessary to reflect on their meaning. It was possible to interpret the obtained results and by analysing the related literature, specifically on the subject of Business demography and Firm survival, it was possible to justify the interpretations made. A critical discussion has also been made, arguments in the evaluation and discussion of results are supported by the literature in this matter. Literature that approximated to this area was found, but in the exact context of projects that use EU structural funds is non-existent.

In summary, while some difficulties were encountered with the implementation of XAI methods, as well as in the predictive process due to the presence of overfitting, it was possible to generate explanations for the black-box models. The explanations given by different XAI techniques proved successful, with the majority of these methods being in agreement regarding the importance of features. The major contributions of this dissertation are twofold: (i) a proposal for the categorization of XAI models through a simplifying taxonomy and the timely collection of XAI techniques in finance; (ii) the successful implementation of these models on a real-world data case study.

The contributions made by this dissertation helped answer the three investigative questions. For the first question, "What is XAI, and what is its relevancy?", it was possible to define XAI as a relatively recent research area whose main purpose is to better understand black-box models. Moreover, there are legal implications such as GDPR that further motivate the development of the area.

It was possible to answer the second research question, "How should XAI techniques be classified?", by solidifying existing knowledge through the literature review, resulting in the categorization of XAI methods in a detailed yet simple taxonomy, proposed in the literature review.

Finally, the third question was also answered, "Are existent XAI methods relevant for real-world applications?", by applying several XAI methods in different Experiments. It was seen that the simultaneous application of explored methods helped in the interpretation of obtained results.

5.1. Limitations and future work

One of the limitations of this dissertation has been the complexity of XAI methods themselves. Since there is no standard way to make code repositories available, the task of adapting XAI techniques to this dissertation proved difficult because their structures, as methods, are inherently different. One technique that fit as a candidate for usage but ultimately had to be discarded was Anchors, since it was found to be incompatible with the process that was built for the other XAI models. Another limitation found while developing the experiments was the lack of data, both for the German Credit dataset

experiment (Experiment 1) and for the IAPMEI dataset. The low number of observations severly limits the implementation of more complex sets of hyperparameter tuning and leads to overfitting. In the case of the experiment with the IAPMEI dataset, this occurred with the usage of under-samplers. Since under-samplers were used for the balancing of the number of projects in the majority class with the numbers of the minority class (the positive instances), the under-sampling results in a decrease in the number of overall projects available for usage in the training data, which was already low originally.

In terms of future work, we could see that XAI is a growing area of research. Namely, it in need of standardization of definitions and taxonomies of XAI methods and of a general agreement on its foundations. This standardization helps researchers already working in the area and also the ones who are starting to work with XAI. Most important, there is also a need to ease the replication of XAI models. While author's code repositories, when available, usually demonstrate how their XAI techniques can be used with public datasets, there needs to be a consensus on the nomenclature of methods. For instance, by grouping XAI models in a package similar to the predictive models present in Scikit-learn. The code used for this dissertation was made public[1].

[1] `https://github.com/tiagoafonsomartins/thesis_mcd`

Sources

Decree-Law No. 57/75
Decree-Law No. 136/2012, 2nd of July

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Rajshree Agarwal and Michael Gort. The determinants of firm survival. *SSRN Electronic Journal*, April 1999.

[3] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. *Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms*, 8:1027–1035, 01 2007.

[4] Susan Athey. *The Impact of Machine Learning on Economics*, pages 507–547. University of Chicago Press, January 2018.

[5] Gustavo E. A. P. A. Batista, Ana Lúcia Cetertich Bazzan, and Maria Carolina Monard. Balancing training data for automated annotation of keywords: a case study. In *WOB*, 2003.

[6] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.*, 6(1):20–29, jun 2004.

[7] Jacobo Chaquet-Ulldemolins, Francisco-Javier Gimeno-Blanes, Santiago Moral-Rubio, Sergio Muñoz-Romero, and José-Luis Rojo-Álvarez. On the black-box challenge for fraud detection using machine learning (ii): Nonlinear analysis through interpretable autoencoders. *Applied Sciences*, 12:3856, 4 2022.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[9] Dangxing Chen, Weicheng Ye, and Jiahui Ye. Interpretable selective learning in credit risk, 2022.

[10] Xolani Dastile and Turgay Celik. Making deep learning-based predictions for credit scoring explainable. *IEEE Access*, 9:50426–50440, 2021.

[11] Xolani Dastile, Turgay Celik, and Hans Vandierendonck. Model-agnostic counterfactual explanations in credit scoring. *IEEE Access*, 10:69543–69554, 2022.

[12] Tanusree De, Prasenjit Giri, Ahmeduvesh Mevawala, Ramyasri Nemani, and Arati Deo. Explainable ai: A hybrid approach to generate human-interpretable explanation for deep learning prediction. *Procedia Computer Science*, 168:40–48, 2020. "Complex Adaptive Systems"Malvern, Pennsylvania-aNovember 13-15, 2019.

[13] Klest Dedja, Felipe Kenji Nakano, Konstantinos Pliakos, and Celine Vens. Bellatrex: Building explanations through a locally accurate rule extractor, 2023.

[14] Klest Dedja, Felipe Kenji Nakano, Konstantinos Pliakos, and Celine Vens. Explaining random forest prediction through diverse rulesets, 2023.

[15] Houtao Deng. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, 7(4):277–287, 2018.

[16] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[17] Timothy Dunne, Shawn D Klimek, and Mark J Roberts. Entrant experience and plant exit. *Working paper*, Working Paper Series(10133), December 2003.

[18] Timothy Dunne, Mark Roberts, and Larry Samuelson. Patterns of firm entry and exit in u.s. manufacturing industries. *RAND Journal of Economics*, 19:495–515, 02 1988.

[19] Ossama Embarak. Decoding the black box: A comprehensive review of explainable artificial intelligence. In *2023 9th International Conference on Information Technology Trends (ITT)*, pages 108–113, 2023.

[20] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

[21] P.A. Geroski. What do we know about entry? *International Journal of Industrial Organization*, 13(4):421–440, 1995. The Post-Entry Performance of Firms.

[22] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning, 2019.

[23] Vasilis Gkolemis, Theodore Dalamagas, and Christos Diou. Dale: Differential accumulated local effects for efficient and accurate global explanations. In Emtiyaz Khan and Mehmet Gonen, editors, *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pages 375–390. PMLR, 12–14 Dec 2023.

[24] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34:14–23, 11 2019.

[25] Masoud Hashemi and Ali Fathi. Permuteattack: Counterfactual explanation of machine learning credit scorecards, 2020.

[26] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328, 2008.

[27] Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

[28] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlì. Castle: Cluster-aided space transformation for local explanations. *Expert Systems with Applications*, 179:115045, 2021.

[29] Valerio La Gatta, Vincenzo Moscato, Marco Postiglione, and Giancarlo Sperlì. Pastle: Pivot-aided space transformation for local explanations. *Pattern Recognition Letters*, 149:67–74, 2021.

[30] Jorma Laurikkala. Improving identification of difficult small classes by balancing class distribution. In Silvana Quaglini, Pedro Barahona, and Steen Andreassen, editors, *Artificial Intelligence in Medicine*, pages 63–66, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[31] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[32] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, jan 2020.

[33] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[34] Ece Çiğdem Mutlu, Niloofar Yousefi, and Ozlem Ozmen Garibay. Contrastive counterfactual fairness in algorithmic decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 499–507, New York, NY, USA, 2022. Association for Computing Machinery.

[35] Alcina Nunes and Elsa Sarmento. Business demography dynamics in portugal: a semi-parametric survival analysis. *In Global Conference on Business and Finance*, 2010.

[36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

[38] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges, 2021.

[39] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. CERTIFAI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, feb 2020.

[40] Colin Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22, 2000.

[41] Yue Tian and Guanjun Liu. Mane: Model-agnostic non-linear explanations for deep learning model. In *2020 IEEE World Congress on Services (SERVICES)*, pages 33–36, 2020.

[42] Ivan Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, 1976.

[43] Alberto Neto Vilas. *Previsão de anulação de projetos financiados por fundos públicos*. Msc thesis, Iscte - Instituto Universitário de Lisboa, December 2021. Available at `http://hdl.handle.net/10071/24119`.

[44] David Watson. Rational shapley values. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022.

[45] Adam White and Artur d'Avila Garcez. Measurable counterfactual local explanations for any classifier. *arXiv e-prints*, August 2019.

[46] Dennis L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

[47] I-Cheng Yeh. Default of credit card clients. UCI Machine Learning Repository, 2016. DOI: https://doi.org/10.24432/C55S3H.

Appendices

APPENDIX A

Experiment 1

Table A.1: Statistical description of features used in the German credit dataset

Feature	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**25%**	**50%**	**75%**	**Maximum**
Status of existing checking account	862	0.54	0.42	0	0	0.33	1	1
Duration in month	862	0.37	0.23	0	0.21	0.37	0.53	1
Credit amount	862	0.30	0.22	0	0.14	0.24	0.41	1
Savings account or bonds	862	0.31	0.25	0	0.25	0.25	0.25	1
Present employment since	862	0.60	0.30	0	0.50	0.50	1	1
Install. rate (%) of disposable income	862	0.67	0.37	0	0.33	0.67	1	1
Present residence since	862	0.61	0.37	0	0.33	0.67	1	1
Age in years	862	0.35	0.23	0	0.18	0.31	0.49	1
No. of existing credits at this bank	862	0.13	0.19	0	0	0	0.33	1

Continued on next page

Table A.1: Statistical description of features used in the German credit dataset (Continued)

Feature	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**25%**	**50%**	**75%**	**Maximum**
Job	862	0.62	0.21	0	0.67	0.67	0.67	1
No. people being liable for	862	0.15	0.36	0	0	0	0	1
Risk	862	0.26	0.44	0	0	0	1	1
Credit history_all_paid_duly	862	0.05	0.21	0	0	0	0	1
Credit history_critical	862	0.30	0.46	0	0	0	1	1
Credit history_delay	862	0.08	0.27	0	0	0	0	1
Credit history_existing_duly_until_now	862	0.54	0.50	0	0	1	1	1
Credit history_none_paid_duly	862	0.03	0.18	0	0	0	0	1
Purpose_business	862	0.08	0.28	0	0	0	0	1
Purpose_car_new	862	0.24	0.43	0	0	0	0	1
Purpose_car_used	862	0.08	0.28	0	0	0	0	1
Purpose_domestic_appliances	862	0.01	0.11	0	0	0	0	1
Purpose_education	862	0.05	0.21	0	0	0	0	1
Purpose_furniture_equipment	862	0.20	0.40	0	0	0	0	1
Purpose_others	862	0.01	0.08	0	0	0	0	1

Continued on next page

Table A.1: Statistical description of features used in the German credit dataset (Continued)

Feature	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**25%**	**50%**	**75%**	**Maximum**
Purpose_radio_television	862	0.30	0.46	0	0	0	1	1
Purpose_repairs	862	0.02	0.15	0	0	0	0	1
Purpose_retraining	862	0.01	0.10	0	0	0	0	1
Personal status and sex_female_divorced_ separated_married	862	0.32	0.47	0	0	0	1	1
Personal status and sex_male_divorced_separated	862	0.05	0.22	0	0	0	0	1
Personal status and sex_male_married_widowed	862	0.10	0.30	0	0	0	0	1
Personal status and sex_male_single	862	0.52	0.50	0	0	1	1	1
Other debtors or guarantors_coapplicant	862	0.04	0.20	0	0	0	0	1
Other debtors or guarantors_guarantor	862	0.06	0.23	0	0	0	0	1
Other debtors or guarantors_none	862	0.90	0.29	0	1	1	1	1
Property_car_other	862	0.33	0.47	0	0	0	1	1

Continued on next page

Table A.1: Statistical description of features used in the German credit dataset (Continued)

Feature	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**25%**	**50%**	**75%**	**Maximum**
Property_real estate	862	0.31	0.46	0	0	0	1	1
Property_soc_savings_life_insurance	862	0.25	0.43	0	0	0	0	1
Property_unknown	862	0.12	0.32	0	0	0	0	1
Other installment plans_bank	862	0.13	0.34	0	0	0	0	1
Other installment plans_none	862	0.82	0.39	0	1	1	1	1
Other installment plans_stores	862	0.05	0.21	0	0	0	0	1
Housing_free	862	0.08	0.26	0	0	0	0	1
Housing_own	862	0.74	0.44	0	0	1	1	1
Housing_rent	862	0.19	0.39	0	0	0	0	1
Telephone_none	862	0.62	0.49	0	0	1	1	1
Telephone_yes	862	0.38	0.49	0	0	0	1	1
Foreign worker_no	862	0.04	0.19	0	0	0	0	1
Foreign worker_yes	862	0.96	0.19	0	1	1	1	1

TABLE A.2. Results of the predictive models on training data

Model	ROC AUC	Accuracy	F1-score	Precision	Recall
DT	56.24%	76.29%	23.53%	78.57%	13.84%
GNB	58.52%	41.29%	46.04%	30.38%	94.97%
LR	70.21%	81.09%	56.82%	71.43%	47.17%
MLP	46.45%	24.71%	39.30%	24.96%	92.45%
RF	99.69%	99.83%	99.68%	100.0%	99.37%
XGB	100.0%	100.00%	100.00%	100.00%	100.00%

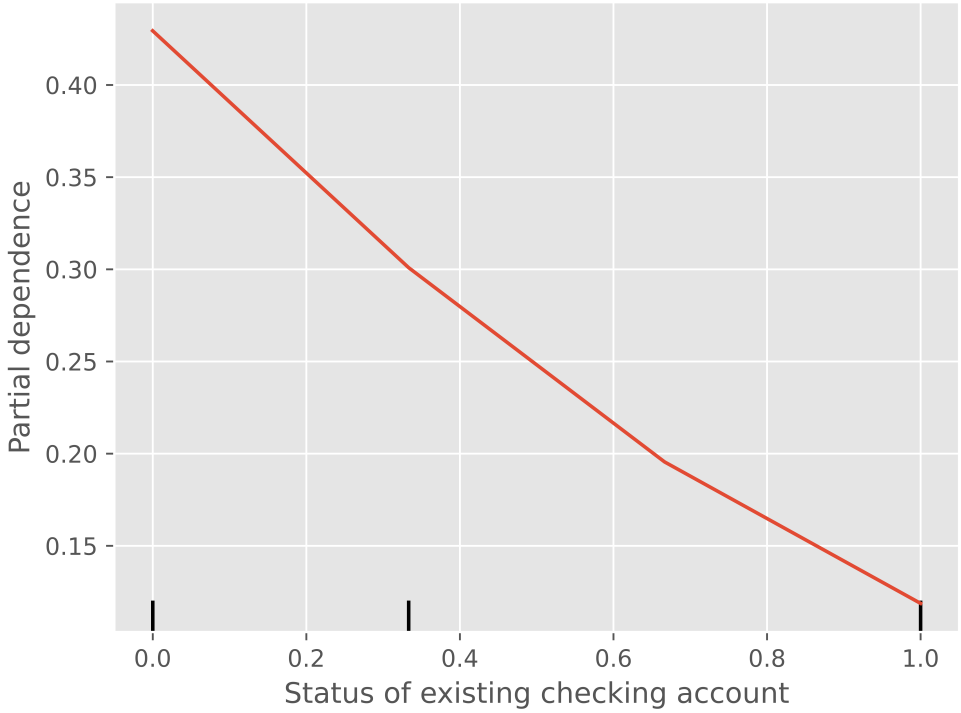


FIGURE A.1. Partial Dependence Plot for the feature Status of existing checking account

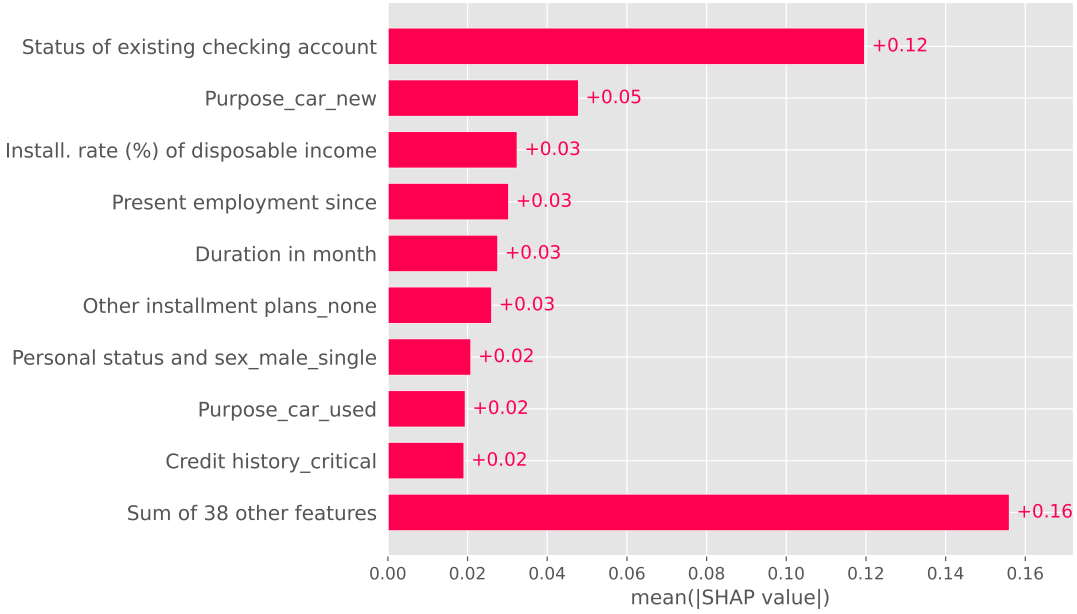


FIGURE A.2. Summary of SHAP for the most important features

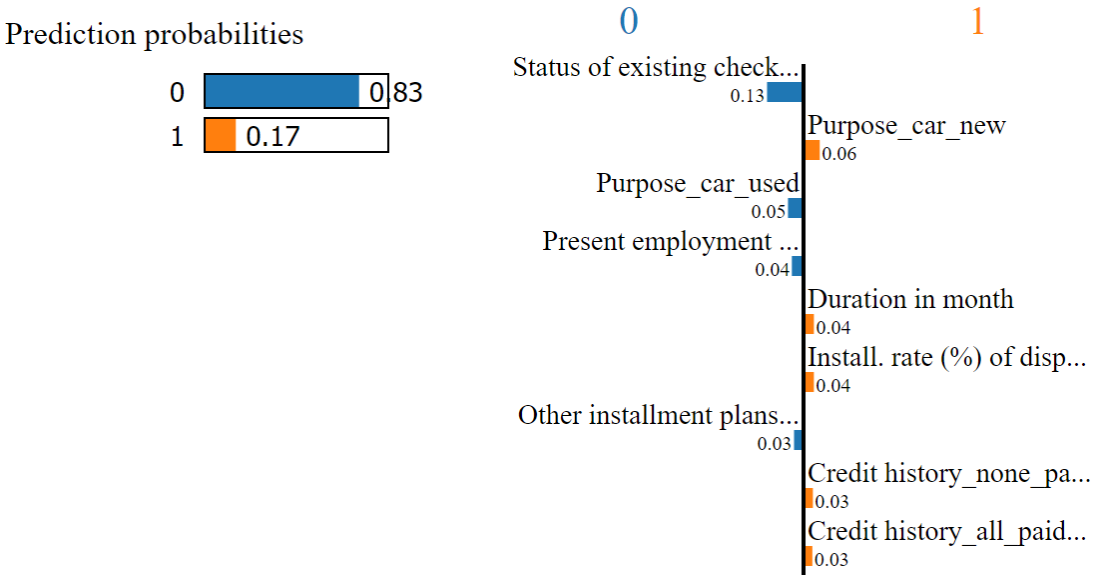


FIGURE A.3. Summary of LIME for the most important features

TABLE A.3. Best hyper-parameters for Decision Trees.

Hyper-parameter	**Value**
criterion	entropy
max_depth	4
max_features	auto
min_samples_leaf	1
min_samples_split	10

Table A.4. Best hyper-parameters for Gaussian Naive-Bayes.

Hyper-parameter	Value
var_smoothing	1e-09

Table A.5. Best hyper-parameters for Logistic Regression.

Hyper-parameter	Value
C	1
max_iter	100
penalty	l1
solver	liblinear

Table A.6. Best hyper-parameters for Multi-Layer Perceptron.

Hyper-parameter	Value
activation	identity
alpha	1e-10
hidden_layer_sizes	[1, 3, 4]
learning_rate	constant
learning_rate_init	0.2
max_iter	200
solver	sgd

Table A.7. Best hyper-parameters for Random Forest.

Hyper-parameter	Value
bootstrap	false
criterion	gini
max_depth	20
max_features	auto
min_samples_leaf	2
min_samples_split	2
n_estimators	100

Table A.8. Best hyper-parameters for XGBoost.

Hyper-parameter	Value
booster	gbtree
colsample_bytree	0.7
learning_rate	0.01
max_depth	10
min_child_weight	1
n_estimators	200
n_thread	-1
objective	reg:squarederror
subsample	0.7

APPENDIX B

Experiment 2

TABLE B.1. Statistical description of features used in the Default credit card dataset

Feature	**Count**	**Mean**	**Standard Deviation**	**Minimum**	**25%**	**50%**	**75%**	**Maximum**
Given credit (NT$)	22816	0.27	0.23	0.00	0.08	0.20	0.38	1.00
Education	22816	0.49	0.26	0.00	0.25	0.50	0.50	1.00
Age	22816	0.45	0.28	0.00	0.25	0.50	0.50	1.00
Past, monthly payment (-1)	22816	0.36	0.23	0.00	0.18	0.31	0.51	1.00
Past, monthly payment (-2)	22816	0.26	0.22	0.00	0.08	0.17	0.36	1.00
Past, monthly bill (-1)	22816	0.29	0.21	0.00	0.12	0.21	0.39	1.00
Past, monthly bill (-2)	22816	0.22	0.20	0.00	0.06	0.18	0.31	1.00
Prev. payment in NT$ (-1)	22816	0.21	0.20	0.00	0.04	0.16	0.29	1.00
Prev. payment in NT$ (-2)	22816	1.84	0.71	1.00	1.00	2.00	2.00	4.00
Gender_female	22816	0.61	0.49	0.00	0.00	1.00	1.00	1.00
Gender_male	22816	0.39	0.49	0.00	0.00	0.00	1.00	1.00
Marital status_married	22816	0.45	0.50	0.00	0.00	0.00	1.00	1.00
Marital status_others	22816	0.01	0.11	0.00	0.00	0.00	0.00	1.00
Marital status_single	22816	0.54	0.50	0.00	0.00	1.00	1.00	1.00
Target	22816	0.23	0.42	0.00	0.00	0.00	0.00	1.00

TABLE B.2. Results on training data

Model	ROC AUC	Accuracy	F1-score	Precision	Recall
DT	64.32%	81.26%	44.72%	69.87%	32.93%
GNB	65.46%	78.79%	46.95%	55.37%	40.74%
LR	60.52%	79.76%	36.13%	66.18%	24.84%
MLP	63.92%	80.63%	43.75%	66.56%	32.92%
RF	66.95%	83.22%	50.24%	79.31%	36.77%
XGB	64.89%	82.13%	45.91%	75.83%	32.92%

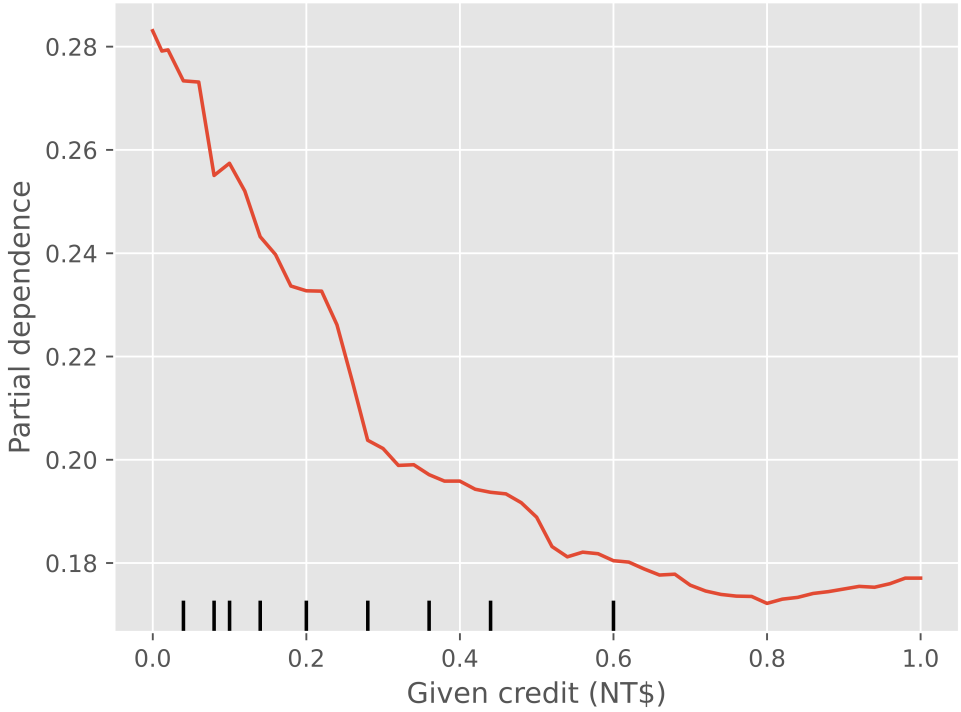


FIGURE B.1. Partial Dependence Plot for the feature Given Credit (NT$)

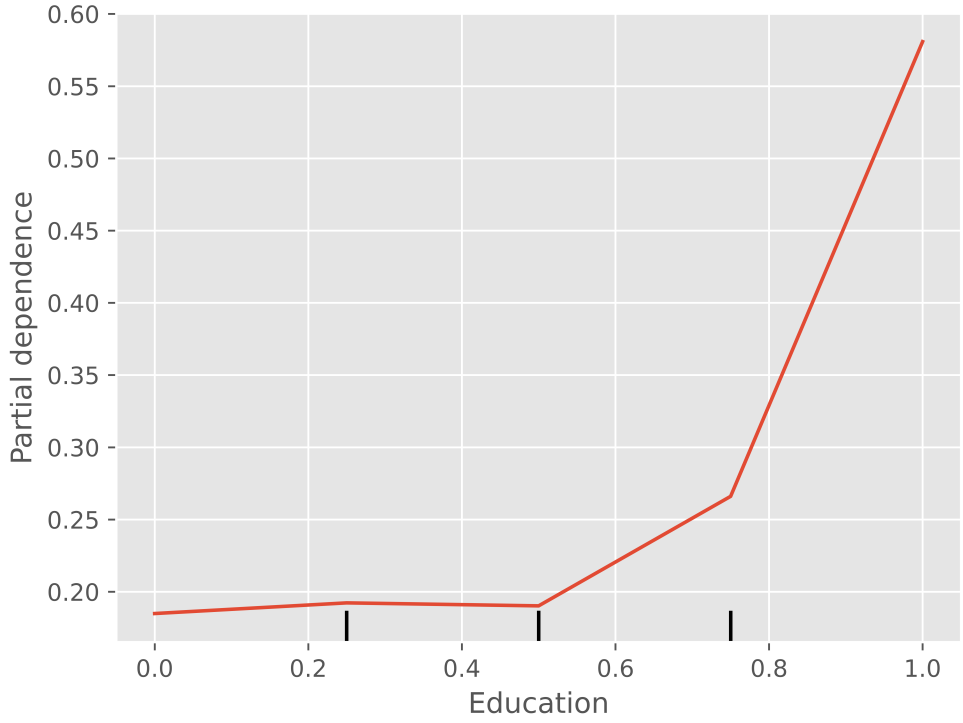


FIGURE B.2. Partial Dependence Plot for the feature Education

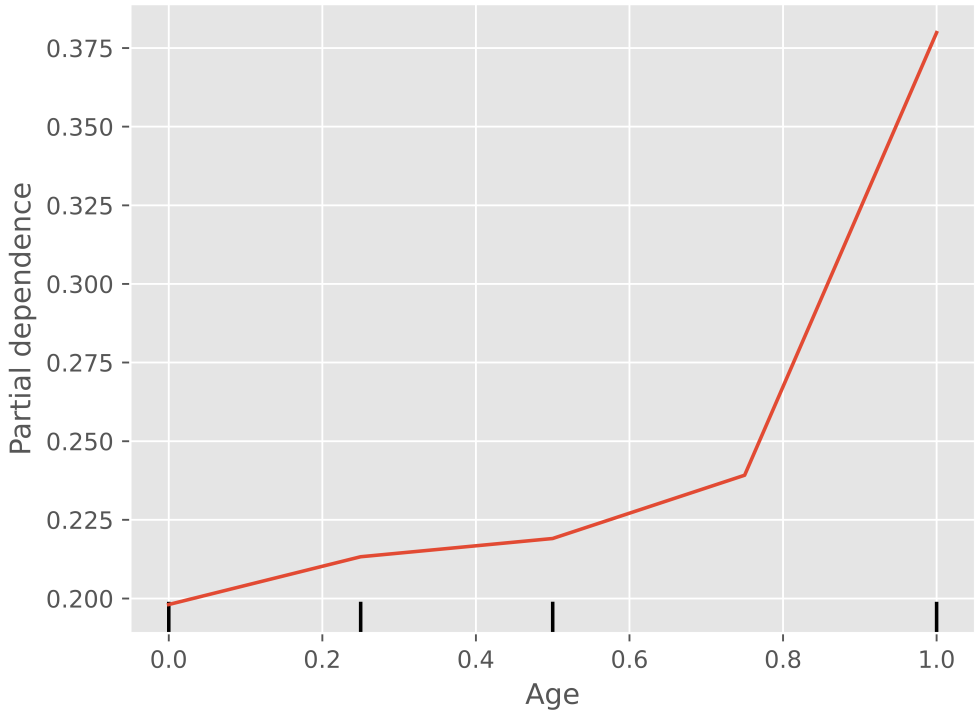


FIGURE B.3. Partial Dependence Plot for the feature Age

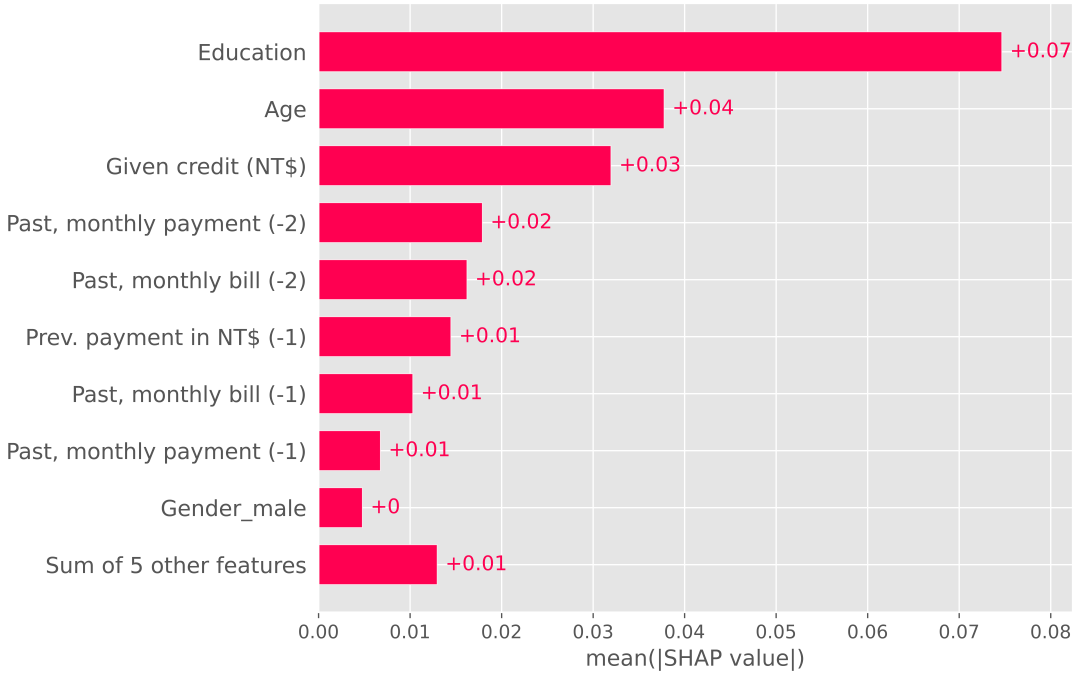


FIGURE B.4. Summary of SHAP for the most important features

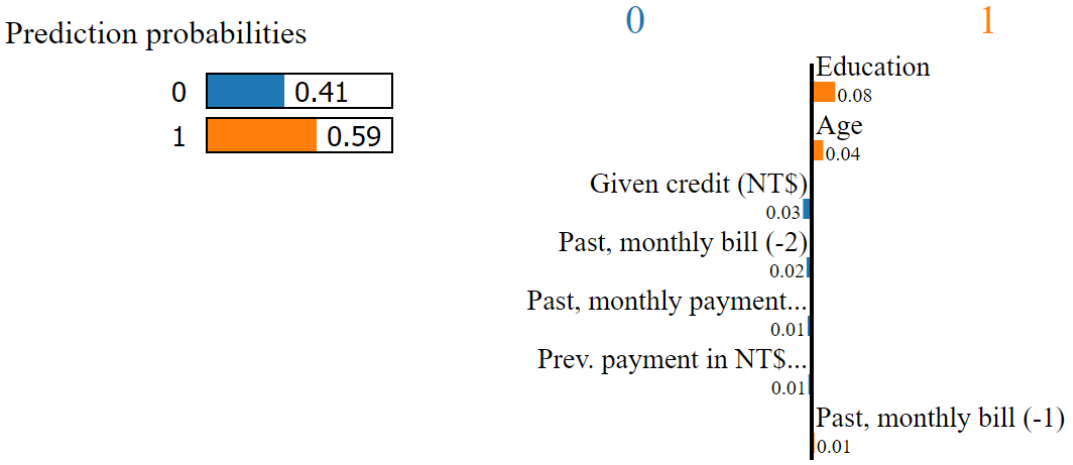


FIGURE B.5. Summary of LIME for the most important features

TABLE B.3. Best hyper-parameters for Decision Trees.

Hyper-parameter	**Value**
criterion	entropy
max_depth	7
max_features	auto
min_samples_leaf	4
min_samples_split	10

Table B.4. Best hyper-parameters for Gaussian Naive-Bayes.

Hyper-parameter	Value
var_smoothing	1e-09

Table B.5. Best hyper-parameters for Logistic Regression.

Hyper-parameter	Value
C	10
max_iter	100
penalty	l2
solver	lbfgs

Table B.6. Best hyper-parameters for Multi-Layer Perceptron.

Hyper-parameter	Value
activation	tanh
alpha	0.001
hidden_layer_sizes	[50, 1]
learning_rate	adaptive
learning_rate_init	0.2
max_iter	200
solver	sgd

Table B.7. Best hyper-parameters for Random Forest.

Hyper-parameter	Value
bootstrap	true
criterion	entropy
max_depth	10
max_features	auto
min_samples_leaf	2
min_samples_split	10
n_estimators	100

Table B.8. Best hyper-parameters for XGBoost.

Hyper-parameter	Value
booster	dart
colsample_bytree	0.5
learning_rate	0.01
max_depth	5
min_child_weight	3
n_estimators	500
n_thread	-1
objective	reg:squarederror
subsample	0.5

APPENDIX C

IAPMEI

Feature meaning and complete data profiling

Table C.1: Features regarding the proposal that were used for the dissertation

Main group	**Name in the dataset**	**Original name of the feature**	**Implementation of the feature**
Proposal	Young company	emp_menos_4_anos_cand	Is older_date_vs_4years <= 4? If any of the required dates is null, then the feature value is null
Proposal	Is micro enterprise	micro_emp	If Resumo/Dimensao == 1: 1, else: 0
Proposal	Is small company	pequena_emp	If Resumo/Dimensao == 1: 2, else: 0
Proposal	Is medium company	media_emp	If Resumo/Dimensao == 1: 3, else: 0
Proposal	Is non-SME	nao_pme	If Resumo/Dimensao == 1: 4, else: 0
Proposal	Investment (€)	proj_investimento	Investment for the project defined in the proposal

Continued on next page

Table C.1: Features regarding the proposal that were used for the dissertation (Continued)

Main group	**Name in the dataset**	**Original name of the feature**	**Implementation of the feature**
Proposal	NUTS II of Project = NUTS II of head office	mesma_sede_nuts_ii	If nuts_ii_sede == nuts_ii_op: 1, else: 0
Expenditure of the proposal	Weight - Equity	PesoCP	rubrica_1 / total
Expenditure of the proposal	Weight - Self-financing	PesoAutofinanciamento	rubrica_2 / total
Expenditure of the proposal	Weight - Foreign capital	PesoCapitais Alheios	if rubrica_408 is not null: (rubrica_4 - rubrica_408) / total else: (rubrica_4 - rubrica_407) / total
Expenditure of the proposal	Weight - Partners	Peso dos Sócios	rubrica_403_102 / total
Expenditure of the proposal	Weight - Total incentive	Peso Incentivo Total	if rubrica_408 is not null: rubrica_408 / total else: rubrica_407 / total
Expenditure of the proposal	Weight - Reimbursable incentive	Peso Incentivo N Reembolsável	rubrica_40701 / total
Expenditure of the proposal	Eligibility (%)	%elegibilidade	rubrica_92 / total

Continued on next page

Table C.1: Features regarding the proposal that were used for the dissertation (Continued)

Main group	**Name in the dataset**	**Original name of the feature**	**Implementation of the feature**
Expenditure of the proposal	Errors in weights	erros_racios_inv	If any of the Weights is larger than 1, fill with 1, otherwise, 0
Additional Information on the company	Number of workers	N_Linha	Number of workers for each company
Expenses on external services	Value of training	Val_Calc	Filter values where Id == 90, sum all values for each company

TABLE C.2. Profiling of the features created from the proposal information.

Feature	**Mean**	**Standard Deviation**	**Minimum**	**1st Quartile**	**2nd Quartile**	**3rd Quartile**	**Maximum**
Investment (€)	2025738.00	2786257.00	29780.00	576168.60	1123362.00	2375546.00	24995250.00
Young company	0.21	0.41	0	0	0	0	1.00
NUTS II of Project = NUTS II of head office	0.99	0.11	0.00	1.00	1.00	1.00	1.00
Value of training	1101.13	7770.67	0	0	0	0	146125.60
Number of workers	11.06	5.81	2.00	7.00	10.00	14.00	36.00
Weight - Equity	0.15	0.64	0	0	0.15	0.21	19.90
Weight - Self-financing	0.18	0.20	0	0	0.13	0.30	1.79
Weight - Foreign capital	0.11	0.16	0	0	0.02	0.20	1.40
Weight - Partners	0.13	0.15	0	0	0.10	0.22	0.80
Weight - Total incentive	0.59	0.13	0	0.53	0.60	0.70	1.45
Weight - Reimbursable incentive	0.05	0.14	0	0	0	0	0.75
Eligibility (%)	0.96	0.10	0.30	0.98	1.00	1.00	1.00
Errors in weights	0.01	0.09	0	0	0	0	1.00
Is micro enterprise	0.21	0.41	0	0	0	0	1.00
Is small company	0.44	0.50	0	0	0	1.00	1.00
Is medium company	0.33	0.47	0	0	0	1.00	1.00
Is non-SME	0.02	0.15	0	0	0	0	1.00

TABLE C.3. Features regarding the expenses that were used for the dissertation

Name in the dataset	**Original name of the feature**	**Implementation of the feature**
Expenses - IT	tipodesp_Software _Equipamentos_Informaticos	despesa_total_Software_Equipamentos_Informaticos / Investimento
Expenses - Civil construction	tipodesp_Construcao - Remodelacao_Edificios	despesa_total_Construcao/Remodelacao_Edificios / Investimento
Expenses - Engineering services	tipodesp_Estudos _Diagnosticos_Licencas_ServicosEngenharia	despesa_total_Estudos_Diagnosticos _Licencas_ServicosEngenharia / Investimento
Expenses - Equipment	tipodesp_Maquinas _Equipamentos	despesa_total_Maquinas_Equipamentos / Investimento
Expenses - Other	tipodesp_Outras Despesas	despesa_total_Outras Despesas / Investimento

TABLE C.4. Profiling of the features created from the expenditure.

Feature	**Mean**	**Standard Deviation**	**Minimum**	**1st Quartile**	**2nd Quartile**	**3rd Quartile**	**Maximum**
Expenses - Civil construction	0.14	0.14	0	0.00	0.10	0.23	0.70
Expenses - Engineering services	0.02	0.03	0	0	0.01	0.03	0.29
Expenses - Equipment	0.73	0.19	0	0.62	0.75	0.87	1.00
Expenses - IT	0.04	0.09	0	0	0.02	0.05	0.99
Expenses - Other	0.07	0.11	0	0.01	0.02	0.08	0.85

TABLE C.5. Features regarding the financial indicators that were used for the dissertation

Name in the dataset	**Original name of the feature**	**Implementation of the feature**
Mean turnover	prom_vol_negocio_med_anual	Mean of all values of _5001_VENDAS_SERVICOS_PRESTADOS for the company
Asset	prom_ativo_total_t-1	_5127_ATIVO_TOTAL
Turnover	prom_volume_negocios_t-1	_5001_VENDAS_SERVICOS_PRESTADOS
Liabilities (credit)	prom_emprestimo_obtidos_passivo_ncor_t-1	_5143_PASSIVO_NC_FINANCIAMENTOS_OBTD
Liquidity	liquidez_geral_t-1	(_5113_ATIVO_COR_INVENTARIOS + _5114_ATIVO_COR_ACTIVOS_BIOLOGICOS + _5115_ATIVO_COR_CLIENTES + _5116_ATIVO_COR_ADIANTAMENTOS_FORNEC + _5117_ATIVO_COR_ESTADO_OUT_ENTES_PUB)/ (”_5148_PASSIVO_COR_FORNCEDORES” + _5149_PASSIVO_COR_ADIANTA_DE_CLIENTES + _5150_PASSIVO_COR_ESTADO_OUT_ENT_PUB)
Equity	rentabilidade_capitais_proprios_t-1	_5025_RESULTADO_LIQUIDO_PERIODO / _5141_CP_TOTAL
Net profit or loss	resultadoliquido_ativo t-1	_5139_CP_RESULTADO_LIQUIDO_PERIODO / _5127_ATIVO_TOTAL
Financial mean turnover	prom_financ_vol_negocio_med_anual	Investment present in the proposal dividing by the mean turnover: Dadosprojecto/Investimento / prom_vol_negocio_med_anual

TABLE C.6. Profiling of the features created from IES.

Feature	**Mean**	**Standard Deviation**	**Minimum**	**1st Quartile**	**2nd Quartile**	**3rd Quartile**	**Maximum**
Mean turnover	5345291.00	8116707.00	0	935741.80	2547636.00	6335655.00	84261754.00
Asset	4286591.00	6758712.00	0	528896.60	1759049.00	4960884.00	63025587.00
Turnover	3685638.00	6009630.00	0	347164.80	1501980.00	4246600.00	64111773.00
Liabilities (credit)	607762.90	1628532.00	0	0.00	100576.70	577379.00	35527902.00
Liquidity	inf		0	1.32	2.18	3.61	inf
Equity	inf		0	0	0.02	0.09	inf
Net profit or loss	0.02	0.04	0	0	0.01	0.03	0.68
Financial mean turnover	inf		0.01	0.19	0.42	1.02	inf

TABLE C.7. Social-economic features from INE that were used for the dissertation

Name in the dataset	**Original name of the feature**	**Implementation of the feature**
Gini Index NUTS	Gini_NUTS_prom	Median of Gini coefficient for each NUTS II, for the interval from 2017 to 2021. 5 attributes were used, one for each year
Company closure NUTS	estemp	Mean of the number of company closures for each NUTS II, for the interval from 2014 to 2021. It was not possible to extract the number of closures on a yearly basis, only monthly. In total, 96 attributes were used, one for each month, which were aggregated, and summed on a yearly basis in order to extract the yearly mean.
Population density NUTS	densidade_pop_NUTS_prom	Median of population density for each NUTS II, for the interval from 2014 to 2021. 8 attributes were used, one for each year
College Network NUTS	Rede_Universitaria_distrito	Median of the number of institutes for higher education for each NUTS II, for the interval from 2014 to 2021. 8 attributes were used, one for each year
Number of SME NUTS	n_pme_NUTS_prom	Median of the number of companies for each NUTS II, for the interval from 2014 to 2021. 8 attributes were used, one for each year
Mean salary NUTS	remuneracao_mensal_media_NUTS_prom	Median of the monthly salary for each NUTS II, for the interval from 2014 to 2021. 8 attributes were used, one for each year
Unemployment rate NUTS	Tx_desemp_NUTS_prom	Median of the unemployment rate for each NUTS II, for the interval from 2014 to 2020 (2021 not available). 7 attributes were used, one for each year
Poverty rate NUTS	Risco_pobreza_NUTS_prom	Median of the unemployment rate for each NUTS II, for the interval from 2017 to 2020 (2014-2016 not available). 5 attributes were used, one for each year

Table C.8. Profiling of the features created from INE's available data.

Feature	Mean	Standard Deviation	Minimum	1st Quartile	2nd Quartile	3rd Quartile	Maximum
Gini Index NUTS	30.99	0.71	30.30	30.30	31.30	31.30	32.70
Company closure NUTS	5265.50	1961.13	1150.50	3887.50	5254.00	6620.50	8786.50
Population density NUTS	770.29	435.18	88.70	426.40	733.75	1041.10	1685.90
College Network NUTS	73.44	26.02	11.50	54.00	72.25	99.50	99.50
Number of SME NUTS	338655.70	103553.80	75669.50	274746.00	325722.80	441321.50	441321.50
Mean salary NUTS	896.98	93.85	836.09	851.95	869.70	887.44	1187.08
Unemployment rate NUTS	8.43	1.37	6.90	6.90	8.95	9.80	9.80
Poverty rate NUTS	17.35	1.73	12.30	17.30	17.30	18.60	18.70

Table C.9. Features created after the merge of the dataset

Name in the dataset	Original name of the feature	Implementation of the feature
Project cancelled	proj_nulled	If the unique identifier for the proposal is present in the file/table "Anulações-Resposta.csv": 1, else: 0
Historical frequency of cancellation CAE	freq_target_anul_CAE	Divide the number of projects cancelled with a given CAE by the total number of projects with that CAE. The attribute for CAE is present in the proposal, under "Resumo/Cae"
Historical frequency of cancellation NUTS	freq_target_anul_distrito	Divide the number of projects cancelled with a given NUTS II by the total number of projects with that NUTS II. NUTS II is given by "nuts_ii_sede"

TABLE C.10. Distribution of project cancelations for age of company - Micro-sized entreprises

Project cancelled	Not young	Young
0	66	45
1	36	82

Young company: year of birth is less than four years apart from the application year.

TABLE C.11. Distribution of project cancelations for age of company - Small-sized companies

Project cancelled	Not young	Young
0	329	34
1	82	35

Young company: year of birth is less than four years apart from the application year.

TABLE C.12. Distribution of project cancelations for age of company - Medium-sized companies

Project cancelled	Not young	Young
0	282	30
1	46	9

Young company: year of birth is less than four years apart from the application year.

Table C.13. Distribution of projects by CAE

Initial CAE number	**Final CAE number**	**Group**	**Description of the Group**	**Number of projects**
10	33	C	Manufacturing industries	989
36	39	E	Water collection, treatment and distribution; sanitation, waste management and remediation	<10
41	43	F	Construction	<10
45	47	G	Wholesale and retail trade; repair of motor vehicles and motorcycles	15
49	53	H	Transportation and storage	<10
58	63	J	Information and communication activities	11
69	75	M	Professional, scientific and technical activities	14
77	82	N	Administrative and support service activities	16
85	85	P	Education	<10
86	88	Q	Human health and social support activities	<10
-	-	NA	-	<10

Due to several CAEs having few companies with projects, the table was anonymized to not include exact number lower than 10. The number of projects canceled and closed was not included as well for this reason.

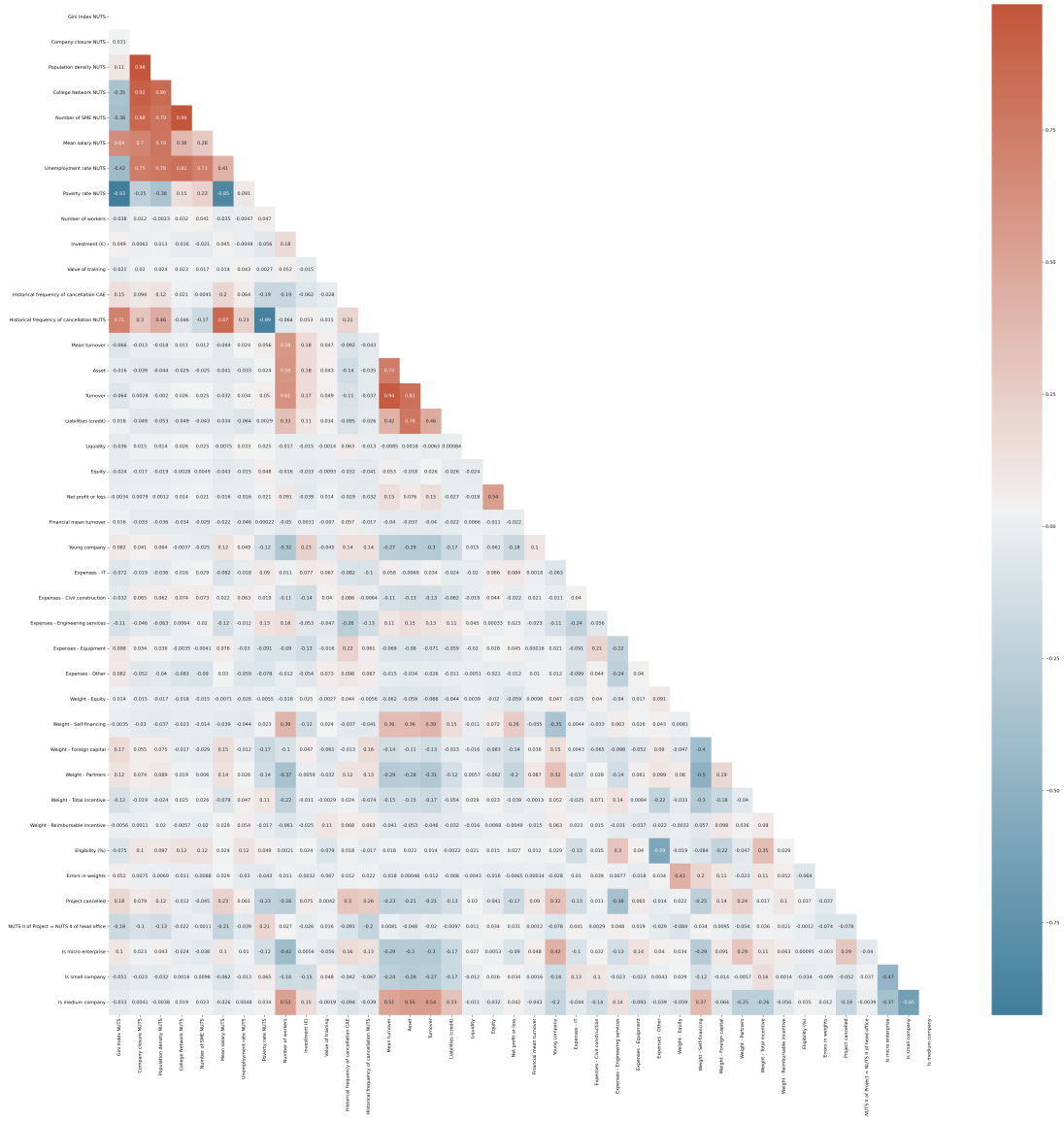


FIGURE C.1. Correlation matrix for all features

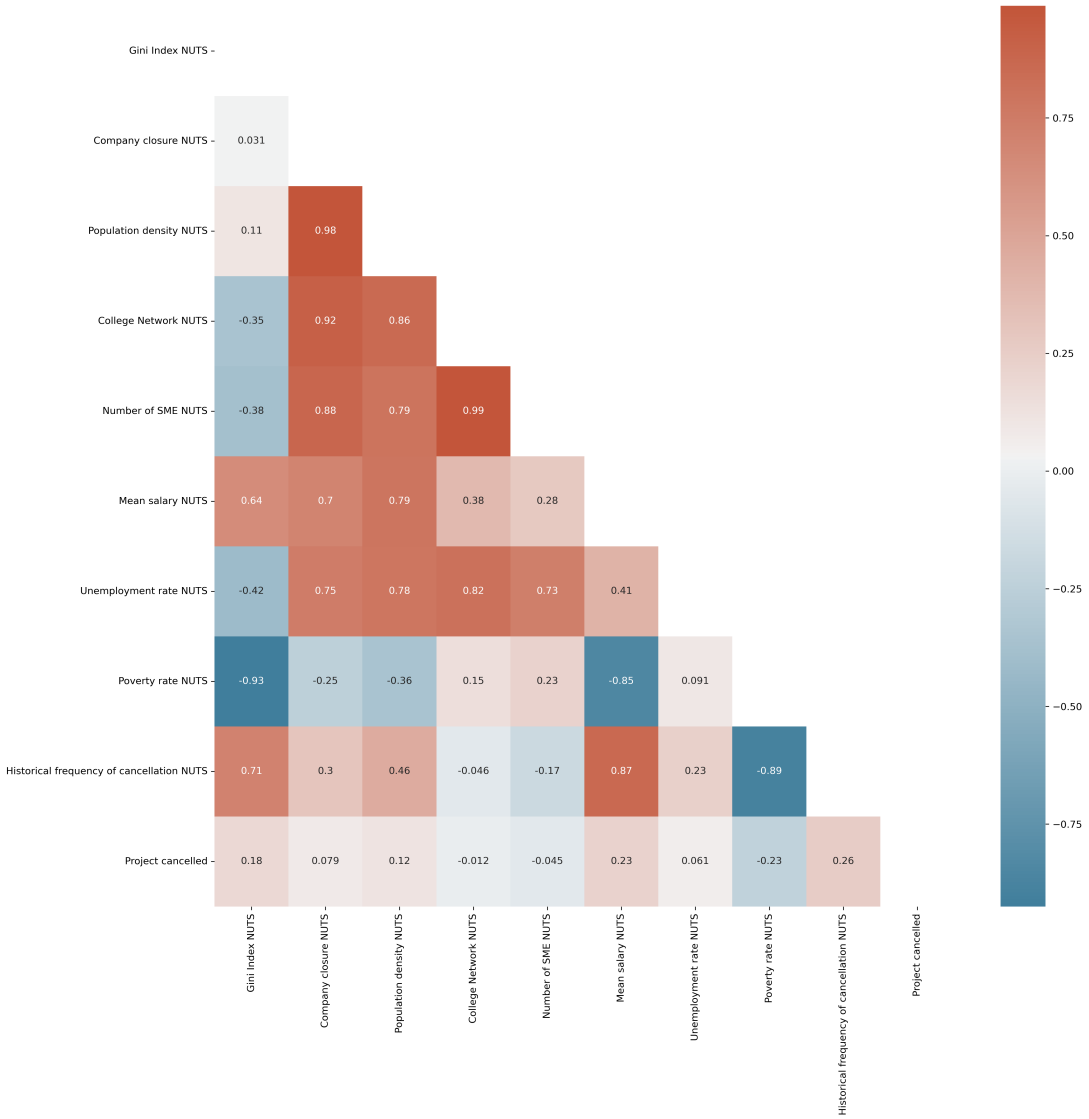


FIGURE C.2. Correlation matrix for the features related to INE

8]

Table C.14: Train results for experiments 17-27

Experiment	Sampler	Model	ROC AUC	Accuracy	F1-score	Precision	Recall
17	None	DT	70%	82%	56%	85%	42%
		GNB	71%	66%	56%	43%	82%
		LR	64%	80%	46%	81%	32%
		MLP	71%	83%	54%	84%	45%
		RF	90%	95%	89%	100%	80%
		XGB	86%	92%	84%	98%	73%
18	Under	DT	76%	76%	75%	81%	71%
		GNB	73%	73%	76%	69%	85%
		LR	80%	80%	79%	82%	77%
		MLP	73%	73%	61%	68%	55%
		RF	95%	95%	95%	95%	95%
		XGB	100%	100%	100%	100%	100%
19	Over	DT	100%	100%	100%	100%	100%
		GNB	70%	70%	74%	65%	88%
		LR	81%	81%	81%	83%	79%

Continued on next page

Table C.14: Train results for experiments 17-27 (Continued)

Experiment	**Sampler**	**Model**	**ROC AUC**	**Accuracy**	**F1-score**	**Precision**	**Recall**
		MLP	80%	80%	80%	82%	78%
		RF	100%	100%	100%	100%	100%
		XGB	100%	100%	100%	100%	100%
20	Tomek	DT	74%	83%	64%	79%	55%
		GNB	73%	68%	59%	46%	82%
		LR	68%	81%	54%	80%	41%
		MLP	77%	85%	69%	84%	59%
		RF	91%	95%	90%	99%	82%
		XGB	90%	94%	89%	100%	80%
21	Smote	DT	100%	100%	100%	100%	100%
		GNB	69%	69%	74%	63%	89%
		LR	82%	82%	81%	83%	80%
		MLP	79%	79%	78%	83%	75%
		RF	100%	100%	100%	100%	100%
		XGB	100%	100%	100%	100%	100%
22	Adasyn	DT	97%	97%	97%	99%	95%
		GNB	65%	65%	72%	60%	88%
		LR	78%	78%	78%	79%	77%

Continued on next page

Table C.14: Train results for experiments 17-27 (Continued)

Experiment	Sampler	Model	ROC AUC	Accuracy	F1-score	Precision	Recall
		MLP	75%	75%	72%	82%	66%
		RF	100%	100%	100%	100%	100%
		XGB	100%	100%	100%	100%	100%
23	SmoteTomek	DT	96%	96%	96%	96%	96%
		GNB	74%	74%	73%	76%	72%
		LR	82%	82%	82%	83%	80%
		MLP	80%	80%	80%	82%	78%
		RF	100%	100%	100%	100%	100%
		XGB	100%	100%	100%	100%	100%
24	SmoteTeenn	DT	94%	94%	95%	96%	95%
		GNB	83%	80%	81%	96%	70%
		LR	91%	91%	92%	92%	92%
		MLP	89%	89%	91%	93%	88%
		RF	100%	100%	100%	100%	100%
		XGB	100%	100%	100%	100%	100%
25	Cluster	DT	89%	89%	89%	91%	88%
		GNB	71%	71%	73%	68%	80%
		LR	79%	79%	79%	79%	79%

Continued on next page

Table C.14: Train results for experiments 17-27 (Continued)

Experiment	**Sampler**	**Model**	**ROC AUC**	**Accuracy**	**F1-score**	**Precision**	**Recall**
		MLP	74%	74%	76%	72%	82%
		RF	99%	99%	99%	100%	99%
		XGB	100%	100%	100%	100%	100%
26	NeighbourhoodClean	DT	85%	88%	82%	89%	76%
		GNB	79%	82%	74%	80%	69%
		LR	80%	84%	76%	90%	65%
		MLP	84%	86%	79%	86%	75%
		RF	100%	100%	100%	100%	100%
		XGB	97%	97%	96%	100%	93%
27	NearestNeighbours	DT	89%	91%	87%	94%	81%
		GNB	75%	81%	66%	89%	55%
		LR	81%	85%	77%	92%	66%
		MLP	85%	88%	82%	91%	75%
		RF	99%	99%	99%	100%	98%
		XGB	99%	99%	99%	100%	99%

12]

Table C.15: Test results for experiments 17-27

Experiment	Sampler	Model	ROC AUC	Accuracy	F1-score	Precision	Recall	TN	FP	FN	TP
17	None	DT	62%	76%	42%	64%	32%	220	16	60	28
		GNB	72%	67%	58%	44%	83%	144	92	15	73
		LR	62%	77%	42%	70%	30%	225	11	62	26
		MLP	66%	78%	47%	67%	41%	218	18	52	36
		RF	73%	82%	62%	71%	55%	216	20	40	48
		XGB	71%	80%	59%	68%	52%	214	22	42	46
18	Under	DT	70%	72%	57%	52%	67%	174	62	29	59
		GNB	72%	66%	58%	44%	85%	138	98	13	75
		LR	75%	75%	62%	53%	73%	180	56	24	64
		MLP	69%	76%	48%	45%	53%	200	36	42	46
		RF	78%	78%	66%	58%	78%	185	51	19	69
		XGB	76%	77%	63%	56%	73%	185	51	24	64
19	Over	DT	65%	71%	49%	48%	51%	187	49	43	45
		GNB	68%	60%	54%	39%	86%	117	119	12	76
		LR	75%	76%	62%	54%	74%	181	55	23	65
		MLP	73%	75%	60%	53%	71%	179	57	26	62
		RF	73%	83%	62%	75%	53%	221	15	41	47
		XGB	73%	82%	62%	75%	53%	221	15	42	46

Continued on next page

Table C.15: Test results for experiments 17-27 (Continued)

Experiment	Sampler	Model	ROC AUC	Accuracy	F1-score	Precision	Recall	TN	FP	FN	TP
20	Tomek	DT	66%	75%	50%	57%	45%	205	31	49	39
		GNB	72%	67%	57%	44%	82%	145	91	16	72
		LR	62%	76%	42%	59%	33%	216	20	59	29
		MLP	72%	79%	60%	64%	56%	208	28	38	50
		RF	74%	81%	62%	67%	58%	211	25	37	51
		XGB	74%	81%	62%	66%	59%	209	27	36	52
21	Smote	DT	66%	70%	51%	46%	57%	177	59	38	50
		GNB	67%	58%	53%	38%	86%	111	125	12	76
		LR	74%	75%	61%	53%	73%	180	56	24	64
		MLP	73%	75%	60%	54%	71%	180	56	26	62
		RF	77%	83%	67%	71%	64%	213	23	32	56
		XGB	76%	83%	66%	72%	61%	215	21	34	54
22	Adasyn	DT	63%	69%	47%	44%	50%	180	56	44	44
		GNB	65%	53%	51%	36%	90%	93	143	9	79
		LR	73%	72%	60%	49%	75%	168	68	22	66
		MLP	73%	75%	60%	54%	69%	181	55	27	61
		RF	77%	81%	66%	66%	66%	206	30	30	58
		XGB	75%	81%	64%	67%	61%	209	27	34	54

Continued on next page

Table C.15: Test results for experiments 17-27 (Continued)

Experiment	Sampler	Model	ROC AUC	Accuracy	F1-score	Precision	Recall	TN	FP	FN	TP
23	SmoteTomek	DT	68%	72%	53%	49%	59%	181	55	36	52
		GNB	73%	74%	60%	55%	70%	179	57	26	62
		LR	78%	80%	67%	61%	74%	194	42	23	65
		MLP	77%	79%	66%	60%	73%	192	44	24	64
		RF	78%	84%	68%	73%	64%	215	21	31	57
		XGB	76%	83%	66%	71%	62%	214	22	34	54
24	SmoteTeenn	DT	67%	68%	52%	44%	65%	161	75	30	58
		GNB	70%	72%	56%	49%	65%	176	60	31	57
		LR	73%	69%	58%	46%	81%	151	85	17	71
		MLP	73%	70%	59%	48%	79%	159	77	19	69
		RF	76%	77%	64%	55%	76%	181	55	21	67
		XGB	75%	75%	62%	53%	75%	176	60	22	66
25	Cluster	DT	65%	64%	51%	41%	68%	147	89	28	60
		GNB	70%	65%	56%	44%	80%	140	96	17	71
		LR	74%	75%	61%	53%	73%	178	58	24	64
		MLP	72%	69%	59%	48%	80%	152	84	17	71
		RF	73%	70%	59%	47%	80%	155	81	18	70
		XGB	72%	69%	58%	46%	78%	156	80	20	68

Continued on next page

Table C.15: Test results for experiments 17-27 (Continued)

Experiment	Sampler	Model	ROC AUC	Accuracy	F1-score	Precision	Recall	TN	FP	FN	TP
26	NeighbourhoodClean	DT	70%	74%	56%	52%	60%	187	49	35	53
		GNB	74%	76%	62%	56%	69%	186	50	27	61
		LR	74%	79%	62%	62%	62%	203	33	33	55
		MLP	76%	77%	64%	58%	73%	186	50	24	64
		RF	78%	80%	67%	61%	75%	194	42	22	66
		XGB	76%	79%	64%	59%	70%	193	43	26	62
27	NearestNeighbours	DT	71%	74%	57%	52%	64%	183	53	31	57
		GNB	68%	75%	52%	55%	53%	196	40	42	46
		LR	74%	79%	62%	60%	65%	199	37	31	57
		MLP	75%	77%	62%	56%	70%	187	49	26	62
		RF	77%	78%	65%	57%	75%	186	50	22	66
		XGB	76%	78%	64%	57%	73%	187	49	23	65

Hyper-parameters for the predictive models

This section contains the best set of hyper-parameters found for experiment 23 which was considered for the evaluation and discussion of results.

TABLE C.16. Best hyper-parameters for Decision Trees.

Hyper-parameter	Value
criterion	gini
max_depth	10
max_features	auto
min_samples_leaf	1
min_samples_split	2

TABLE C.17. Best hyper-parameters for Gaussian Naive-Bayes.

Hyper-parameter	Value
var_smoothing	1e-09

TABLE C.18. Best hyper-parameters for Logistic Regression.

Hyper-parameter	Value
C	500
max_iter	100
penalty	L1
solver	liblinear

TABLE C.19. Best hyper-parameters for Multi-Layer Perceptron.

Hyper-parameter	Value
activation	identity
alpha	1e-10
hidden_layer_sizes	[1, 3, 4]
learning_rate	adaptive
learning_rate_init	0.005
max_iter	300
solver	adam

Table C.20. Best hyper-parameters for Random Forest.

Hyper-parameter	**Value**
bootstrap	false
criterion	gini
max_depth	20
max_features	log2
min_samples_leaf	1
min_samples_split	2
n_estimators	100

Table C.21. Best hyper-parameters for XGBoost.

Hyper-parameter	**Value**
booster	gbtree
colsample_bytree	0.7
learning_rate	0.1
max_depth	5
min_child_weight	3
n_estimators	200
n_thread	-1
objective	reg:squarederror
subsample	0.7

APPENDIX D

Accompanying Article for the Literature Review

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Explainable Artificial Intelligence (XAI): a Systematic Literature Review on Taxonomies and Applications in Finance

Tiago Martins[1], Ana de Almeida[1,2,3], Senior Member, IEEE, Elsa Cardoso[1,4], Luis Nunes[1,2]

[1]Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[2]Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisboa, Portugal
[3]CISUC – Center for Informatics and Systems of the University of Coimbra
[4]CIES-Iscte - Centro de Investigação e Estudos de Sociologia, Lisboa Portugal

Corresponding author: Tiago Martins[1] (e-mail: Afonso_Martins@iscte-iul.pt).

This work was partially supported by Fundação para a Ciência e a Tecnologia, I.P. (FCT) [ISTAR Projects: UIDB/04466/2020; UIDP/04466/2020 and POAT-01-6177-FEDER-000059].

ABSTRACT Artificial Intelligence and the progress of Machine Learning led to significant growth in applications to real-world problems. However, many Machine Learning models are complex and often used without a clear and transparent understanding of the logic behind what happens: the so-called black-box models. We present a systematic literature review on Explainable Artificial Intelligence (XAI) methods for tabular data with a focus on the financial domain. Recent applications of XAI in the area of Finance will be presented along with a review of the most popular methods used. For the sake of the uniformization of taxonomies, we propose a categorization of the XAI methods found. This new organization results in a more concise definition of existing explainable methods and techniques only using the most common categories found in the reviewed literature. Moreover, we pinpoint which of the works apply which of the methods, as well as the most used open datasets within the financial domain.

INDEX TERMS AI, Artificial Intelligence, Financial applications, Explainable Machine Learning, Systematic Literature Review, XAI

I. INTRODUCTION

Explainable Artificial Intelligence, XAI[1], is an area that aims to improve the interpretation and explanation of Machine Learning (ML) algorithms and their results. Due to the growing relevance of ML algorithms in recent decades, mainly through black-box approaches such as neural networks or random forests, interest in the ability to interpret and explain these approaches has increased in several application areas, with emphasis in areas related to Health and Finance [1]. The most complex models that learn from examples, that is, supervised learning using neural networks or randomization, where one is expected to input the characteristics of the example and its output, exhibit no or limited transparency. Such models, high in performance yet low in comprehension, need to be explained so that users can understand the (reasons for the) outputs of these models and, consequently, informed decisions can be supported.

Although no universal definition of explainability exists, numerous works related to XAI, with different purposes and levels of detail for explainability, enable the definition of the main objectives of this area. The purpose of an XAI technique is to understand the behavior of an ML model and its output, as mentioned by M.Turek [2]: "...XAI aims to help users understand, appropriately trust, and effectively manage the emerging generation of Artificial Intelligence (AI) systems." The Assessment List for Trustworthy Artificial Intelligence (ALTAI) defines explainability as a "feature of an AI system that is intelligible to non-experts. An AI system is intelligible if its functionality and operations can be explained non-technically to a person not skilled in the art." Besides these definitions, experts are also highly interested in understanding what is happening inside a model, which can be defined as the interpretability of ML models. Christoph Molnar proposes to define Interpretable ML as the methods and models that make the behavior and predictions of machine learning systems understandable to humans [22]. In general, explanations are meant for humans to trust black-box methods, and explainability mainly focuses on models that can summarize the reasons for the model's results or give insights about the causes of the decisions that have been

[1] Acronym popularized by the Defense Advanced Research Projects Agency (DARPA), in 2016, when an announcement was made to potentially fund research proposals in ML towards Explainable Artificial Intelligence (https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf)

made and be auditable [63]. Other relevant works in the attempt at a definition of explainability and interpretability can be found in [10][64][65].

While the primary purpose of this area of study is to help understand ML models, there is also a legal motivation to help further this area, namely the General Data Protection Regulation (2016/679, GDPR), which is a privacy and data protection regulation[2]. Within the European Union and Economic Area, projects envolving personal data must comply with this regulation and the possible legal repercussions [3]. GDPR is the European Union's effort to serve the interests of its citizens regarding how their personal data is used by third parties, as well as defining the obligations of the parties and establishing citizens' rights. Among these, in Article 17, we can find the "right to forget," where the data subject, typically the citizen, can ask the data holder to erase his/her personal data, or, according to Article 21, the right to object to the processing of his/her personal data. While the phrasing of these articles is open to interpretation, some pave the way for XAI as an obligation rather than an optional feature. The GDPR clearly defines that personal data should be processed in a "…lawfully, fairly and in a transparent manner in relation to the data subject…," as seen in Article 5. While there might be some doubt regarding the applicability of this article, there is a more detailed definition of transparency applied to ML models in the right for a data subject to have the information regarding "...the existence of automated decision-making…" as well as the process, importance, and consequences behind such decision-making, according to Article 14, paragraph 2.g). The need for an explanation of ML models becomes even more apparent because it implies that the prediction and the logic behind it should be made available to the user.

The purpose of this paper is two-fold: firstly, to integrate current knowledge regarding XAI techniques and methods, specifically for tabular data; secondly, based on the results of a systematic literature review, introduce the specific XAI methods and techniques that have been applied in the financial domain. The paper is organized as follows: Section 2 describes the methodology used for the systematic search of articles; Section 3 presents a quantitative analysis of the search results; in Section 4, a qualitative analysis of the reviewed surveys is made and a more concise taxonomy is proposed; in Section 5, the analysis is employed to understand what are the XAI methods that are currently being applied in the financial sector; finally, in Section 6, conclusions are drawn along with a critical discussion of this work's contributions, as well as the limitations of this study.

II. METHODOLOGY

The search for relevant scientific papers follows the PRISMA methodology for systematic literature reviews [4]. The general methodology has been adapted for this paper's aims, exclusion criteria, and search engines used. Two distinct searches have been performed: the first served to seek a definition of XAI and to understand the different characteristics and implications of this ML area. The second one is a systematic search for practical finance applications of XAI methods to bring to light current trends in XAI methods within the financial sector.

A. SEARCH FOR EXISTING LITERATURE SURVEYS ON XAI

In search of surveys and literature reviews, the SCOPUS citation database was chosen, as it is more restrictive than Google Scholar or other engines which do not have a validation component. The query used breaks down into two search elements: first, the definition of the area of study; second, the filter for surveys or literature reviews:

TITLE ("Explainable Artificial Intelligence" OR "Explainable AI" OR "XAI" OR "Interpretable Artificial Intelligence") AND TITLE ("Systematic Review" OR "Review" OR "Survey")

This search was performed without specifying the domain of applications. This is deemed as not relevant as the purpose of this search is to get a general view of the definition of XAI as well as a clear specification of the methods' categories. As such, the results seen in Fig. 1 reflect works that are either generic in nature or applied specifically for a type of data (i.e., tabular data).

FIGURE 1. PRISMA methodology for surveys

After obtaining a batch of original 68 results, an exclusion filter was applied to keywords and abstracts, with the purpose of only including papers focusing on XAI and without any specificity regarding subject areas, resulting in 20 papers for revision, of which one was found to be inaccessible. Finally, two criteria were established to exclude further papers in case that XAI was not covered in depth or if it was not the paper's focus, resulting in a final count of 15 documents to be reviewed. Two additional papers were retrieved from a manual search, which increased the total number of surveys to seventeen.

The final list of documents contains 17 surveys whose core concept relates to XAI. These results have been used in Section 4 to sustain the proposal of a taxonomy.

B. SEARCH FOR PRACTICAL IMPLEMENTATIONS OF XAI METHODS

As in the previous search, the SCOPUS citation database was chosen as the data source. We needed to define the essential terms for searching for papers on XAI while differentiating between more generic and domain-free approaches and specific applications to finance. Building on previous knowledge, notably of the XAI concept, as well as works

[2] 2016/679 GDPR Regulation: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

describing specific implementations [5], the following domain-free query was constructed, comprising two parts - explainable artificial intelligence and based and generic or specific implementations of XAI methods:

*(TITLE-ABS-KEY("Explainable Artificial Intelligence" OR "Explainable AI" OR xai)) AND (TITLE-ABS-KEY(counterfactual OR *explanation* OR lime OR "Local Surrogate" OR anchors OR "Individual Conditional Expectation" OR ice OR "Accumulated Local Effects" OR ale OR clear OR "Counterfactual Local Explanations for any classifier" OR dice OR permuteattack OR lore OR "Local Rule-Based Explanations" OR dale OR "Differential Accumulated Local Effects" OR pdp OR "Partial Dependence Plot" OR intrees OR treeexplainer OR shap OR "shapley additive explanation" OR "difference net"))*

The 1984 papers obtained show that XAI has gained some traction over the current years. The following search term was added to the query to filter out all papers not related to Finance:

AND (TITLE-ABS-KEY(financ OR loan OR market OR credit))*

As shown in Fig. 2, from the first batch of 1984 papers, 1855 of which are not subject-specific and were excluded by the automatic filter applied based on the financial domain keywords. For the remaining 129 papers, a manual analysis of title, abstract, and keywords was made only to include papers on the subject area of Finance and with focus on XAI, excluding 60 papers. This exclusion rendered 69 documents, from which nine were directly excluded due to their unavailability. After analyzing the contents of the 60 accessible documents, another filter was applied. This filter excluded documents that did not specify the XAI method used or were unrelated to a practical application of XAI. A final manual analysis concluded if a paper was unrelated to or not in the area of Finance, namely relevant for credit risk and business failure prediction acted as a final criterion for exclusion. In total, only 27 papers were found to obey the inclusion criteria and selected for deeper analysis.

These final papers mainly focus on practical applications of XAI methods in the financial domain, although there were some that did not quite fit into this category as they addressed the legal domain. This domain has gained traction in recent years, notably with the wider adoption of GDPR, resulting in these papers being considered important and thus included and mentioned in Section 1.

FIGURE 2. PRISMA methodology for practical applications

Note that some documents were obtained manually. Fifteen of those were obtained from multiple sources, such as ArXiv[3], Springer[4], and IEEE Explorer[5] databases, with an emphasis given to ArXiv due to its characteristic of hosting very recent studies, which allows it to be on par with the current state of specific implementations of XAI methods. In fact, XAI is an area with an increasing and recent trend in interest [10], and with studies being published rapidly, ArXiv enables to know what investigators are working on without waiting for the peer review process. Finally, after analyzing the respective papers based on their content, these resulted in a final list of six out of 15 documents.

The final list of papers contains 33 documents: 27 resulting from the systematic search and six from the manual search.

III. SEARCH RESULTS

The result of both searches totaled 2069 papers. This section presents an analysis to characterize the rising popularity of XAI as a field of study.

We start with a visual analysis produced with the help of the VOSViewer[6] software. The list of results of both the search for practical applications and for surveys in the SCOPUS database were combined and passed through VOSViewer by filtering out the most frequent and distinct keywords. This resulted in a co-occurrence network of keywords that can be observed in Fig. 3. In the graph visualization is easily perceived the expected closeness between XAI and Artificial Intelligence, interpretability, and decision-making. Furthermore, it is possible to identify, not only the connections with these research areas but also the different implementations of XAI methods, such as SHAP, LIME, Decision Trees, and counterfactual methods. As for application areas, the Health domain is highlighted, with connections to nodes such as medical imaging, diagnosis and diseases.

FIGURE 3. Co-occurrence of keywords

TABLE I
DIACHRONIC OVERVIEW OF PAPERS ON XAI

Year	No. of publications
Pre-2018	10
2018	55
2019	153
2020	318
2021	695
2022	828
2023 (5th January)	10
Total:	2069

Table I presents chronological information regarding the publishing years of the papers, showing a definite rise of popularity in recent years: while until 2018, only ten papers regarding XAI were found, as many as the ones that had already been published over the first five days of January 2023. From then on, there has been a stable increase in the number of published papers related to XAI, with each new year approximately doubling the number of publications from the previous year. Interestingly, one of the surveys, Adadi and Berrada's [6], published in 2018, had been cited 999 times at the time of the search (January 5, 2023), which

[3] https://arxiv.org/
[4] https://link.springer.com/
[5] https://ieeexplore.ieee.org/Xplore/home.jsp
[6] https://www.vosviewer.com/

helps point to 2018 as a turning point for the popularity of XAI in general.

TABLE II
TYPES OF PAPERS PUBLISHED

Type	Count
Article	782
Book Chapter	44
Conference Paper	1125
Conference Review	41

The type of document is important as, typically, more importance is given to scientific articles than conference papers due to the greater difficulty in publishing the former. Table II presents the type of documents obtained, and it is possible to observe that below half the papers are articles, the majority are conference papers, which stresses the recent and developing interest in the theme. Still, the number of articles is deemed sufficiently large for this analysis.

TABLE III
MAIN SUBJECT AREA OF THE JOURNAL THE PAPER IS IN

Domain	Subject area	Count
Health Sciences	Medicine	62
	Health Professions	10
	Dentistry	1
Life Sciences	Biochemistry, Genetics, and Molecular Biology	25
	Immunology and Microbiology	3
	Neuroscience	11
	Agricultural and Biological Sciences	10
Physical Sciences	Computer Science	235
	Engineering	127
	Materials Science	21
	Physics and Astronomy	19
	Chemical Engineering	12
	Mathematics	56
	Chemistry	13
	Environmental Science	21
	Energy	15
	Earth and Planetary Sciences	15
Social Sciences	Social Sciences	157
	Arts and Humanities	18
	Psychology	10
	Decision Sciences	24
	Business, Management, and Accounting	15
	Economics, Econometrics, and Finance	7
	General	8
	Total:	895

To determine the journal's subject area, we used the SCOPUS Journal List[7], which encompassed 43014 journals at the time of the access (November 15, 2022). There were only 527 conference proceedings with corresponding subject areas, so we decided to use only the journal's subject areas. Table III presents the subject area and the domains given by the SCOPUS Journal List and displays the counts of papers found by subject area contained in one of the four domains found in the papers: Health Sciences, Life Sciences, Physical Sciences, and Social Sciences. For these subject areas, albeit the fact that (i) this analysis specifically searches for scientific papers and (ii) the focus is on the area of Finance might limit the perspective on other areas, we can still infer that most of papers arise in the domains of Computer Science, Social Sciences, and Engineering. This is hardly a surprise since these are areas closely related to XAI, especially Computer Science. Other very relevant areas are those of Medicine and Mathematics, where the need for an explanation for any automated decision is most important. While the number of papers classified in these subjects is much less than for the former areas (62 for Medicine and 56 papers for Mathematics), the quantities are still expressive. As for journals in Economics, Accounting, and Finance, a few journals do present papers on this subject (22 papers in total), suggesting that these areas are not yet explored in-depth or, which is common, use AI techniques that are explanatory by default.

IV. LITERATURE REVIEW AND ANALYSIS OF EXISTING SURVEYS

Adadi *et al.* raised the need for XAI for several reasons: the need for ML models to comply with existing legislation to provide a better comprehension of developed systems, which in turn gives a better insight into the flaws or vulnerabilities of such systems [6]. The authors also propose explainability to make model improvements easier and because of the explicit need for an explanation since it helps to extract knowledge. Five domains of application are highlighted, including Finance. The paper also presents a detailed taxonomy for characterizing XAI methods, concluding that this area is still in need of further research work.

A historical perspective of XAI is the focus of Angelov *et al.*, which also detail several XAI methods, which are categorized based on their taxonomy proposal [7]. The authors also describe several key applications for XAI, ranging from the criminal justice system to fraud detection. They conclude with three main points: the importance of the area, how to fill the gap between Deep Learning and Neuroscience with XAI, and finally, future directions for work.

Islam *et al.* present a systematic review that identifies specific domains and applications of XAI methods based on 137 reviewed papers [8]. From these, only three are found to be in the financial domain. The authors conclude with the proposal of a taxonomy for XAI techniques that, albeit new, is largely influenced by the work of other authors. A similar

[7] https://www.scopus.com/sources.uri

study, which also classified papers in specific domains and applications, analyzed and classified 350 papers based on these authors' taxonomy proposal, that has been created based on the analysis of literature and of previously proposed classification systems [9]. Linardatos *et al.* also propose a taxonomy built upon previous work, emphasizing the application of XAI methods to specific areas of AI and reviewing several techniques, some specifically for Deep Learning, while others with a more general approach, including white-box XAI methods [10].

Minh *et al.* focus on a review of the theoretical background for XAI [11]. Each paper is categorized in terms of the type of explanation provided, and the advantages and disadvantages of each of the XAI approaches described is discussed. The authors also propose a taxonomy to classify the papers, where the categories are independent between themselves [11].

In [12], the authors explore an in-depth review of specific implementations and respective categorization along with some practical applications based on the justifications raised previously in [6]. Finally, the authors discuss the practical applications of XAI per domain, current limitations, and future work for this area.

Lin *et al.* introduce a hierarchical taxonomy, focusing on XAI approaches with emphasis in Deep Learning. The authors also raise some issues, namely the trade-off between model interpretability and performance when using Deep Learning [13].

The definition of a taxonomy for XAI methods, largely adapted from other papers and with several categories which include but are not limited to the domain of application of the method is presented in [14]. After a review of previous work, another paper with focus on the definition of a proper taxonomy of XAI methods concludes with a proposal for a taxonomy trying to adapt the taxonomies found in their review [15].

Darias *et al.* [16] perform an analysis of XAI methods libraries and compare each one of the approaches found. The authors' focus is on how each of the XAI methods generates explanations and not how they fit in a taxonomy, hence its exclusion from Table IV. In a systematic literature review the authors systematically analyze papers looking for ways to tackle the problem of cognitive bias or the "systematic error in judgment and decision-making common to all human beings" (as defined in [21]) that has been found in XAI methods used in decision-making systems [17]. While the authors do not provide a taxonomy for XAI methods, it is a relevant paper that helps understand how we use and trust XAI methods. An exploration of the ethical principles of XAI can be found in [18], with focus on reviewing current methods used in the area and providing a taxonomy for these based on previous works. In one other survey, Stepin *et al.* discuss contrastive and counterfactual explanations and propose a taxonomy for these methods [19]. Finally, Lopes *et al.* created a taxonomy, not for XAI methods but rather for the evaluation of such methods [20].

Based on the reviewed literature, we can conclude that no standard categorization of XAI methods still exists. This opinion is supported by Vilone and Longo that, in 2020, with basis on an extensive search, conclude that no proper definition of what an explanation in ML is exists and that the task of having a formalization of XAI is a complex one due to the cross-domain applicability of XAI [9]. This disagreement in achieving an unified taxonomy comes from comparing the approaches of Islam *et al.* [8] and Molnar [22], where the former proposes four main categories, while the latter suggests only three. Nevertheless, two of the categories considered are shared in both approaches.

TABLE IV
PROPOSED TAXONOMY

Category for XAI methods	Works who support the category definition
Stage	[6], [8]–[11], [13], [15], [22], [23]
Model	[6], [8], [10], [12], [13], [15], [22], [23]
Scope	[6], [8], [10], [14], [15], [22]

In the remainder of this section, we will analyze the findings in the literature directly related to a categorization of XAI methods in terms of supporting an integrative taxonomy. Considering only the most relevant and more frequent categories found, we propose three main categories: Stage, Model, and Scope. Table IV shows these categories along with the works that fully support this division, thus excluding, for instance, the approach found in [18], where the authors contemplate only model-agnostic methods and not model-specific ones. The summary table, Table IV, helps to strengthen the argument for a more straightforward and concise taxonomy.

A "post-hoc" XAI method is named after the fact that it acts after predictions are made, not knowing how the predictor model made its decisions (e.g., LIME ([24]). It is a surrogate model since it tries to simplify the function of the black-box model by sampling, perturbing data, and weighing the distance between instances to generate an approximation of the black-box model. "ante-hoc" techniques, such as Decision Trees, and more specifically, the CART technique ([25]) as used in ML, derive their explainability from their clear approach and logic: a tree where an internal node (attribute) is split based on a specific condition. While the complexity of such a model can become large, thus suffering in terms of interpretability by displaying many nodes and depth, it is always possible to inspect the first levels where the most relevant decisions are made.

These findings suggest our first category, Stage, that indicates if the method is used after the prediction is made - post-hoc - or if the XAI model is intrinsically explainable -

ante-hoc. We find evidence for this category in references [6], [8]–[11], [13], [15], [22], [23].

Some works do not make this distinction clearly, as is the case with the approach found in [18], where intrinsically explainable methods are detailed, such as Linear Regression or kNN, a technique initially proposed by Hodges and Fix [26] and since then widely used in Machine Learning, but post-hoc methods are not presented in the same capacity. The authors conclude that Linear Regression and kNN methods can be applied to complex problems but are inadequate for understanding ML models [26]. Barredo Arrieta *et al.* [23] define a taxonomy based on the reviewed literature. Contrary to taxonomies on previously mentioned works, where no general order of importance is mentioned, this work presents a hierarchical structure. The first level of the taxonomy tree, with ante-hoc models being referred to as "Transparent Models" and post-hoc models as "Post-Hoc Explainability," can be encompassed into the Stage category.

The second category that we propose is Model, referring to whether an XAI method is defined for a single or restricted group of models, that is, if it is model-specific, or if the method can be applied generally to any predictive model, that is, is model-agnostic. Evidence for this category can be found in references [6], [8], [10], [12], [13], [15], [22], [23].

Model-specific techniques tend to be the most well-known and established models, like in the case of Decision Trees. The intrinsic explainability of this model is one of its downsides since, when compared with the performance of a neural network, it may leave a lot to be desired. While model-specific methods can be great as they have the unique ability to access the predictive model's internals, they suffer greatly in terms of interoperability due to their lack of adaptation for a more general usage [12].

Model-agnostic methods, such as LIME [24], are the opposite. Its general purpose makes it suitable for any predictive model, as shown by the authors, that present examples of explanations of predictive models, such as SVM (as defined in [27]) for text classification. We can find evidence for Model as a category in [9] and [15], where this categorization is proposed as being a subcategory of the type post-hoc category. However, for Linardatos *et al.* [10], this category is named "Model Specific vs. Model Agnostic" and is presented in a non-hierarchical taxonomy. The same is seen in the work of Molnar [22] and Sahakyan [12], named "Model-specific or Model-agnostic." On a different approach, the authors of [18] only explored model-agnostic approaches and not model-specific ones.

The final proposal for a category for XAI methods is Scope, intending to separate XAI methods on whether they are used to help understand the general behavior of the model, that is, if these techniques provide global interpretability or if they try to explain singular or a limited group of instances of data, that is, local interpretability [6]. This category is largely accepted within the reviewed literature, where it is found as a main category for classifying XAI methods [6], [8], [10], [14], [15], [22].

Local interpretability encapsulates methods such as LIME, that introduces explainability by choosing relevant features, along with the features' respective importance, for a subset of the data to help understand singular instances of data. Global interpretability techniques focus on explaining the behavior of the model. One such example is SHAP ([28]), that returns a graphical importance of the used features [22]. In some of the works found only local explanations are mentioned, like in the example of [23], or where an XAI taxonomy is explicitly stated and Scope is considered as being a sub-class of the model-agnostic class [11], [13], [18].

The three previous categories - Stage, Model, and Scope - were presented based on what the relevant literature shows as most generally used for the reviewed taxonomies for the categorization of XAI methods. Nonetheless, there are a couple more relevant categories to discuss, as they might be studied more in-depth by other authors, thus gaining the relevancy necessary to become a main category in the near future.

Molnar [22] points out "Result" as a category, where importance is given to how the output of the XAI method is categorized. The author points out several possible sub-classes, from feature summary statistics and feature importance to data points. This category is a contender for relevancy when defining a taxonomy, as other authors support this category even if under different names [15]. Another work favoring the categorization of results is [14], although this category is named "Presentation Format," showing two sub-classes on whether the generated explanation is textual (when explanations are generated using natural language techniques), or visual, focusing on providing a visual explanation, for example, via graphs or images. We can also find Result among other categories mentioned in [23].

Some authors consider "Output Format" as a proper category for XAI methods. This classification is somewhat similar to the Result category, but in [15], we can find a difference between these two: while the Result class categorizes the explanation about the type of result provided, the Output Format looks at whether the explanation is of a particular type of data, such as numeric, textual, visual, among others. Such a difference is deemed relevant to define the purpose of the explanation for the different stakeholders, i.e., to whom the explanation is intended [8], [9].

In [10], the category "Purposes of Interpretability" is defined as "the purpose that these methods were created to serve and the ways through which they accomplish this purpose." The authors propose four subcategories within Purposes of Interpretability. Two of these categories, intrinsic and post-hoc, serve as references to the category Stage as previously stated in this section. However, in this case, these categories are inserted as sub-classes in the 'Purpose' category to explain complex black-box models, or

post-hoc purpose, and to create white-box models, following an ante-hoc or intrinsic purpose. However, another author separates this purpose into two sub-classes, one for explaining how something works and another for explaining why something happened [14].

Other categories try to include stakeholders, i.e., to whom the explanation will serve. Hu *et al.* mention three types of users: developers, the ones who build the algorithm; observers, typically those who examine the system in place; and finally, end-users, people who are affected by the systems' results [14]. Another category proposed by the same authors is "Domain," which defines the subject area or domain for which XAI explanations are generated. Yet another category, 'Functioning,' is referred to by the authors of [15] to categorize how information is extracted from ML models. For instance, some XAI methods focus on perturbations of the data to gain insights for their explanatory process. In contrast, others focus on leveraging structures, which tend to result in feature importance attributes, among other sub-classes.

One last emerging category is "Type of Problem," which defines for what purposes the XAI method is useful to cover (classification or regression problems) and can be found in [9], [15].

V. LITERATURE REVIEW AND ANALYSIS OF PRACTICAL APPLICATIONS

XAI methods have gained much traction over the past few years as depicted in Table I. This section will explore findings related to applications of XAI restricted to the financial sector, with special emphasis on credit-related problems and fraud detection. However, the latter is significantly less explored, as remarked earlier in Section 3. The following section presents the specific applications of XAI methods in the financial domain that have been found in our search, starting with a brief description of these methods and presenting a table summarizing the XAI method with examples of applications.

In general, SHAP tends to be one of the most widely used XAI methods for this domain. SHAP is a model-agnostic technique which has the possibility of providing explanations both on a local, and on a global scope. Although we can find slight differences with how it is implemented, with a mixture of studying feature importance with clustering and decision trees [29] or a simple application of the method on predictions [30], [31]. Some works follow a more complex approach, with a detailed procedure on how the treatment of data is made along with the phases related to the prediction/explanation, culminating in explanations given by a sequence of steps, like feature selection followed by clustering [32]. One approach combines counterfactual explanations with SHAP [41]. The feature importance provided by SHAP is used to provide counterfactual explanations in a localized region in the data, resulting in a more detailed explanation than by simply using either method independently. This method is model-agnostic and works on the local scope.

SHAP is not the only popular method used, with LIME also being a popular choice. Both methods differ in the Scope category, as SHAP is mostly used globally, while LIME tends to be used locally. Overall, the value in the explanations of SHAP and LIME comes in the form of feature importance, where calculations are made to determine the weight in contribution that features bear for the prediction process. Some of the articles mentioned employing both these XAI methods to explain the models used [37], [38]. In summary, LIME is a model-agnostic approach which presents explanations on a local scope.

While SHAP and LIME employ explanations in the form of feature importance, counterfactual methods create explanations for predictive models through the generation of what-if examples where certain feature values are changed to alter the predicted result [22]. Regarding counterfactual methods, PermuteAttack was found in the manual search for practical applications [5]. This method consists in using a genetic algorithm that perturbs data by changing randomly selected features and goes through an optimization process to find an instance with the least number of permuted features, resulting in a counterfactual explanation. Another counterfactual method was found in reference [54], where a genetic algorithm is also implemented to produce explanations. As for the optimization process, it works only with features showing a correlation with the targe, and for each iteration, the distance between the counterfactual example and the original instance is constrained. The explanation come in the form of visual explanations, showing what features needed changes to alter the prediction. PermuteAttack is a model-agnostic approach and provides explanations on a local scope.

One widely used method for explainability is Partial Dependence Plots or PDP [34], which helps interpret how one feature affects another. This aids in the explanation for the target feature, where the visual representation of this plot makes this relationship more understandable. PDP can be implemented regardless of the predictive model used and provides explanations in a global scope.

Two other methods are PASTLE [49] and CASTLE [50], created by the same authors. The first method introduces explainability through the reduction of the sample space into pivots or points, while the second identifies clusters in the data that have common behavior and classification, finalizing in the extraction of rule-based explanations. Both methods are model-agnostic and provide explanations on a local basis.

Anchors [51] is a model-agnostic method which provides explanations on a local scope by calculating the set of predicates or rules that are most relevant for the predictive outcome. It is an iterative process, starting with a general approach and finalizing in a filtered set of the most relevant rules presented as if-then clauses.

Specifically for deep-learning methods, MANE [52] works by processing features to extract cross features, and where linear regression is then applied to approximate the nonlinear decision boundary or the curve that separates two classes of data. This aids in the understanding of behavioral patterns of the instances of data, resulting in a model-agnostic method on a local scope.

There are two model specific approaches, that of LTreeX [56] and inTrees [57]. LTreeX is a local method and creates a surrogate model directly from the Random Forest, resulting in the presentations of rules that explain the outcome of any given instance. inTrees, on the other hand, is a global method that provides explanations by extracting rules from tree ensembles such as Random Forests or boosted trees.

DALE [58] is an XAI method which makes the calculations made by Area of Local Effects or feasible through an approximation of ALE. Similarly to the explanations provided by PDP, DALE's explanations come in the form of plots where it is possible to see the effect a feature has on the target. DALE is a model-agnostic technique and presents explanations in a global scope.

The final method for XAI found in the literature review is a model-agnostic approach where TREPAN trees are combined with neural networks to explain localized instances [46]. After clustering the data using a neural network, TREPAN is applied to build decision trees on a cluster level, resulting in explanations of the target feature by sets of rules defined by the trees. This hybrid model works on any predictive model and locally in terms of its scope.

Table V summarizes the XAI methods found, along with their respective categorization based on the taxonomy defined in Section 4. One obvious conclusion is that all methods being used in the financial domain are post-hoc, with their explanations being formed after the predictions have been made. However, it is important to point out this distinction since XAI methods exist that do not work on a post-hoc basis, such as the Decision Trees, in which the method is not only explanatory in how decisions are made for the prediction process but the method itself predicts the outcome in question. Therefore, these methods are intrinsically explanatory in nature, thus, are ante-hoc methods and our searches only targeted post-hoc explanability.

TABLE V
CATEGORIZATION OF XAI METHODS

XAI Method	Author	Stage	Model	Scope
SHAP	[28]	Post-hoc	Agnostic	Global/Local
LIME	[24]	Post-hoc	Agnostic	Local
Counterfactuals	[5], [54]	Post-hoc	Agnostic	Local
PDP	[34]	Post-hoc	Agnostic	Global
PASTLE	[49]	Post-hoc	Agnostic	Local
CASTLE	[50]	Post-hoc	Agnostic	Local
Anchors	[51]	Post-hoc	Agnostic	Local
MANE	[52]	Post-hoc	Agnostic	Local
LTreeX	[56]	Post-hoc	Specific	Local
inTrees	[57]	Post-hoc	Specific	Global
DALE	[58]	Post-hoc	Agnostic	Global
Rational Shapley Values	[41]	Post-hoc	Agnostic	Local
TREPAN/Hidden-layer-clustering	[46]	Post-hoc	Agnostic	Local

Next, we present a description of the practical applications that have been found in the literature review.

Hastie *et al.* [33] introduced explainability for the prediction of financial distress through XAI methods such as SHAP, PDP [34], and Counterfactuals [22]. On another work, using a dataset containing data from Chinese companies, Zhang *et al.* introduce Counterfactuals on the three most important features, analyzed via SHAP, where the specific instance of data has its features values changed. Through a cyclical prediction process, a check is made on the variation prediction to see if its result has changed [35]. Some other works focus on explainability by combining LIME and SHAP applied to predictive models, such as Random Forests and XGBoost. Mandeep *et al.* worked with a dataset from Yahoo Finance companies' shares, filtered for the most relevant companies [36]. The authors combined the excellent predictive performance with intuitive explanations from LIME and SHAP to support the predictions results. Park et al. [39] investigated reliable prediction explanations for the predictive model built using XGBoost applied to a Korean companies' dataset containing 110 features. For evaluating the reliability of LIME, they analyzed, instance by instance, the number of features present for the top ten most important instances when LIME was applied globally to the entire dataset. Another application involving the use of LIME for the explanation of the predictions made by a Multi-Layer Perceptron on a transactions dataset can be found in [40].

In the work of Watson, Rational Shapley Values were introduced [41]. This hybrid method uses Shapley values and Counterfactuals, built to reap the benefit from both methods. The process was tested using the German Credit dataset.

On a different note, Hadash *et al.* focused on improving current implementations of LIME and SHAP methods [42] with an experiment performed on a credit dataset where they used 133 users to evaluate the transformations. The improvements focused primarily on semantic changes to make the explanations given by these methods more understandable [42].

Another application proposes the implementation of 2DCNN (Convolutional Neural Networks), typically used for image-related problems, to tabular data. This process partitions the German Credit Dataset into bins, which are then used to create images. Based on these images, the model made its predictions. Subsequently, LIME and SHAP were used to explain such predictions, where the authors

determined that SHAP performance was superior to the one obtained with LIME [43].

Analyzing some more general applications of XAI to Finance, we can find an approach that applied SHAP for the explainability of the model to determine what were the most used features and using loan data that was reviewed using NLP (Natural Language Processing) techniques [44].

De *et al.* proposed the combination of TREPAN [45] and hidden-layer clustering to explain predictions made using a credit dataset and for the predictive goal of determining a default in payment. This method was compared with LIME, and the authors concluded that the TREPAN model outperforms LIME [46].

Huynh *et al.* focused on implementing a framework to answer questions mainly motivated by legal regulations such as GDPR [47]. One of the inquiries relates to the fact that the final decision is "reached solely via automated means," which helps determine whether Article 22 of the GDPR is applicable. The authors worked on a loan scenario, concluding that their developed framework successfully answered eight out of the 13 questions which explain the decisions in the loan scenario, encompassing individual concerns or the individual data subject, and institutional concerns or the data controller. While the process of selection of features is clear and explained, one of the limitations of this paper is that the ML algorithm itself is not explained [47].

Chromik implements SHAP onto the predictive model XGBoost to create an interface for personal loan applications [48]. This experiment shows mixed results when tested through user queries, with the users finding the interface overwhelming due to the presentation of several types of explanations calculated through SHAP. However, by complementing the experiment through several and different elements, it was possible to determine that the explanations were detailed enough to understand the system's behavior in a prediction scenario.

A novel XAI method, PASTLE, was introduced by Gatta *et al.* and used to decrease the dataset to the points representing regions where the predictive model behaves differently. As for the data used, many experiments were performed, including the use of a financial dataset [49]. The same authors also developed another XAI method, CASTLE, whose main difference is what is used to decrease the number of instances used: while the first method uses pivots, the new method utilizes clustering [50]. When compared to Anchors [51], the authors found it less taxing on computational resources.

While the applications seen so far are primarily model-agnostic, the authors of [52] propose an XAI method specifically for deep learning models called MANE. Using a dataset of private transactions, they evaluated the proposed method against LIME, concluding that the performance was similar for both, albeit with a small difference when compared with the proposed approach. When testing the fidelity, i.e., the degree of correctness of selected features of MANE, only five features were used to create the explanations contrasting with LIME, which needed 25 features for the same goal [52].

Lesser-known approaches use Feature Importance and Partial Dependence Plots to improve the interpretability of the predictive model, like in the case of XGBoost [53]. In another approach, the authors utilize an XAI method they developed to create counterfactual explanations, resulting in a low number of features needed to change the given outcome [54]. In [55], using the Home Equity Line of Credit (HELOC) dataset, the authors extended Shapley Values to mixed features without assuming them to be independent, concluding that no model outperformed the others.

Dedja *et al.* implemented another method, LTreeX, testing it over several datasets, although none was described as a financial dataset [56]. Nonetheless, this very recent approach deserves to be evaluated for possible implementation in the financial domain since the value of the explanation comes from the summarization of Random Forests, a common technique employed in modeling. In this regard, Deng explains Random Forests and Boosted Trees by expanding on known methods such as Area of Local Effects ([58]), even if not for the specific area of Finance [57]. Within the related literature, we can also encounter different implementations of counterfactuals [59]–[61] and a similar approach presenting a combination of Linear Regression and Neural Networks in order to explain the predictions [62].

The summary table below (Table VI) describes the most predominant XAI methods emerging from this literature review and ordered by descending popularity. SHAP is by far the most popular method, being referred to in most of the works here reviewed. The novel approaches, such as the LTreeX defined in [56], are placed in the category 'Others' that encompasses several more recent and thus less used methods.

TABLE VI
RECENT APPLICATIONS OF XAI METHODS

XAI Method	Works that make use of the method
SHAP	[29-33, 35-38, 41, 43, 48, 55]
LIME	[36-40, 43, 46, 52]
Counterfactuals	[33, 35, 41, 59-61]
Hybrid models	[41, 45]
CERTIFAI	[53, 54]
Others	[33, 49, 50-52, 56, 57]

Finally, when reviewing related work, it is important to discuss the datasets used. All the datasets found are from the financial domain, but only a handful are publicly available. One of the publicly available datasets is the German Credit[8] dataset, which contains 21 features and 1000 instances. This dataset encompasses financial information of clients and used

[8] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

to predict the risk posed when credit is granted [32, 41, 43, 54, 57, 59, 61]. Another public dataset is the Default of Credit Card Clients in Taiwan[9], containing 30,000 samples (customers) and information on 25 variables related to credit, default, billing, payments, and demographic factors [46, 60, 62]. Three more datasets were found that were publicly available. The first dataset[10] was used in [59, 61] and has the goal of predicting whether a person makes over $50,000 a year, containing 14 features and 48,842 records. The second[11] is an anonymized credit card transactions' dataset where the target is to determine whether a transaction is legitimate or not. This second dataset was used in [37, 40], has 31 features in total and 284,807 transactions. Finally, the third dataset was found only in [62] and has the goal of determining the probability that a person will experience financial distress in the next two years. This dataset contains 11 features and totals 251,503 records.

VI. CONCLUSION

The present analysis presents a review of existing literature on the application of XAI methods with focus on works pertaining the financial domain. First, a search was made exclusively for surveys relating with XAI. A second search was performed to discover practical applications of XAI specifically for finances. From the data obtained with both searches, we were able to point out what are the major categories for XAI methods. This research results in the proposal of a simple taxonomy that resumes the main characteristics of known methods of explainability. While deemed adequate and based in a significant part of the reviewed literature, the proposed holistic taxonomy is yet subject to change, given the fast pace of progress in this area.

In a second contribution, we present the methods that are used to achieve explainability for models applied in the financial sector. The existing literature seems to favor SHAP and LIME as the preferred explainability methods. The applications found demonstrate that different methods can be employed simultaneously, helped by the fact that the generality of the XAI techniques here reviewed are applied post-hoc, thus providing the ability to function independently and be used together. Though the popularity of LIME and SHAP in this domain seems to prevail, numerous new approaches are being proposed, broadening the spectrum of XAI methods available, from counterfactual explanations to partial dependence plots or more novel approaches which repurpose techniques used in image classification for tabular data.

This work reflects the current understanding of the state-of-the-art regarding XAI methods in financial applications and presents a solid proposal for categorizing the existing XAI methods.

Due to the recent rise in the search for explainable methods for artificial intelligence applications, it is expected that new developments will be arising in the near future, paving the way for new anthological descriptive research to emerge.

REFERENCES

[1] M. Attaran and P. Deb, "Machine Learning: The New 'Big Thing' for Competitive Advantage," International Journal of Knowledge Engineering and Data Mining, vol. 5, no. 1, p. 1, 2018, doi: 10.1504/IJKEDM.2018.10015621.

[2] M. Turek, "Explainable Artificial Intelligence (XAI)," https://www.darpa.mil/program/explainable-artificial-intelligence.

[3] Official Journal of the European Union, "Regulation 2016/679," The protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016. Accessed: December 16, 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[4] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, p. n71, Mar. 2021, doi: 10.1136/bmj.n71.

[5] M. Hashemi and A. Fathi, "PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards," Aug. 2020.

[6] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[7] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," WIREs Data Mining and Knowledge Discovery, vol. 11, no. 5, Sep. 2021, doi: 10.1002/widm.1424.

[8] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks," Applied Sciences, vol. 12, no. 3, p. 1353, Jan. 2022, doi: 10.3390/app12031353.

[9] G. Vilone and L. Longo, "Explainable Artificial Intelligence: a Systematic Review," May 2020.

[10] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[11] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," Artif Intell Rev, vol. 55, no. 5, pp. 3503–3568, Jun. 2022, doi: 10.1007/s10462-021-10088-y.

[12] M. Sahakyan, Z. Aung, and T. Rahwan, "Explainable Artificial Intelligence for Tabular Data: A Survey," IEEE Access, vol. 9, pp. 135392–135422, 2021, doi: 10.1109/ACCESS.2021.3116481.

[13] K.-Y. Lin, Y. Liu, L. Li, and R. Dou, "A Review of Explainable Artificial Intelligence," pp. 574–584, 2021. doi: 10.1007/978-3-030-85910-7_61.

[14] Z. F. Hu, T. Kuflik, I. G. Mocanu, S. Najafian, and A. Shulner Tal, "Recent Studies of XAI - Review," in Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation, and Personalization, pp. 421–431, Jun. 2021. doi: 10.1145/3450614.3463354.

[15] T. Speith, "A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods," in 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 2239–2250, Jun. 2022. doi: 10.1145/3531146.3534639.

[16] J. M. Darias, B. Díaz-Agudo, and J. A. Recio-Garcia, "A systematic review on model-agnostic XAI libraries," in Workshops for the 29th International Conference on Case-Based Reasoning, ICCBR-WS 2021, pp. 28–29, Sep. 2021.

[17] A. Bertrand, R. Belloum, J. R. Eagan, and W. Maxwell, "How Cognitive Biases Affect XAI-assisted Decision-making," in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 78–91, Jul. 2022. doi: 10.1145/3514094.3534164.

[9] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[10] https://archive.ics.uci.edu/dataset/2/adult
[11] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[18] A. Hanif, X. Zhang, and S. Wood, "A Survey on Explainable Artificial Intelligence Techniques and Challenges," in 2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW), pp. 81–89, Oct. 2021. doi: 10.1109/EDOCW52865.2021.00036.

[19] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Farina, "A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence," IEEE Access, vol. 9, pp. 11974–12001, 2021, doi: 10.1109/ACCESS.2021.3051315.

[20] P. Lopes, E. Silva, C. Braga, T. Oliveira, and L. Rosado, "XAI Systems Evaluation: A Review of Human and Computer-Centred Methods," Applied Sciences, vol. 12, no. 19, p. 9423, Sep. 2022, doi: 10.3390/app12199423.

[21] D. Kahneman, P. Slovic, A. Tversky, and (eds.), Judgment Under Uncertainty: Heuristics and Biases. Cambridge University Press, 1982.

[22] C. Molnar, Interpretable Machine Learning, 2nd ed. 2022. Accessed: 08/01/2023 [Online]. Available: https://christophm.github.io/interpretable-ml-book

[23] A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "' Why Should I Trust You?'," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144, Aug. 2016. doi: 10.1145/2939672.2939778.

[25] D. H. Moore, "Classification and regression trees, by Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. Brooks/Cole Publishing, Monterey, 1984,358 pages, $27.95," Cytometry, vol. 8, no. 5, pp. 534–535, Sep. 1987, doi: 10.1002/cyto.990080516.

[26] J. L. Hodges and E. Fix, "Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties," Randolph Field, Texas, Feb. 1951.

[27] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[28] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[29] C. Maree, J. E. Modal, and C. W. Omlin, "Towards Responsible AI for Financial Transactions," in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 16–21, Dec. 2020. doi: 10.1109/SSCI47803.2020.9308456.

[30] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," Comput Econ, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.

[31] S. Kim and J. Woo, "Explainable AI framework for the financial rating models," in 2021 10th International Conference on Computing and Pattern Recognition, pp. 252–255, Oct. 2021. doi: 10.1145/3497623.3497664.

[32] J. Chaquet-Ulldemolins, F.-J. Gimeno-Blanes, S. Moral-Rubio, S. Muñoz-Romero, and J.-L. Rojo-Álvarez, "On the Black-Box Challenge for Fraud Detection Using Machine Learning (II): Nonlinear Analysis through Interpretable Autoencoders," Applied Sciences, vol. 12, no. 8, p. 3856, Apr. 2022, doi: 10.3390/app12083856.

[33] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[34] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," The Annals of Statistics, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.

[35] Z. Zhang, C. Wu, S. Qu, and X. Chen, "An explainable artificial intelligence approach for financial distress prediction," Inf Process Manag, vol. 59, no. 4, p. 102988, Jul. 2022, doi: 10.1016/j.ipm.2022.102988.

[36] Mandeep, A. Agarwal, A. Bhatia, A. Malhi, P. Kaler, and H. S. Pannu, "Machine Learning Based Explainable Financial Forecasting," in 2022 4th International Conference on Computer Communication and the Internet (ICCCI), pp. 34–38, Jul. 2022. doi: 10.1109/ICCCI55554.2022.9850272.

[37] I. Ullah, A. Rios, V. Gala, and S. Mckeever, "Explaining Deep Learning Models for Tabular Data Using Layer-Wise Relevance Propagation," Applied Sciences, vol. 12, no. 1, p. 136, Dec. 2021, doi: 10.3390/app12010136.

[38] S. Tyagi, "Analyzing Machine Learning Models for Credit Scoring with Explainable AI and Optimizing Investment Decisions," Sep. 2022.

[39] M. S. Park, H. Son, C. Hyun, and H. J. Hwang, "Explainability of Machine Learning Models for Bankruptcy Prediction," IEEE Access, vol. 9, pp. 124887–124899, 2021, doi: 10.1109/ACCESS.2021.3110270.

[40] T.-Y. Wu and Y.-T. Wang, "Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection," in 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), pp. 25–30, Nov. 2021. doi: 10.1109/TAAI54685.2021.00014.

[41] D. Watson, "Rational Shapley Values," in 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1083–1094, Jun. 2022. doi: 10.1145/3531146.3533170.

[42] S. Hadash, M. C. Willemsen, C. Snijders, and W. A. IJsselsteijn, "Improving understandability of feature contributions in model-agnostic explainable AI tools," in CHI Conference on Human Factors in Computing Systems, pp. 1–9, Apr. 2022. doi: 10.1145/3491102.3517650.

[43] X. Dastile and T. Celik, "Making Deep Learning-Based Predictions for Credit Scoring Explainable," IEEE Access, vol. 9, pp. 50426–50440, 2021, doi: 10.1109/ACCESS.2021.3068854.

[44] A. Stevens, P. Deruyck, Z. van Veldhoven, and J. Vanthienen, "Explainability and Fairness in Machine Learning: Improve Fair End-to-end Lending for Kiva," in 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1241–1248, Dec. 2020. doi: 10.1109/SSCI47803.2020.9308371.

[45] M. Craven and J. Shavlik, "Extracting Tree-Structured Representations of Trained Networks," in Advances in Neural Information Processing Systems, vol. 8, 1995.

[46] T. De, P. Giri, A. Mevawala, R. Nemani, and A. Deo, "Explainable AI: A Hybrid Approach to Generate Human-Interpretable Explanation for Deep Learning Prediction," Procedia Comput Sci, vol. 168, pp. 40–48, 2020, doi: 10.1016/j.procs.2020.02.255.

[47] T. D. Huynh, N. Tsakalakis, A. Helal, S. Stalla-Bourdillon, and L. Moreau, "Addressing Regulatory Requirements on Explanations for Automated Decisions with Provenance—A Case Study," Digital Government: Research and Practice, vol. 2, no. 2, pp. 1–14, Apr. 2021, doi: 10.1145/3436897.

[48] M. Chromik, "Making SHAP Rap: Bridging Local and Global Insights Through Interaction and Narratives," pp. 641–651, 2021. doi: 10.1007/978-3-030-85616-8_37.

[49] V. la Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "PASTLE: Pivot-aided space transformation for local explanations," Pattern Recognit Lett, vol. 149, pp. 67–74, Sep. 2021, doi: 10.1016/j.patrec.2021.05.018.

[50] V. la Gatta, V. Moscato, M. Postiglione, and G. Sperlì, "CASTLE: Cluster-aided space transformation for local explanations," Expert Syst Appl, vol. 179, p. 115045, Oct. 2021, doi: 10.1016/j.eswa.2021.115045.

[51] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11491.

[52] Y. Tian and G. Liu, "MANE: Model-Agnostic Non-linear Explanations for Deep Learning Model," in 2020 IEEE World

Congress on Services (SERVICES), pp. 33–36, Oct. 2020. doi: 10.1109/SERVICES48979.2020.00021.

[53] Y. Zou, C. Gao, and H. Gao, "Business Failure Prediction Based on a Cost-Sensitive Extreme Gradient Boosting Machine," IEEE Access, vol. 10, pp. 42623–42639, 2022, doi: 10.1109/ACCESS.2022.3168857.

[54] X. Dastile, T. Celik, and H. Vandierendonck, "Model-Agnostic Counterfactual Explanations in Credit Scoring," IEEE Access, vol. 10, pp. 69543–69554, 2022, doi: 10.1109/ACCESS.2022.3177783.

[55] A. Redelmeier, M. Jullum, and K. Aas, "Explaining Predictive Models with Mixed Features Using Shapley Values and Conditional Inference Trees," pp. 117–137, 2020. doi: 10.1007/978-3-030-57321-8_7.

[56] K. Dedja, F. K. Nakano, K. Pliakos, and C. Vens, "Explaining random forest prediction through diverse rulesets," Mar. 2022.

[57] H. Deng, "Interpreting tree ensembles with inTrees," Int J Data Sci Anal, vol. 7, no. 4, pp. 277–287, Jun. 2019, doi: 10.1007/s41060-018-0144-8.

[58] V. Gkolemis, T. Dalamagas, and C. Diou, "DALE: Differential Accumulated Local Effects for efficient and accurate global explanations," Oct. 2022.

[59] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, and F. Turini, "Factual and Counterfactual Explanations for Black Box Decision Making," IEEE Intell Syst, vol. 34, no. 6, pp. 14–23, Nov. 2019, doi: 10.1109/MIS.2019.2957223.

[60] A. White and A. d'Avila Garcez, "Measurable Counterfactual Local Explanations for Any Classifier," Aug. 2019.

[61] E. Ç. Mutlu, N. Yousefi, and O. Ozmen Garibay, "Contrastive Counterfactual Fairness in Algorithmic Decision-Making," in Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 499–507, Jul. 2022. doi: 10.1145/3514094.3534143.

[62] D. Chen, W. Ye, and J. Ye, "Interpretable Selective Learning in Credit Risk," Sep. 2022.

[63] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, October 2018, pp. 80–89, doi: 10.1109/DSAA.2018.00018.

[64] Z.C. Lipton, "The mythos of model interpretability," Queue, vol.16, no. 3, pp. 31–57, May-June 2018, doi: 10.1145/3236386.3241340.

[65] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv*, March 2017, doi:10.48550/arXiv.1702.08608.