INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Predictive Modeling of Financed Amount of Personal Loans in the context of Outbound Marketing

Clara Alexandra dos Santos Martins Carvalho Baptista

Master's in Data Science

Supervisor:
PhD, Ana de Almeida, Professora Associada,
ISCTE- Instituto Universitário de Lisboa

Co-Supervisor:
PhD, Luís Nunes, Professor Associado,
ISCTE- Instituto Universitário de Lisboa

October, 2023

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

Predictive Modeling of Financed Amount of Personal Loans in the context of Outbound Marketing

Clara Alexandra dos Santos Martins Carvalho Baptista

Master's in Data Science

Supervisor:
PhD Ana de Almeida, Professora Associada,
ISCTE- Instituto Universitário de Lisboa

Co-Supervisor:
PhD Luís Nunes, Professor Associado,
ISCTE- Instituto Universitário de Lisboa

October, 2023

Acknowledgments

I would like to thank my two supervisors, professors Ana de Almeida and Luís Nunes, for their guidance, always showing compassionate help, and understanding, and providing me with valuable insights throughout this work.

I extend my heartfelt thanks to all my work colleagues for their empathetic support throughout my master's, whose assistance was crucial at the commencement of my thesis, and for the valuable and insightful exchanges that greatly contributed to this research. In particular to my manager whose support and understanding were pivotal from the beginning to the completion of my thesis.

I would like to mention the support that I felt throughout the writing of this thesis of all my friends, which was an important source of motivation.

I am immensely thankful to my family for their unwavering emotional support throughout my academic journey. They not only provided me with valuable advice but also served as a crucial driving force in my endeavors and always helped me when I most needed it.

I am grateful for my boyfriend's emotional support and motivation, and for consistently being there for me when I needed it the most. I deeply appreciate his efforts to ensure my well-being and help overall, which greatly assisted me in completing this research.

I want to express my profound gratitude to my hard-working and compassionate brother, whose sacrifices provided me with the opportunity to pursue my academic aspirations. My deepest thanks for being my safety net. My deepest thanks for being my safety net.

Resumo

A otimização de *direct marketing*, visa maximizar lucros e reduzir custos. Este estudo investigou a previsão dos montantes financiados de empréstimos após a comunicação de marketing, comparando esses resultados com a previsão da resposta em termos de receita em menores amostras de clientes.

Neste estudo foram utilizados dados de campanhas de marketing direto de uma empresa de finanças pessoais entre 2019 e 2021. Foram comparadas técnicas de seleção de variáveis e diferentes algoritmos, incluindo Random Forest, Xgboost e Redes Neuronais, para prever respostas e montantes financiados. Métricas como AUC, precisão, recall, MAE e R^2 foram usadas para variáveis categóricas e numéricas. Dada a raridade de respostas positivas, o estudo abordou o desequilíbrio de dados, aplicando técnicas como SMOTE, subamostragem e uma abordagem de bagging, melhorando consideravelmente as previsões do modelo. A transformação de Box Cox foi utilizada para lidar com assimetria nos montantes financiados, e foram explorados a afinação de hiperparâmetros e funções de perda personalizadas. O estudo destaca o valor adicional de prever o valor do cliente em *marketing* direto, permitindo otimizar a receita quando se pretende reduzir número de contactos. A adequação da abordagem depende de restrições como tempo, recursos, orçamento e volume mínimo de comunicações. Se houver recursos, a previsão do valor do cliente fornece informações valiosas à tomada de decisão, mesmo que não seja usada para limitar o número de comunicações. Esta pesquisa oferece um quadro para otimizar estratégias de telemarketing, equilibrando lucro e eficiência de recursos.

Palavras-Chave: Empréstimos Pessoais; *Marketing* Direto; Montante Financiado; Retorno sobre o Investimento; Redes Neuronais; Dados não balanceados

Abstract

The optimization of direct marketing aims to maximize profits and reduce costs.

This study investigated predicting the financed amounts of loans following marketing communication, comparing these results with response prediction in terms of revenue with smaller customer samples. Data from direct marketing campaigns of a personal finance company between 2019 and 2021 were utilized. Variable selection techniques and algorithms, including Random Forest, Xgboost, and Neural Networks, were compared to predict responses and financed amounts. Metrics such as AUC, precision, recall, MAE, and R-squared were used for both categorical and numerical variables. Due to the rarity of positive responses, the study addressed data imbalance by applying techniques like SMOTE, undersampling, and a bagging approach, significantly enhancing model predictions. The Box Cox transformation was employed to address skewness in financed amounts, and hyperparameter tuning and custom loss functions were explored. The study emphasizes the added value of predicting customer value in direct marketing, enabling revenue optimization when aiming to reduce the number of contacts. The appropriateness of the approach depends on constraints like time, resources, budget, and minimum communication volume. When resources allow, predicting customer value offers valuable insights for decision-making, even if not used to limit the number of communications. This research provides a framework for optimizing telemarketing strategies, balancing profit and resource efficiency.

Keywords: Personal loans; Direct Marketing; Financed Amount; Return on Investment; Neural Networks; Data imbalance

Index

Figure Index

Table Index

Abbreviations

AI – Artificial Intelligence

ANN – Artificial Neural Network

AUC – Area Under the Curve

CLV– Customer Lifetime Value

CNN – Convolutional Neural Networks

CRISP – DM – Cross Industry Standard Process for Data Mining

DCNN – Deep Convolutional Neural Networks

DL – Deep Learning

DT – Decision Tree

FN – False Negatives

FP – False Positives

GRU – Gated Recurrent Unit

K-NN – K-Nearest Neighbours

LR – Logistic Regression

LSTM – Long Short-Term Memory

MAE – Mean Absolute Error

ML – Machine Learning

MLP – Multilayer Perceptron

MLPNN– Multilayer Perceptron Neural Network

NB – Naïve Bayes

NN – Neural Network

PL – Personal Loan

RC – Revolving Card

RF – Random Forest

RFM – recency, frequency and monetary

RMSE – Root Mean Squared Error

ROI – Return on Investment

SMOTE – Synthetic Minority Over-sampling Technique

SVM – Support Vector Machine

TN – True Negatives

TP – True positives

CHAPTER 1

Introduction

Regarding new information systems and technology, the financial sector has been particularly innovative. A great deal of research has been done especially on the topics of prediction of credit risk used to decide credit approval (Moro *et al*.,2015), being that the first research on credit risk was published in the nineties (Durango-Gutiérrez *et al*., 2021).

Artificial Intelligence (AI) solutions have been implemented in several business areas of the financial sector like marketing and risk management causing them to evolve.

This AI implementation enables cost reduction, and efficiency boost and can significantly improve a financial institution's market competitiveness (Liu & Han, 2022; Ghatasheh, *et al*., 2020). One potential area of application of AI in the marketing field can be direct marketing. Direct Marketing is gaining popularity in recent years as it is an approach that allows cutoff costs and maximizes efficiency on campaigns (Ładyżyński *et al*.,2019). It is so popular that it has become the primary strategy of several businesses in the financial sector, like banks, for customer interaction (Koumétio *et al*., 2018) Additionally, there has been an increase in purchasers thus an increase in loan demand, and that consequently increases the manpower need (Agarwal *et al*., 2021). This creates the need to identify and select the best customers to mitigate the need for manpower and make the loans easier to manage.

One common approach to this is the response modeling inserted in the target marketing strategy. This type of modeling is used to identify which customers are more likely to have a positive response to a certain campaign to target them through different communication channels like email or phone.

The banking sector is under increasing pressure to boost profits and save expenses, therefore optimizing targeting for telemarketing is a critical issue (Moro *et al*., 2014).

This optimized choice of targets is usually conducted with data mining techniques like Support Vector machines (SVM) and neural networks (NN) (Pan & Tang,2014). Because of their flexibility in computing nonlinear relations in the data., NN and SVM frequently produce reliable predictions (Moro *et al*., 2014).

The results of the optimization of the targeting are quite important as it allows for a higher percentage of clients that will most probably convert after receiving communication regarding the total of clients that will be contacted, which permits an increased Return on Investment (ROI) (Piatetsky-Shapiro & Masand, 1999).

There are previous studies about target optimization using loan predictions such as credit risk, probability of default, or propensity of making a term deposit for the outbound marketing strategy of lending institutions and banks. However, to this author's knowledge, there has not been research using the loan amount prediction for the target selection in an outbound marketing campaign. As a lending institution's profit of a loan comes from the loan interest, consequently the profit increases if the loan amount is higher, making customers with higher loans more valuable than clients who purchase loans of lower amounts. For that reason, the target selection, meaning, the selection of the best customers and maximization of the ROI can be improved if the dimension of the loan amount is added. There are several studies that show that there is a gain to companies in contacting the customers that will convert after a communication, if there are rigid budget constraints there can be an increase in the Roi if not only the company contacts people with a high probability of conversion but also the ones with highest monetary value. Since the loan amount is not a variable available prior to the direct communication, but it might be important for the selection as mentioned above, it can be of added value to the prediction of the loan amount of a customer. This compresses the main motivation of this research, which is the prediction of the Financed Amount of a person after a direct marketing communication to help spot the best customers to contact.

One of the research questions is to what extent does predicting the customer value in direct marketing within the domain of personal finance contribute to enhancing ROI and reducing overall marketing costs?

Consequently, the other that follows is what approach or model offers the most effective means to predict customer value in the context of direct marketing for personal loans?

This dissertation will address the following topics: Section 1 will undertake a literature analysis to synthesize all the work previously done; Section 2 will describe the applied methodology, including the prediction models and performance metrics used in this analysis; Section 3 will present the results obtained in the course of this study, and finally, the main conclusions, limitations and potential future work will be outlined.

CHAPTER 2

Literature Review

For this literature review, the main source of research of previous works was the Scopus website. Scopus is a database of abstracts and citations of scientific articles, conferences, and book chapters. In Scopus, we also have access to information that helps evaluate the relevance of the article like the number of citations and year of publication. This information is used to compute a metric called Field-Weighted Citation Impact (FWCI) that intends to express through a number of how frequently the article is cited when compared to similar works. Articles with a FWCI higher than 1 are more cited than the average similar studies.

For the articles search, several term combinations were used in Scopus's search engine and all terms were searched within article, abstract, and keywords.

Firstly, the scope of the literature analysis was defined. As the main objective of this literature analysis was to gather insights that concerned trends and gaps in the literature in the context of loan amount prediction with neural networks for outbound marketing, an analysis of the underlying themes that are contained in this objective was performed. The main terms were Personal Loan, Loan Amount Prediction, Outbound Marketing, and Neural Networks.

Although it was defined the main terms of the search, it was also taken into consideration that there are synonyms and related terms to the ones chosen, for example, credit and loan are intimately related so saying "personal credit" is the same as "personal loan". The more general terms are summarized in Table 1. As it is important to filter the number of articles that are returned in a search query to make the final selection of relevant articles easier, it is also important to not filter articles that have a high likelihood of being relevant to our study due to the use of very limited terms. Scopus's search engine is designed to accommodate this need as it is possible to compute complex searches by using specified parameters and Boolean operators "AND" and "OR.", for example, ("loan" OR "credit"). The search, as well as the Boolean operators, it was used quotes to make the search return results with the words together, being an example of the expression "personal loan". Moreover, to include cases like "amount of the personal loan" and to exclude cases where the word amount was not related to the loan, it was used the parameter PRE/6, which forces the two words to have a maximum of 6 words between them, preceding the word amount. Some of the search keys used are summarized in Table 2.

Table 1 Summary of the themes and sub-themes considered relevant for the literature review.

THEME TERMS	MORE GENERAL TERMS OR SIMILAR TERMS
PERSONAL LOAN	Personal Credit Microloan/credit Insurance Loan Retail bank P2P Lending Retail Finance Bank
LOAN AMOUNT PREDICTION	Loan default prediction Loan approval prediction Risk Score prediction Financed amount prediction Loan segmentation Personal Loan propensity
OUTBOUND MARKETING	Telemarketing Direct marketing Precision Marketing CRM
NEURAL NETWORKS	Machine Learning Supervised Learning Data Science Data Mining Classifier Deep Learning (DL) Artificial Neural Networks (ANN) Convolution Neural Networks (CNN) Recurrent Neural Network (RNN) Quantitative methods

Table 2 Number of results per key used in Scopu's engine (January 2023)

SEARCH KEY	Nº OF ARTICLES RETURNED
NEURAL NETWORKS FOR OUTBOUND MARKETING	2
NEURAL AND NETWORKS AND LOAN	509
NEURAL AND NETWORKS AND (AMOUNT PRE/6 LOAN)	7
"LOAN AMOUNT" AND MARKETING	7
MARKETING AND "PERSONAL LOAN"	8
PROFILE AND LOAN AND AMOUNT	39
PROFILE AND "MICRO LOAN "	3
SEGMENTATION AND "PERSONAL LOAN"	4
" LOAN AMOUNT" + AND PREDICT	12

In case the term search returns numerous results, the articles were sorted on descending publication year and on descending number of citations, on two different instances and only the articles featured on the first page results were analyzed. After that, it would add more complexity to the query to filter more and have as a result fewer articles.

To evaluate the trustworthiness of the study the year of publication was considered, being that the most recent publications were considered more relevant as they encapsulate the findings of older studies and contain new findings and developments. Moreover, publications with more citations were deemed more reliable as well. In the case that the article was very recent (2020 onwards), if there were not any citations, the number of citations metric was overlooked as the short period of time might had more impact on not having citations than the quality of the article. Nonetheless, in the cases of 0 citations, it was performed a search on Scimago to consider the h-index of the journal. on which being in the first or second quartile was desirable. There were articles included that were older (before 2010) or had 0 citations due to their relevance to the current study.

To assess the relevance of the study, the abstract, introduction, and conclusion of the collected articles were reviewed to identify their connection to the main themes.

For the Loan amount prediction, it was included studies about loan prediction in general. as it is the case of prediction of the probability of default of a loan, loan approval and risk score. These types of predictions were included, firstly because, as far as this author's knowledge, there is no prior research using the loan amount, this was reaffirmed after a search on Google Scholar to find articles with it with no success, but also because there were recognized as relevant. Even though the loan amount was not the target variable, the variables used in the mentioned predictions are related to it and sometimes include the loan amount itself as is the case in the prediction of loan default conducted by Jin, & Zhu (2015), and for that reason, the modelling techniques can be repurposed for the loan amount prediction. The same rationale was applied to research about loan segmentation, as their findings could be useful to hint at possible features to be used. There was no exclusion on the type of business, so the articles include not only retail finance institutions but also banks and peer-to-peer (P2P) lending institutions. Banks are a business as traditional as retail finance institutions but as Han *et al*.'s (2017) argue P2P marketplaces offer a substantial amount of data for microloan marketing study.

Regarding outbound marketing, studies related to direct marketing and telemarketing were included. In direct marketing, the objective is to forecast a certain client behaviour, according to Piatetsky-Shapiro & Masand (1999) so it can be used in outbound marketing strategy. Outbound Telemarketing is a successful form of direct marketing conducted by phone based on Gu *et al*.'s (2021) study, making telemarketing relevant. Other forms of specific direct marketing such as direct mailing was also included, as the modelling techniques can be reutilized for the general themes.

Referring to Neural networks, the neural network types used were artificial neural networks (ANN) and convolutional neural networks (CNN) were used. Even if the article did not use neural networks, it might not be excluded if it was relevant to loan prediction or outbound marketing, noting that research that did include comparisons of the performance of neural networks to other machine learning algorithms or suggested optimization methods to the neural networks deemed as more relevant.

After a first collection of articles through Scopus's search engine, it was conducted further investigation of articles that were referenced by the firsts. The final reference list comprises almost half of the articles identified through the latest method.

There is a considerable amount of research that mixes several sub-themes, having found articles that mix all sub-themes (Zhang, 2022), (Ładyżyński *et al*., 2019; Shih *et al*., 2014; Youqin Pan & Zaiyong Tang, 2014).

Regarding the models and techniques used in the literature, in 2014 Moro *et al*. did a follow-up research with a Telemarketing Portuguese bank data set of the University of California at Irvine (UCI), this time pretending to not use features that were not known before the direct marketing approach such as duration of the call, as these features cannot be used for success prediction, although may conceal valuable insights about the success of the campaign as shown in the previous study. In this research feature preprocessing was conducted to reach 22 features out of 150 features with a 2-step method that consists of combining the knowledge of business experts and an automated forward selection method. The models used were logistic regression (LR), decision trees (DT), neural networks (NN), and support vector machines (SVM). The authors tackled the issue of SVMS and NNs being black-box models by mentioning two methods; one focuses on uncovering the underlying rules made by the more complex models to reach their predictions by leveraging models like decision trees, and this method is rule extraction. The other method consists of response sensitivity analysis, which analyses the impact of varying each input to the model's output. The best results in Area Under the Curve (AUC) and ALIFT were achieved with NN, 0.80 and 0.60 on rolling window evaluation, respectively.

Moro *et al*. (2015), did a new follow-up research on the topic and tried to improve the NN used in the previous study by incorporating historical information about marketing campaigns' success and features that reflect the Customer Lifetime Value (CLV). Part of the features that corporate CLV are features related to Recency, Frequency, and Monetary (RFM), the recency is linked to how long has been since the customer interacted with the company (months since last purchase, e.g.), frequency to how frequently the customer interacts with the company (number of times client has purchased an item, e.g.) and monetary factor is meant to translate how valuable the customer is (mean expenditure per order, e.g.). The model enriched with the new features had an increase of 6 pp in the ROC area and of 4 pp in the cumulative lift curve on the predictions for clients with historical data when compared to the baseline model.

Elsalamony and Elsayad (2013), compared Multilayer Perceptron Neural Network (MLPNN) and DT (C5.0) performances on UCI's bank dataset (without external data) and achieved better results in sensitivity, specificity and Accuracy on the test sample with the MLPNN algorithm.

Asare-Frempong & Jayabalan (2017), also used the Portuguese bank dataset. In the research, no external data was added, and the dependent variable class imbalance was dealt with by down sampling the 'No' class to be the same size as the 'Yes' class. The models used were decision tree, random forest (RF), logistic regression and MLPNN and the model with the highest AUC and Accuracy was Random Forest,86.8% and 90.00%, respectively. They performed a cluster analysis and found that one important feature to distinguish customers was variable call duration, which is not available prior to a call, so this model could not be used for prediction purposes.

Kim *et al*. (2015) proposed the use of Deep Convolutional Neural Network (DCNN) as it can be useful to uncover the underlying relations of inputs, for example, uncover that age and income together indicate retirement, etc. The DCNN was tested on the Portuguese telemarketing dataset and presented higher accuracy when compared to six other classifiers: DT, K-Nearest Neighbours (K-NN), Naïve Bayes (NB), LR, MLP and SVM.

In another work on the same dataset, Koumétio *et al.* (2018), have proposed a new classification approach that consists of identifying the class of an instance based on the Euclidean distance to the centre of each class. The method was compared to four other algorithms DT, NB, ANN and SVM on 21 features of the 150 available in the dataset, ANN had the best accuracy among the five algorithms when applied with normalization, but the proposed approach had a higher f-measure, without normalization the ANN outperformed the other algorithms in both metrics. Interestingly, the neural network performance in terms of f-measure worsened with the data normalization, would have been interesting to compare the results with other scaling methods such as standardization.

Although the previous works mentioned did not pay special attention to cater for the problem of imbalanced classes, except for Asare-Frempong & Jayabalan (2017), Ghatasheh *et al*. (2020) tried reducing the False Negative rate caused by class imbalance with cost sensitivity analysis. Used MLP as the baseline algorithm and tried two methods of cost sensitivity, cost-sensitivity classification, and cost-sensitivity learning. The two approaches were compared to the MLP algorithm without cost sensitivity, Deep learning for MLP classifier, random forest and others including from related works, and the MLP algorithm with cost sensitivity learning had the best true positive rate (0.808), but the random forest achieved the highest accuracy (89.82), mainly due to favouring the prediction of the most frequent class as the true negative rate was 0.98. Turkmen (2021), predict the telemarketing success with the same by using the algorithms Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU) and simple Recurrent Neural Network (RNN). The Synthetic Minority Over-Sampling Technique (SMOTE) approach was used to deal with data imbalance. When comparing the predictions results in terms of accuracy and F1-Score, all the algorithms performed better when used with data that was over sampled than with imbalanced data in both performance metrics. The three algorithms with SMOTE scored an accuracy of 0.87, and the GRU method had the highest F1-Score with 0,85.

Y. Pan & Z. Tang (2014) did research on the imbalanced data issue in direct marketing. They studied the effects of using two different ensemble methods in the prediction of direct marketing response on the Portuguese bank dataset. The ensemble methods used were boosting and bagging. The models implemented were bagged neural networks, bagged logistic regression and gradient-boosting. When comparing sensitivity (TPR), specificity (TNR) and ROC curve, the bagged NN showed the best performance and the Gradient Boosting the worst. Although the boosting method did not perform the best, the authors claim boosting can be a good method to deal with the class imbalance and for this study's results the factor of the sample size of the train set, in this case, 45212, should be taken into consideration as it affects both boosting and bagging results. Boosting was initially designed to perform on large data sets, whereas bagging on smaller ones.

Gu *et.al* (2021), made use of the data set to test the advantages of Deep Learning on a high-dimensional data set. The authors of the previous study argue that feature selection comes with information loss, so they propose to leverage the capabilities of Deep Learning models to learn the complex relations of a data set with high number of features. The data set had originally 25 categorical variables that were one-hot-encoded into 82 numerical features. SMOTE was also used to turn the class of 'Loan Executed' from 19.2% of the data to 50%. In this study it was compared the performance between machine learning models such as SVM, RF, Xgboost, LightGBM and GBM, as well as Deep Learning models like a DNN, three different CNN and an author's proposed model called XmCNN.
The proposed model identifies important features with a "CancelOut" layer that assigns negative weights to non-important features, the feature extractor in the model's architecture includes

convolutional layers for feature extraction and pooling layers for sub-sampling. Average pooling is also employed after convolution to reduce the dimensionality of features and introduce invariance, aiding in classification.

LightGBM showed the best performance among ML models, with an accuracy of 0.8384 and the proposed XmCNN model outperformed both traditional DNN and convolutional neural network models CNNS. Ensemble models were created with different combinations of the previously mentioned models, one ensemble model composed by the three CNNs and the proposed XmCNN outperformed both ML a DL models. The ensemble approach demonstrated robustness against class imbalance issues in the dataset.

In most experiments with neural networks in the referenced literature, neural networks outperform other supervised learning algorithms. The outperformance and tendency of utilization of neural networks in these types of tasks does not come as a surprise as it was mentioned in the literature beforehand in articles like Shih *e*. (2014). In 2014, Shih *et al*. conducted a model comparison between logistic regression, decision trees, neural networks, and support vector machines in the task of classifying customers between two classes: applying for personal loans and not applying in another dataset. The dataset used included variables related to customer behaviour, financial indicators, transaction history, and other relevant attributes. The performance metrics used were Cumulative lift (10%), Gini coefficient and the Misclassification rate. In this article NN outperformed the other algorithms in the three metrics, the misclassification rate was approximately 0.302, Gini coefficient of 0.522 and the cumulative lift of 1.77. The LR had the second-best performance and DT was the worst algorithm in this experiment.

Upon analysis of the articles collected it is possible to outline some trends in the research field, as well as some gaps. Table 3 provides a summary of the models employed, the application of class imbalance treatment, and the number of input features. Half of the articles addressed class imbalance, and, except for one study, none utilized more than 30 features.

Table 3 Comparison of works that used UCI's dataset

Authors	Title	Models	Class unbalance treatment	Number of Input Features
Moro *et al.*, 2014	A data-driven approach to predict the success of bank telemarketing	Logistic regression Decision tree Artificial neural networks Support vector machine	No	22
Moro *et al.*, 2015	Using CLV and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns	Artificial neural networks	No	27
Elsalamony & Elsayad, 2013	Bank Direct Marketing Based on Neural Network	MLPNN Decision Tree	No	16
Y. Pan and Z. Tang, 2014	Ensemble methods in bank direct marketing.	Neural Network Logistic Regression Gradient Boosting	Yes	16
Kim *et al.*, 2015	Predicting the success of bank telemarketing using deep convolutional neural network	Deep convolutional neural networks	No	16
Asare-Frempong & Jayabalan, 2017	Predicting customer response to bank direct telemarketing campaign	Artificial neural networks Decision tree Logistic regression	Yes	16
Koumétio et al., 2018	Optimizing the prediction of telemarketing target calls by a classification technique	Nalve Bayes classifiers Decision tree Artificial neural networks Support vector machine	No	21
Ghatasheh *et al.*, 2020	Business analytics in telemarketing: cost-sensitive analysis of bank campaigns using artificial neural networks	Artificial neural networks Support vector machine Random Forest	Yes	16
Turkmen, 2021	Deep learning-based methods for processing data in telemarketing-success prediction	Long short-term memory Gated recurrent unit - Simple recurrent neural networks	Yes	20
Gu *et al., 2021*	Predicting Success of Outbound Telemarketing in Insurance Policy Loans Using an Explainable Multiple-Filter Convolutional Neural Network	Random forest Support vector machine Gradient boosting machine Extreme gradient boosting light gradient boosting machine Deep neural networks Deep convolutional neural networks	Yes	210

Neural Networks have great potential to make predictions of the quality of loan amounts and optimize a marketing campaign's ROI. This idea is prominent in the studies, as older publications would have the tendency to compare neural networks' performance to other supervised learning algorithms such as SVM's and random forest, whereas more recent publications such as Gu *et al*. (2021) and Youqin Pan & Zaiyong Tang (2014) use neural networks with proposed optimizations methods.

Although Neural networks tend to have greater performances, it is important to mention that they are to be used in a Marketing context, so it is important to keep the neural networks understandable and avoid the black-box problem to avoid local marketing teams distrusting the models. So, people must understand the decisions made by the models to some extent to avoid the latter (Zhang *et al*. ,2022). Neural Networks also require more attention as they are prone to overfitting problems and the tuning of hyperparameters is harder (Baesens *et al*., 2002) Another observed pattern was the type of data used in direct marketing articles, most used external data and marketing campaigns-related data. This observable trend had already been spotted previously in literature with studies like Bose & Chen (2009), that enforces the idea that using only behavioural data has become unpopular and the utilization of the latter is of extreme importance.

One identifiable gap was the lack of research on predictions of loan amounts. Several studies used the Portuguese bank dataset for direct marketing usage of loan prediction, The utilization of this dataset would be interesting to direct comparison of other studies results, but the loan amount variable is not included.

Other than the gap regarding the target feature prediction, no research was found that debated the importance of loan amount in the calculation of the ROI of a direct marketing campaign.

CHAPTER 3

CRISP-DM METHODOLOGY

This research follows the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, which has proven to boost the success of data mining projects (Moro *et al*., 2011). The methodology consists of 6 sequential steps: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment as seen in Figure 1. CRISP-DM is cyclical and non-rigid, as we can revisit a previous step at any point and do several rounds in each stage. In the next sections, we will go into more detail on the implementation of the methodology given the context of this research.

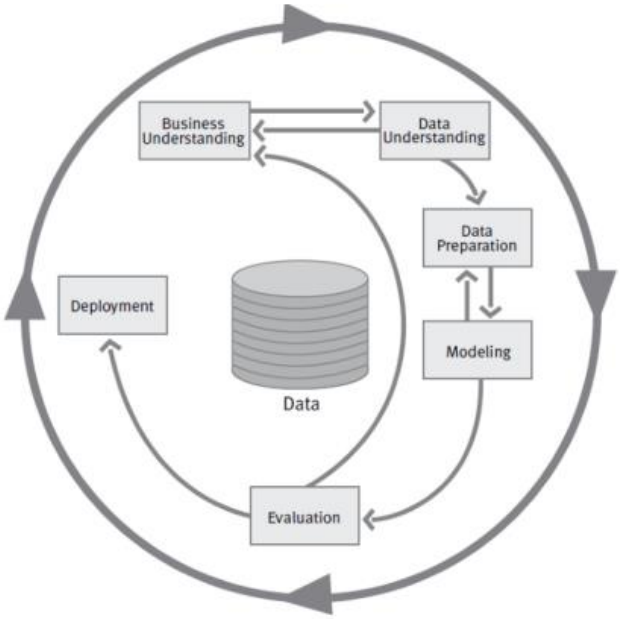

Figure 1 The CRISP-DM Process - Source: Chapman et al. (2000)

3.1. Business Understanding

Research in the field of marketing strategies has revealed the significant impact of precise targeting in marketing campaigns. Selecting customers with a high likelihood of conversion results in improved ROI by optimizing budget allocation and maintaining or increasing the number of valuable customers reached. While targeting customers based on response prediction enhances ROI, there is an additional opportunity for optimization through understanding customer value.

This study aims to predict the financed amount of personal loans following marketing campaign communications. It seeks to compare the theoretical ROI of contacting clients based on a predicted positive response rate against selecting customers based on a predicted financed amount within the given sample size. Two distinct approaches of predicting the financed amount will be examined. The first approach involves predicting the response to the marketing campaign and then forecasting the financed amount for those campaigns with a positive response prediction. In contrast, the second approach directly predicts the financed amount. The latter approach might exhibit lower performance due to the broader sample size and increased variance, including zero values.

The primary objectives of these marketing campaigns lie in cross-selling and upselling strategies. Cross-selling involves promoting new products that complement prior purchases, while upselling aims to enhance the previously acquired product. For instance, in the context of the dataset, cross-selling might involve offering a personal loan to a customer holding a credit card, while upselling could involve proposing an increased expenditure limit for that same card. Additionally, the company engages in promoting consolidation, wherein customers with multiple loans/credits from various entities can transfer their debts to the company. This consolidation simplifies the payment process for the customer and presents an opportunity for the company to earn interest from a credit previously unrelated to their clientele.

The project's success will be measured by demonstrating that selecting smaller samples of customers by ranking them according to the predicted financed amount yields a higher ROI than selecting random samples of customers predicted to have a positive response.

The primary business objective is to optimize the Return on Investment by accurately identifying valuable customers through precise targeting strategies. Understanding customer value through the predicted financed amount aligns with the company's overarching strategy for cost optimization and market competitiveness.

The prediction of the financed amount will be based on real data from a Portuguese Personal Finance branch affiliated with a French Bank. A potential risk identified is the complexity of predicting the financed amount, a numerical feature that may pose challenges in accurate estimations.

Overcoming these complexities is critical for the success of the project and the achievement of the outlined business objectives.

3.2. Data Understanding & Data Preparation

The dataset represents a compilation of observations spanning from December 2018 to July 2021, excluding May 2020. These observations encapsulate diverse customer interactions and financial

transactions relevant to the Personal Finance company's services. The dataset amalgamates information sourced from various tables, reflecting a robust repository derived from multiple sources.

As this data belongs to a company, this section will refrain from providing detailed information about feature distribution and unique values.

The company's primary products revolve around personal loans, typically capped at 75,000 euros over a 10-year period, and revolving cards. These revolving cards function as credit cards with a balance that allows for deferred payments, offering customers flexibility in managing their payments over time. Additionally, the company has established partnerships with other entities, enabling consumers to pay for products through installment plans. When these arrangements occur, the company covers the full product cost to the service provider, while the end consumer repays the installments to the company.

Frequently, these loans are short-term and maintain a 0% interest rate, serving as a means for the company to expand its potential client base for targeted marketing campaigns. Marketing campaigns by the company focus on direct customers who acquire products directly from the company and indirect customers who obtain products without direct contact, such as those making purchases in affiliated stores and choosing to pay in installments.

The data underwent an iterative, non-sequential process, incorporating feature engineering on individual tables as well as on the integrated composite dataset formed by merging multiple tables. This approach allowed for a comprehensive analysis and preparation of the dataset for subsequent data mining procedures.

3.2.1. Database Creation

To achieve the final dataset to be used during the modeling phase, there were 5 tables used. The different table's original features can be found in the Tables 4, 5, 6, 7, 8. The five tables were merged using the date and the customer's ID.

Table 4 Marketing Campaign's Features

Table	Features	Detail
Marketing Campaigns	CAMPAIGN_ID	Identification of campaign
	CAMPAIGN_TYPE	Objective of campaign (e.g., cross-sell)
	CHANNEL	Channel of communication (e.g., email)
	CUSTOMER_ID	Identification of customer
	MKC_COMMUNICATION_DATE	Data of sent communication
	MKC_CONTRACT	Identification of contract
	MKC_PRODUCT	If Personal Loan (PL) or Revolving Card (RC)
	MKC_PRODUCT_TYPE	Code of product type
	OFFER	Promotional offer
	PLANO	Type of campaign
	SUB_SEGMENT	Marketing Campaign segment

Table 5 Loan Equipment's Features

Table	Features	Detail
Loan Equipment	CONTRACT_ID	Identification of contract
	CREATION_DATE	Date of contract creation
	CUSTOMER_ID	Identification of customer
	DURATION	Duration of contract
	FINANCED_AMOUNT	PL Financed Amount
	INSTALLMENT_AMOUNT	Amount the client pays each month
	NUMFIN	Number of financings
	OBSERVATION_DATE	Year and month - YYYYMM
	OUTSTANDING	Value that the client still needs to pay
	POS	Point of situation of contract - if it is regular or not
	PRODUCT_TYPE	Code of product type
	TYPEPROD	= PL

Table 6 Card Equipment's Features

Table	Features	Detail
Card Equipment	CMA	Card Maximum Total - maximum total allowed for the customer
	CONTRACT_ID	Identification of contract
	CRU	Total Credit Used - maximum the client has chosen
	CUSTOMER_ID	Identification of Customer
	DISPO	Available amount
	DURATION	Duration of contract
	FINANCED_AMOUNT	RC Financed Amount
	INSTALLMENT_AMOUNT	Amount the client pays each month
	NUMFIN	Number of financings
	OUTSTANDING	Value that the client still needs to pay
	POS	Point of situation of contract - if it is regular or not
	PRODUCT_TYPE	Code of product type
	TOTAL_FINANCED_AMOUNT	Historical Financed Amount
	TYPEPROD	=RC

Table 7 Card Transaction's Features

Table	Features	Detail
Card Transaction	CONTRACT_ID	Identification of contract
	CUSTOMER_ID	Identification of Customer
	OBSERVATION_DATE	Year and month - YYYYMM
	TRANSACTION_AMOUNT	-
	TRANSACTION_DATE	Date of Transaction
	TRANSACTION_MODE	Transaction Mode (e.g. ATM)
	TRANSACTION_TYPE	Type of Transaction - product code

Table 8 Additional Features

Table	Features	Detail
Additional Features	AGE	-
	CUSTOMER_ID	Identification of Customer
	FAMILY_DEPENDENTS	Number of family dependents
	HAS_EMAIL	Has email – 1 if yes
	HAS_FIXEDPHONE	Has fixed phone – 1 if yes
	HAS_MOBILEPHONE	Has mobile phone – 1 if yes
	INCOME	-
	MARITAL_STATUS	-
	OBSERVATION_DATE	Year and month - YYYYM
	SEX	-

To ensure data accuracy, reliability, and consistency several data-cleaning steps were undertaken on the tables separately.

- **Granularity:** The tables "Loan Equipment" and "Card Equipment" initially operated at the Customer and Contract level of granularity. However, for the purpose of this study, which aims to predict the comprehensive impact of communication on customer behaviour rather than on specific contracts, the data was aggregated to the customer level. Similarly, the "Transaction Table," which initially featured data at the customer, contract, and transaction levels, underwent the same granularity adjustment.
- **Data Filtering:** Data filtering was performed to create a month-level view, ensuring that only information available before the communication date was retained.
- **Outlier Treatment in Loan Equipment Data:**
  - Financed Amount Investigation: Observations with loan financed amounts surpassing the 75,000 € limit were reviewed to ensure data accuracy. Less than 0.02% of rows were affected, with no currency-related issues found, leading to the removal of these observations for data integrity maintenance.
  - Duration Exclusion: Observations displaying durations exceeding 120 months, constituting less than 0.001% of the dataset, were considered outliers and subsequently removed to improve dataset quality. Additionally, observations with a duration of zero months, accounting for less than 0.0004% of the data, were eliminated.
  - Shared Contracts Elimination: Contracts shared among multiple individuals, which is atypical for individual-centric marketing campaigns, were removed. This action impacted fewer than 274 rows.
  - Excessive Contracts Removal: Observations with an excessive number of loan contracts, exceeding 3 standard deviations from the mean and constituting 0.12% of observations, were removed. This process reduced the maximum number of active loan contracts per customer to 16 in the dataset.
- **Categorical feature cleaning:**
  - Addressing data uniformity, categorical feature values were standardized. Some values had letter prefixes followed by explanatory sentences, while others contained only letters due to data collection errors. These values were adjusted to retain only the initial letter to ensure consistency.
  - One-hot encoding was utilized to manage high categorical cardinality, applied to the most prevalent values, while other types were consolidated into a single binary feature.

- **New Features Creation:** In line with Moro *et al*.'s (2014) suggestions, the dataset underwent feature engineering, integrating RFM attributes.
- **Integrity checks:** The consistency between contract creation dates and observation dates was verified to ensure data accuracy, and any inconsistent rows were removed.
- **Data Types update:** All features, initially labelled as object data types, underwent conversion to enhance data readability. Specifically, the creation date features were updated to a datetime format, while numerical attributes were updated to float data types.

3.2.2. Final Data Preparation & Feature Engineering

The dataset with all the information merged counted more than 4 million observations and the following was made:

- **New features were introduced**. One of which relating the income, family dependents, and an external feature called the Harmonised Index of Consumer Prices (HICP). HICP measures the price of goods and services that is representative of the usual price that a family pays in a month. So a new feature was introduced by subtracting the sums of all of the instalment amounts, and the product of the HICP with family dependents to express the theoretical disposable income, with the preposition that for very low disposable incomes a new loan contract would not be feasible. As income had outliers, was performed a winsorization where the values higher than the 95th percentile or lower than the 5th percentile were replaced with values of the 95th percentile and 5th percentile, respectively. The same was done to the theoretical disposable income feature.
- **Missing data cleaned**:
  - There were several missing values due to the merging of the different information sources. On features on which 0 has a specific meaning such as in the case of the feature MIN_NB_MONTH_CREATION, where a 0 means a loan contract has been created in that month, the missing values were replaced with -1 to be account for a different situation and not have a high Euclidean distance to 0 and 1 to not create problems after scaling. In the cases, where 0 had the same meaning for customers that had null values due to different data sources such as the number of loan contracts, the missing values were replaced with 0.
  - The instances in which observations had missing values simultaneously in all socio-demographic features, including income, age, marital status, and sex, accounted for 2.9% of the dataset and were consequently removed from the analysis.

- In the case of the "family dependents" feature, observations with missing values were addressed by categorizing age into four groups and calculating the mode of family dependents for each combination of binned age and marital status, used to fill the missing values. Due to rare instances in one age bin (less than 0.005% of the data), this class was removed. Outliers in the family dependents feature were initially considered for winsorization; however, as this would exclude rare but possible cases like having five children, a more conservative approach replaced higher values with 5 rather than 3. Additionally, missing values in the marketing campaign product type were substituted with the mode of each sub-segment, given its high correlation with the feature.

- **Categorical features cleaning**:
  - The "offer" feature, designed to depict the promotional offer in marketing campaigns, was processed using TfidVectorizer, from the scikit-learn library (Pedregosa *et al*., 2011), and clustered into six groups based on the algorithm's matrix output. It was then replaced by the cluster code assigned to each sentence. The same approach was tried on the feature Sub_segment, which encapsulates the information of the segment of customers to be contacted (if the customer belongs to the company, for example), but most clusters would be comprised by just one of the unique values in the feature.
  - The last feature was one-hot encoded, and the top values were kept as features, whereas the other were grouped into a single feature, as it had more than 30 unique values.
- **Observations Removal:**
  - The observations of July were removed, as it would not be possible for those to accurately compute the response of the campaigns.
  - There were observations where a customer received more than one communication on the same day, these rows were dropped to avoid having one of the observations of a communication in the train and another communication on test, leading to data leakage regarding the response, as the customer’s information would be the same.

- Upon further analysis, it was discovered that 56% of the communications on which there was a conversion were for PL products, which enforces the decision to keep both types of communications (RC and PL) in the scope of predicting the conversion into a loan contract upon communication. Moreover, 23% of the communications with positive responses, had at least one contact in the last 30 days, in these cases both communications may have a positive response which may give a bias in the models by including the feature previous_pl_response. There were 537 observations where the total instalment amount was higher than the income, these observations were not removed as they might not be an error, but the observations where the total installment amount exceeded the 4000 were removed, as they were considered outliers.

The resulting data frame had a total of 4482634 observations and 142 features. The observations with positive responses are an extremely rare case, with a response rate of 0.106%.

3.2.3. Covid Time Period – Data Drift Analysis

Due to a data extraction error, observations from May 2020 were not included. May 2021 coincided with the peak of the COVID-19 Pandemic. Consequently, an analysis was conducted to assess potential data drift during this period to determine its suitability for inclusion in the training dataset.

Three distinct datasets were created, each containing observations from February, March, and April of different years (2019, 2020, and 2021) to represent the pre-COVID, during-COVID, and post-COVID periods. This allowed for the consideration of seasonal effects in the analysis.

The volume of marketing campaigns experienced a slight 5% decrease during the pandemic in comparison to the equivalent pre-pandemic period there was a significant decrease of 37% from the Covid period to after Covid period. The response rate was 0.002% before COVID-19, and it notably increased during (0.0142%) and after (0.0135%) of the COVID-19 period.

The comparison of means for the 'TARGET_FINANCED_AMOUNT' feature among the different period pair combinations using a T-test (3.1) revealed a p-value lower than 0.05 in all cases. This suggests a statistically significant difference in mean values between the various periods at a 95% confidence level.

$$t = (\bar{x}1 - \bar{x}2)/(\sqrt{((s\_1\hat{}2/n\_1) + (s\_2\hat{}2/n\_2))}) \quad (3.1)$$

All combinations had a p-value lower than 0.05, indicating that for a 95% level of confidence, there is a statistically significant difference in mean between the different periods.

It was also used Alibi-detect. Alibi-detect is an open-source Python library utilized for detecting model drift and anomalies. Throughout this analysis, specific columns—CUSTOMER_ID, OBSERVATION_DATE, and MARKETING_COMMUNICATION_DATE—were excluded from the investigation. Upon comparing the periods before, during, and after COVID-19, both indicated instances of data drift in each case. During the transition from the period before COVID-19 to the during-COVID period, and from the during-COVID period to the after-COVID period, observable data drift was identified.

Drift during the transition from before COVID to during COVID was attributed to the number of communications, the campaign's sub-segment, and the relationship of the last contact to a loan. These features inherently possess historical and incremental elements, making the manifestation of drift over time expected.

Conversely, the shift from the during-COVID period to the after-COVID period witnessed significant drift primarily associated with marketing strategy-related features. Factors such as the offer, category, sub-segment, and days since the last contact exhibited noticeable influence. This aligns with the recognition of the COVID-19 period as an anomaly with substantial economic implications, necessitating potential adjustments in the marketing strategy to adapt to evolving circumstances.

Upon thorough analysis, the decision was made to retain the COVID period in this study, recognizing the significance of considering this exceptional period within the dataset.

3.3. Modeling and Evaluation

3.3.1. Target Values

To address one of the primary research questions, which centers around determining the added value of not only considering the probability of a customer's response but also the potential value the customer brings to maximize the return on investment ROI of a marketing campaign, three distinct targets were employed in the modeling process.

The targets were the following:

- **Response** – Binary feature that assumes the value of 1 when a customer contracted a loan in the space of 60 days after receiving a marketing communication / 0 otherwise.
- **Financed Amount** – Numerical feature that expresses the value of the loan contracted in a 60-day period after receiving a communication.
- **Financed Amount Categorical** – Financed Amount feature binned. One of the advantages of transforming the regression problem into a classification is the fact that the results of a classification problem are easier to interpret and compare than a regression problem. In the context of direct marketing can be particularly useful as there will be probably stakeholders

with non-technical profiles, and it is easier for them to understand classification metrics, for example, how many times the model successfully predicted a top client.

The Financed Amount target was approached through two distinct modeling strategies. The first strategy involved modeling using the resulting dataset from data processing. The second strategy focused on modeling using the observations that had predicted a Response equal to 1 from the best-performing Response model. The second approach bears potential advantages, particularly when the Response model performs well. By utilizing this approach, the model benefits from the Response model's accuracy in identifying positive responses, thereby excluding negative observations. This not only improves runtime efficiency but can also enhance the overall predictive accuracy of the Financed Amount target.

3.3.2. Train/Test Split

In this study, two distinct train/test split methods were employed to assess the model's performance effectively:

- **Random Split:** This approach was adopted with consideration of the potential impact of seasonality on loan conversion rates. It acknowledges that conversion rates can vary from month to month (e.g., conversions in June might differ from those in February). Given that positive responses are relatively rare, including a broad representation of positive cases across different months and years can be advantageous. This allows the model to capture the general trends and seasonality patterns in the data. The train and test datasets were partitioned with a ratio of 60/40%.
- **Time-Based Split** (Testing on the Most Recent Data): The time-based split method was chosen for its distinct advantages, especially in instilling confidence among non-technical teams regarding the model's predictions. Testing the model on the most recent data guarantees that the model has not been exposed to this test data previously. This real-world scenario approach is particularly valuable, as it accounts for potential changes in relationships between variables over time. For instance, the number of previous communications may vary over time, making this method useful in tracking evolving trends and patterns. The last four months were reserved for testing in the temporal train/test split.

3.3.3. Algorithms

The models were trained using three different algorithms: NN, RF and Xgboost. These specific algorithms were chosen due to their promising performance, as revealed in the literature review. The RF offers several benefits, making it appealing and a good fit for this context data; It reaches usually good performance, is sturdy while handling a high number of features, and is resistant to over-fitting

(Koutsoukas *et al*., 2017). NN are found to be good in finding intricate connections between the data, especially when it comes to defining the characteristics of clients and potential clients and predicting their actions (*Guido et al.,* 2011). Xgboost has been found promising in direct marketing applications, although has been proved to overfit (Chlebus & Osika, 2020).

3.3.4. Evaluation Metrics

The main measures to evaluate the performance of the categorical models were based on the confusion matrix which is a visual representation of the following metrics:

- **True Positives (TP)** – The number of positive cases correctly identified.
- **False Positives (FP)** – The number of negative cases incorrectly identified as positive.
- **True Negatives (TN)** – The number of negative cases correctly identified.
- **False Negatives (FN)** – The number of positive cases incorrectly identified as negative.

The performance of each categorical target model was compared using the following metrics:

- **Area Under the Receiver Operating Characteristic Curve (AUC)** – A perfect model has an AUC of 1, indicating it can effectively discriminate between positive and negative classes, while a value of 0.5 suggests random discrimination.
- **Precision (TP / (TP + FP))** – This metric expresses the percentage of correctly predicted communications that would lead to a conversion out of the total number of communications that the model predicted as a positive response. A value of 1 means that all instances predicted as having a positive response were correct.
- **Recall (TP / (TP + FN))** – This metric represents the percentage of communications that would result in a conversion and were correctly predicted as the positive class. A value of 1 indicates that all communications with a positive response were predicted as having a positive response.

The metrics used to evaluate the Regression models were the following:

- **Mean Absolute Error (MAE)** – Average absolute difference between predicted and observed values, in this case, in average, how many euros of difference between the predicted value of the loan and what the customer asked for.
- **Mean Squared Error (MSE)** - Average of the squared differences between the predicted and observed values. Bigger errors have more impact to this measure.
- **Root Mean Squared Error (RMSE)** - RMSE is the square root of MSE, which translates the standard deviation of prediction errors.
- **R-squared (R2)** - calculates the percentage of the dependent variable's (target's) variation that can be predicted from the model's independent variables (features). It has a range of 0 to 1, being the 1 the best value.

In addition to these metrics, it was also plotted graphics where the x-axis was the predicted values, and the y-axis was the observed values following the recommendations of Piñeiro *et al.* (2008).

3.3.5. Feature Selection

Several feature selection techniques were explored in pursuit of enhanced prediction performance:

- **Sequential Forward Selection (SFS)** of the mlxtend library (Raschka, 2018): This method starts with an empty set of features and iteratively assesses the impact of adding each feature. It selects the feature that yields the most significant improvement in model performance, considering the impact on predictions. Subsequently, the process continues, incorporating additional features in order of their contribution. SFS continues until a predefined stop criterion is met, such as a specific number of features or a performance threshold.
- **Recursive Feature Elimination (RFE)** from the scikit-learn library (Pedregosa *et al*., 2011): In contrast to SFS, RFE initiates by training the model using the entire feature set. It then ranks the features according to their importance and systematically removes the least important feature. This procedure iterates until the stop criterion is satisfied, optimizing the feature selection process.
- **Random Forest Features Importance:** The random forest model provides a valuable tool for assessing feature importance, typically quantified using Gini importance. Features that contribute to the reduction of impurity, leading to improved splits within the decision trees, are assigned higher importance values. While this approach is effective, it can exhibit a bias towards high cardinality features. Therefore, in this context, numerical features tend to be prioritized.

Additionally, since highly correlated features within the dataset introduce redundancy, there was a focus on reducing the number of features before subjecting them to feature selection algorithms. By implementing these strategies, the study sought to strike a balance between feature richness and prediction performance, ultimately leading to more effective models while preserving computational efficiency.

3.3.6. Data Imbalance Treatment

SMOTE & Bagging

The target feature class presented an extreme unbalance, as the majority class accounted for more than 99% of the data set, which can affect the performance of machine learning models.

To tackle this issue, SMOTE is often employed, which generates synthetic examples for the minority class to rectify class imbalances. However, in scenarios where the minority class is exceptionally rare, the application of SMOTE requires careful consideration. Oversampling too aggressively can lead to the overrepresentation of synthetic data, potentially causing overfitting and skewing the model's behaviour. To strike a balance between addressing the class imbalance and preventing overfitting, a combination of SMOTE and under-sampling techniques was implemented, following a systematic process:

1. The dataset was initially divided into two categories: observations with positive responses and those with negative responses.
2. A specific sample size, representing approximately 0.002% of the negative observations, was extracted from the majority class.
3. The selected negative sample was combined with the positive observations.
4. The combined dataset was randomly partitioned into training and testing sets, adhering to a 60/40 split ratio.
5. Standardization of independent features was performed to ensure uniformity. This process entailed adjusting all feature means to zero and standard deviations to one. Standardization is critical to mitigate the impact of variables with varying units or variances, thereby preventing model bias.
6. SMOTE was then employed, specifically focused on oversampling the minority class only in the train set. This step aimed to equalize the minority class representation relative to the majority class.

Subsequently, models were trained on this balanced dataset with to address the imbalanced nature of the target feature class.

Recognizing that training models solely on a single sample might introduce sample bias due to its potential lack of representativeness, an experiment was conducted. The models were trained using a single sample from the negative class population, and their performance was compared against an ensemble approach. This ensemble approach can be categorized as "bagging," an ensemble learning technique that leverages multiple models to enhance accuracy by combining their predictions. The final predictions of the bagging models were calculated based on the most frequent scores between each model. By adopting this balanced and systematic approach, the study aimed to improve the model's ability to distinguish between positive and negative responses, ultimately leading to more robust and accurate predictions in the context of imbalanced data.

Cost-sensitive learning

There were other methods tested to improve predictions due to data imbalance such as Cost-sensitive learning techniques. These techniques were briefly explored with NN and Xgboost algorithms to predict Response. Cost-sensitive learning, assigns different costs or weights to various types of errors, thereby intensifying the impact of certain errors during classifier training. Depending on the specific context, it might be more detrimental or costly to misclassify positive cases as negative or vice versa.

In this context, the sole financial cost pertained to communication expenses. Thus, there was a clear advantage in minimizing false positives since these corresponded to costly communications with no return on investment. However, it was also crucial to factor in the cost of missed opportunities – the potential profit that the company foregoes by not sending a communication. In this case, the unit cost of communication (varied by channel, country, and service provider) was significantly lower than the profit generated when a recipient positively responded to the campaign. Consequently, a Neural Network was trained using a custom loss function. This custom loss function assigned a weight equivalent to the mean TARGET_FINANCED_AMOUNT when individuals responded positively to the campaign for the misclassification of a positive observation while attributing a weight of 1 to negative misclassifications. This approach balanced the importance of reducing false positives while considering the potential revenue of successful communications.

The hyperparameter "scale_pos_weight" in the Xgboost algorithm was also tested, employing it as a technique to assign different weights and balance the representation of each class during the training phase.

3.3.7. Data Skewness Treatment

Not only was data imbalance considered, but data skewness was also a subject of study. To address the skewness that was adversely affecting the results of the Financed Amount, an approach involving the Box-Cox transformation was tested. When there is a significant violation of a regression model's assumptions, the Box-Cox transformation is a useful method in linear regression (Zhang and Yang, 2017).

This transformation is applied to the dataset with the aim of bringing its distribution closer to normality, potentially improving the predictive performance for the target feature as it can be seen in the Figure 2.

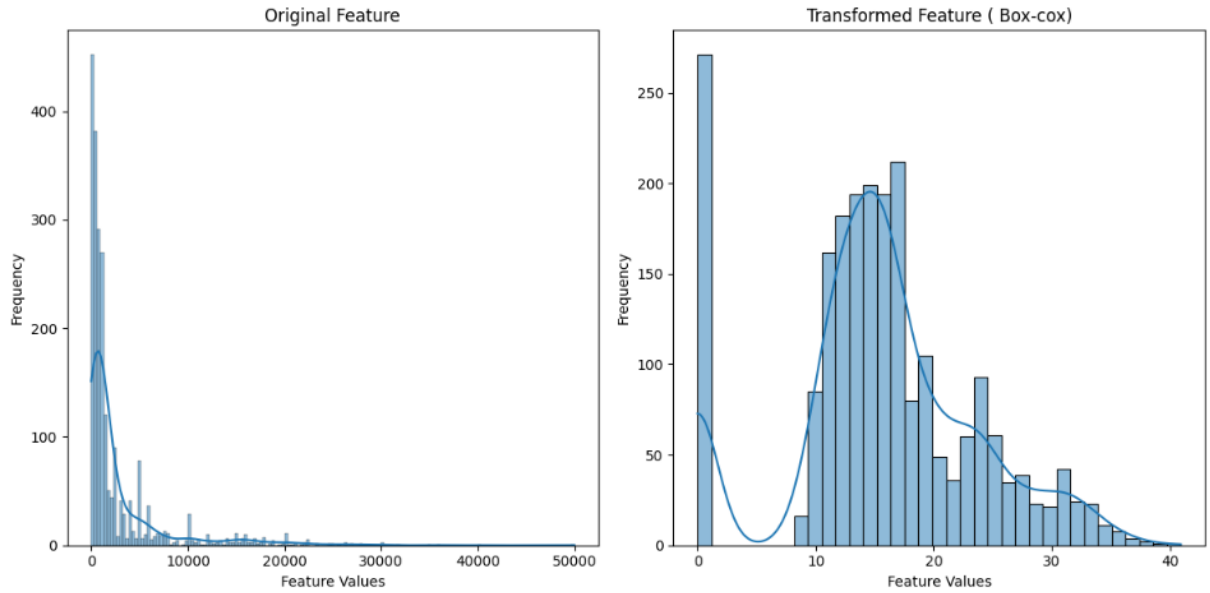

Figure 2 Distribution of the feature Financed Amount before and after box-cox transformation.

3.3.8. Hyperparameter Tuning

To improve the model's performance, the following Random Forest hyperparameters were tuned:

- n_estimators: number of decision trees used for the ensemble model. Increasing the number of trees can improve model's predictions and avoid overfitting.
- max_depth: number of maximum number of nodes or level in each tree. Deeper trees can learn more complex trends in the data but can also lead to overfitting.
- min_samples_split: minimum number of observations it takes to split a node. Helps preventing a too-complex tree with nodes with very few observations.
- min_samples_leaf: minimum number of observations in the final node of a decision tree. It can prevent overfitting by ensuring that leaf (final) nodes have a minimum number of observations.

Neural Networks hyperparameters tuned:

- Hidden layer: number of layers between the input and output layer. More hidden layers may help the model understand more complex trends in the data.
- Neurons: number of neurons in each layer. In this case, the search was just for the neurons on the hidden layers. More neurons help to model non-linear relationships in the data.
- Activation: Mathematical functions applied to the output of each neuron in a layer to introduce non-linearity. In this case, the search was just for the hidden layers, as in this problem the last layer should be sigmoid as it forces the values to be between 0 and 1. The activation also plays a part in modelling underfitting and overfitting.

- Dropout rate: a fraction of how many neurons are randomly set to 0 during training iterations. Can help improve overfitting.
- Learning rate: step size used in updating the model's weights. While a higher learning rate can speed up convergence, it also runs the risk of overshooting the ideal weights and creating instability.
- Batch size: number of observations used in each pass of the training. Although they could cause noisier weight updates, smaller batch sizes can aid in the model's escape from local minima.
- Epochs: number of times the model will learn from the entire training dataset. More epochs give the model more time to optimize its weights and boost performance, but if improperly managed, they also raise the possibility of overfitting.

For the Xgboost model the hyperparameters tuned were n_estimators, max_depth, learning_rate.

The search for optimal hyperparameters was conducted using Bayesian Optimization, involving 50 iterations and a 10-fold cross-validation strategy. Unlike GridSearchCV, Bayesian Optimization doesn't exhaustively test all possible combinations of hyperparameters. Instead, it utilizes probabilistic modeling to make informed choices based on past results. This method provides a less computationally demanding and more effective technique to fine-tune the model. Based on the data obtained from earlier iterations, Bayesian optimization constantly adjusts its search space. Its flexibility enables it to concentrate on more desirable combinations of hyperparameters and prevents the wastage of computational resources on less favourable ones. Iterations save time and computing effort by convergently finding ideal hyperparameters, thanks to the insights and guidance provided by each iteration's results.

3.3.9. Cumulative Gain Comparison

To compare the results of the different models, performance metrics were not used, as the targets varied between numerical and categorical. Instead, a cumulative gain analysis of revenue was conducted using each model for both train/test split methods. The approach was as follows:

- Save the best model for each target result.
- Sort the customers in descending order of their predicted values for each target.
- Select the first N customers from the financed amount results, with N being the number of customers that the best response model predicted would convert into a loan.
- Plot the cumulative gain, considering the value of financed amount according to the sorted order.

CHAPTER 4

Results and Discussion

In this chapter, the outcomes of the methodology outlined in Section 3 will be thoroughly examined. Section 4.1 will scrutinize the results of models developed for the four distinct targets specified earlier. These models were trained and tested using both random split and temporal approaches. Subsequently, Section 4.2 will provide a comparative analysis of the various approaches, employing cumulative gain charts as detailed in the preceding section. This comparative evaluation aims to elucidate the efficacy of each approach and identify patterns that can inform decision-making in marketing strategy optimization.

4.1. Individual Model Approaches Performance

4.1.1. Response Target

The models targeting the response variable, employing the random/test split, underwent training with 8,280 observations and testing using a sample composed of 1,962 positive cases and 3,550 negative cases.

For the temporal train/test split, the most recent 4 months were held out for testing, resulting in a test set with 423,783 observations, comprising 442 positive responses and 423,783 negative responses.

To compare each approach of model improvement through feature selection, bagging, and hyperparameter tuning, it was modelled a baseline model.

Random Forest Baseline Model:

- **Library**: scikit-learn
- **Hyperparameters**: n_estimators=100 (default values for other hyperparameters)

XGBoost Baseline Model:

- **Library:** Xgboost
- **Hyperparameters:** n_estimators=100 (default values for other hyperparameters)

Neural Network Baseline Model:

- **Library**: tensorflow, scikeras
- **Architecture**: Input layer with 20 neurons (relu activation), one hidden layer with 7 neurons (relu activation), and output layer with one neuron (sigmoid activation).
- **Compile Configuration**: Binary cross-entropy loss function, Adam optimizer, and accuracy metric.

Table 9 Summary of results of Response models

Train/Test Split	Algorithm	Model	Precision	AUC	Recall	FN	FP	Bagging - Number of models	Feature Selection	Details
Random	NN	NN1	0.8786	0.9619	1	0	271	10	All features	Baseline
		NN2	0.8685	0.9580	0.9995	1	297	1	All features	
		NN3	0.8786	0.9617	0.9995	1	271	10	RFE without correlated features - 30 features	
		NN4	0.8777	0.9612	1	0	276	10	SFS without correlated features - 30 features	
		NN5	0.8786	0.9619	1	0	271	10	Random forest features	
		NN6	0.8786	0.9619	1	0	271	10	Random forest features	Tuned with parameter search
		NN7	0.8786	0.9619	1	0	271	10	All features	Custom loss function
	RF	RF1	0.8786	0.9619	1	0	271	10	All features	Baseline
		RF2	0.8685	0.9581	0.9995	1	296	1	All features	
		RF3	0.8786	0.9617	0.995	1	271	10	SFS without correlated features - 30 features	
		RF4	0.8788	0.9608	0.9975	5	270	10	RFE without correlated features - 30 features	
		RF5	0.8786	0.9617	0.9995	1	271	10	Random forest features	
		RF6	0.8786	0.9619	1	0	271	10	Random forest features	Tuned with parameter search
	Xgboost	X1	0.6923	0.5017	0.0046	1953	4	10	All features	Baseline
		X2	0.8786	0.9617	0.9995	1	271	10	All features	Tuned
		X3	0.8789	0.9615	0.9990	2	270	10	Random forest features	Tuned
		X4	0.9913	0.9466	0.9648	69	255	10	All features	Undersampling without SMOTE
		X5	0.8804	0.9597	0.9939	12	265	10	All features	With pos_scale_weight= ratio of negative cases/positive cases No SMOTE
Temporal	RF	RF7	0.0130	0.9605	1	0	33520	10	Random forest features	Best random model

In Table 9 the results are summarized, for both train/test split methods, for each algorithm NN, RF and Xgboost. For ease of reference, the column "model" comprises of a code, combining the algorithm abbreviation (e.g., NN) and the model number (also used in Tables 10, 11, and 12). The variations between models are attributed to differences in feature selection and hyperparameters as described in the columns “Feature Selection” and “Details”.

Observing the table, it is possible to observe that the bagging method improved the precision, AUC and recall of the model by comparing models NN1 and NN2. Although in terms of percentage points the difference is not accentuated, when looking at the number of false negatives, even 1 customer that was misclassified as a negative respondent has a big impact on the loss of profit, moreover when positive respondents in direct marketing can be a rare case.

In terms of feature selection, both RFE and random forests had similar runtimes, having the results ready in a matter of few minutes, whereas the SFS could take more than one hour, depending on the number of features selected (would take on average 5 minutes per feature). In terms of performance, RFE underperformed in every model when compared to not doing feature selection, using SFS result or RF’s most important features. The SFS algorithm had a close performance to the RF’s most important features but would tend to increase the number of FP (NN4) or increase the number of false negatives (RF3). The feature selection methodologies applied in these models did not explicitly bolster predictive performance. However, the noteworthy reduction in the number of features from an initial count of 142 down to 30 while preserving the same performance is undeniably valuable. This reduction, though not directly contributing to performance enhancement, adds significant value by virtue of two vital aspects: it substantially curtails runtime and substantially contributes to enhancing the interpretability and explainability of the model. This streamlined feature set not only improves computational efficiency but also empowers a clearer understanding of the model's decision-making process, thus contributing to better insights and trust in its predictions.

The model using all features and cost sensitive learning (NN7) had the same results as the NN baseline model (NN1).

Regarding the algorithm's performance, both NN and RF achieved the highest performance with the baseline models, and model NN6 outperformed the RF5 when using the same feature set without hyperparameter tuning. The baseline Xgboost (X1) underperformed all other models and even with hyperparameter tuning (X2) was still underperforming when compared to the algorithms of NN and RF.

On the subject of data imbalance, when comparing the models X3, X4 and X5, we can observe that the method of both under sampling the majority class and SMOTE the minority class outperforms not using SMOTE and the Xgboost’s hyperparameter pos_scale_weight.

In relation to the temporal train/test split results (RF7), the recall was 1, which means, that if applying this model to real-life scenario data, the model could successfully predict all cases where the response would be positive. Although the precision is significantly small, when considering the low costs of communication and the fact that exists a trade-off between precision and recall, where to improve one the other may decrease, a ratio of 1 conversion out of 100 communications sent, the results are quite good.

Moreover, these results indicate that with a reduction of more than 90% of communications sent, it would be able to communicate with all customers that would target, which would result in an increase of ROI.

4.1.2. Financed Amount after Response

Using the best model's results of the Response approach, in this case, both NN models and RF had the same result, it was predicted the Financed Amount of customers that the response models had predicted would respond. The main metrics are summarized in the Table 10.

Table 10 Results of models of Financed Amount after Response

Train/Test Split	Model	MAE	RMSE	R2	Feature Selection	Details
Random	RF8	3127	4810	-0.0508	All features	Baseline
	RF9	3164	4868	-0.0763	Random forest features	
	RF10	3292	4908	-0.0942	RFE	
	RF11	3145	4862	-0.0735	RFE without correlated features -30 features	
	RF12	3040	4964	-0.1197	SFS -30 features	
	RF13	3066	4850	-0.0684	SFS without correlated features - 30 features	
	RF14	2959	4808	-0.050	SFS -30 features	Tuned with parameter search
	RF15	2932	4680	0.0050	All features	
	RF16	2917	4667	0.0108	Random forest features (60)	
	RF17	2925	4703	-0.0046	Random forest features (20)	
	RF18	2155	4861	-0.0732	All features	Baseline model with box-cox transformation
	RF19	2144	4862	-0.0737	All features	Tuned model with box-cox transformation
	RF20	2143	4873	-0.0787	All features	Tuned model with box-cox transformation and undersampling
	RF21	2312	4992	-0.1317	SFS - 30 features	Baseline model with box-cox transformation
	RF22	2187	4923	-0.1007		Tuned with parameter search
	RF23	2186	4920	-0.0995		
Temporal	RF24	61	944	-0.0039	All features	Best model
	RF25	92	906	0.0730	All features	Undersample

For the 'Financed Amount' target variable, an evaluation of feature selection methods applied to the response model was conducted. The SFS algorithm exhibited superior robustness when compared to RFE, especially in scenarios with highly correlated features. When highly correlated features were removed using the RFE algorithm, we observed a reduction in the average error by €147, along with a decrease in RMSE, signifying fewer significant errors in predictions. Conversely, employing the SFS algorithm and removing highly correlated features led to a slightly higher MAE. Nevertheless, it resulted in a reduced RMSE, indicating that, on average, the model made more errors but with smaller magnitudes, suggesting greater precision in its predictions. The model with the lowest MAE before tuning was the model with SFS feature selection (RF13).

However, after the hyperparameter tuning the lowest MAE and highest R2 were registered by recurring to the top 60 most important RF features (RF16).

The application of the Box-Cox transformation yielded significant benefits in improving our models' predictive capabilities. A notable drop in the MAE of up to 773€ was noted when comparing the model with the Box-Cox transformation (RF19) to the model without the transformation (RF16). By reducing MAE by such a substantial margin, the Box-Cox transformation demonstrates its capacity of mitigation of errors, highlighting the importance of feature engineering.

With a few exceptions (RF15, RF16, and RF24), the model's R-squared values were primarily negative, indicating that they did not demonstrate a good fit to the data.

This downward trend in R-squared values highlights the difficulties in encapsulating the intricate relationships in the dataset and calls for a rigorous analysis of the modelling strategies and features of the data. It was used data visualization approaches to provide deeper insights into the model's performance, particularly in predicting the financed amount. Visually comparing expected and observed values allows to identify areas where the models perform well and those that require work. Using a visual approach helps with decision-making related to feature engineering and model refining by providing a more intuitive view of the model's performance, although it increases complexity.

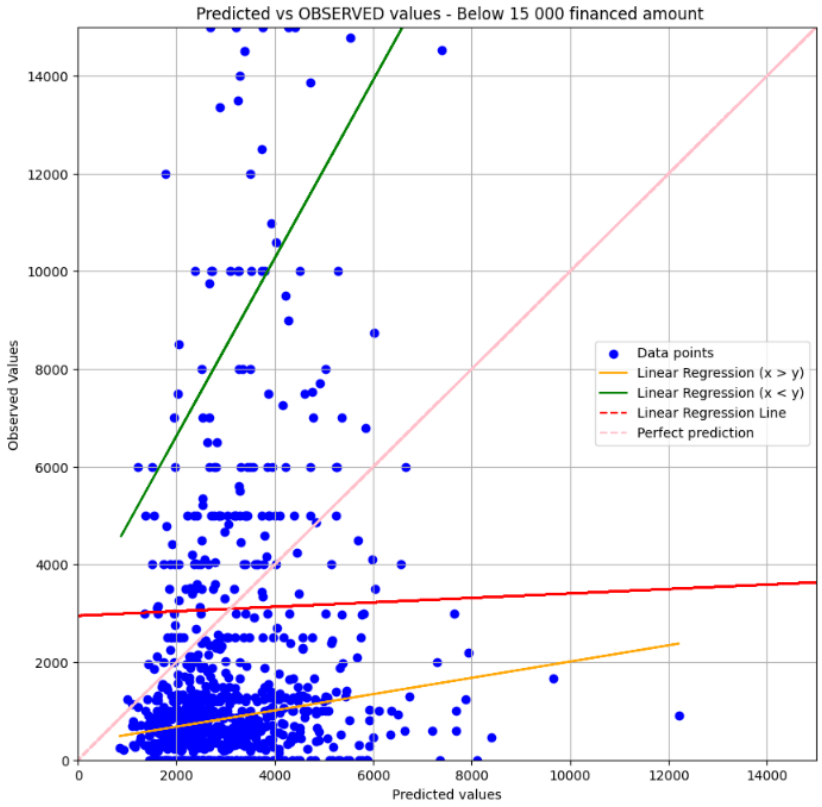

Figure 3 Results of Financed Amount model with the highest R-Squared (RF16) for the observations below 15,000€

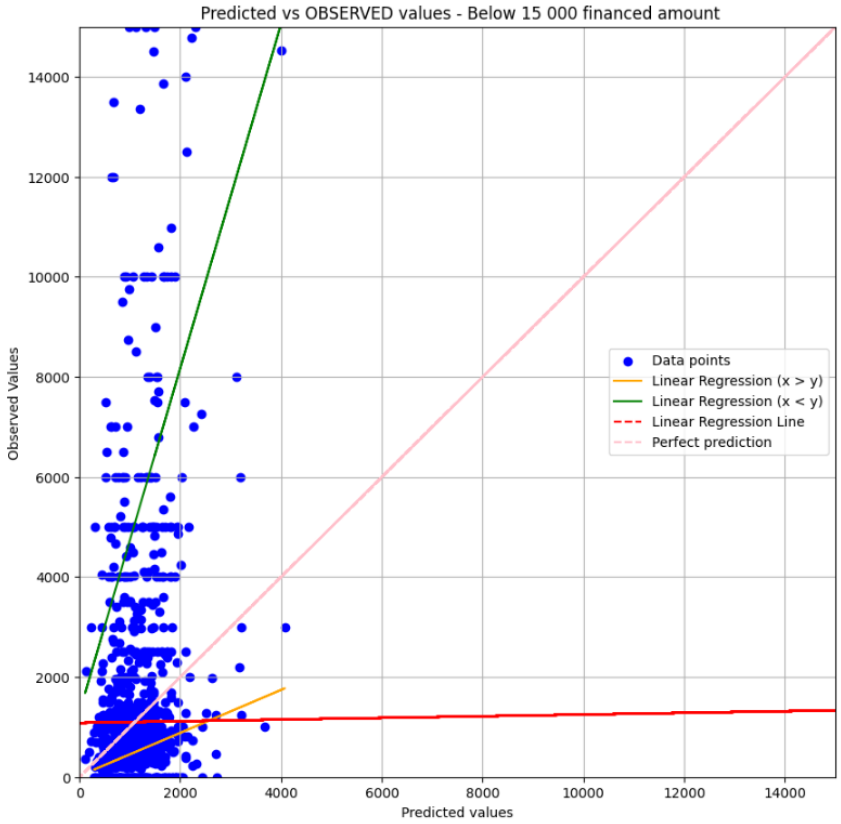

Figure 4 Results of Financed Amount model with the lowest MAE (RF19) for the observations below 15,000€

When comparing the two models, RF16 and RF19, the analysis goes beyond the conventional linear regression approach, as depicted in Figures 3 and 4. Instead, the examination extends to include regressions for overestimated (predicted value higher than the observed value) and underestimated (predicted value lower than observed value) predictions, denoted by the yellow and green lines, respectively. The linear regression (x>y) represents the linear regression calculated for the data points where the predicted values were higher than the observed values. The same logic was applied to the linear regression of data points where the predicted values were lower than the observed values.

In the context of the objective to order customers in terms of value, the regressions of underestimated and overestimated predictions offer additional insights into the ordering dynamics. For instance, if the underestimated line exhibits a slope similar to the perfect prediction (where observed equals the predicted value), it implies that despite errors, there exists a discernible ordering of customers. This nuanced exploration allows for a more refined understanding of how the models capture and prioritize the underlying value distribution among customers, going beyond the conventional evaluation metrics and shedding light on the practical implications of the models in a customer-centric context.

This method allows for uncovering nuances that go beyond the initial observation that RF16 fits the data better than RF19, despite the former having a higher R-squared.

The RF19 model's primary linear regression appears almost horizontal, indicating that, theoretically, it fits the data less effectively than the simple mean of observations. However, a closer look reveals that when observed values fall below the expected predictions, the linear regression of overestimations exhibits an almost identical intercept, along with a slope more similar to the ideal linear regression. In contrast, RF16's linear regression features a slope that more closely resembles the perfect regression line in areas where the model tends to underestimate values. Additionally, RF16 showcases a broader range of predicted values when the observed value is zero.

The model RF19's overall pattern appears more symmetric around the perfect regression line when compared to the nuances exhibited by RF16. In summary, the comprehensive analysis extends beyond the conventional scope, revealing intricate differences between the two models. It highlights how a higher R-squared, in the case of RF16, does not necessarily equate to a superior fit for all data points. Instead, a detailed examination of regression patterns for overestimations, underestimations, and zero-observed values uncovers the unique characteristics of each model, providing valuable insights into their performance and behaviour across the entire data spectrum.

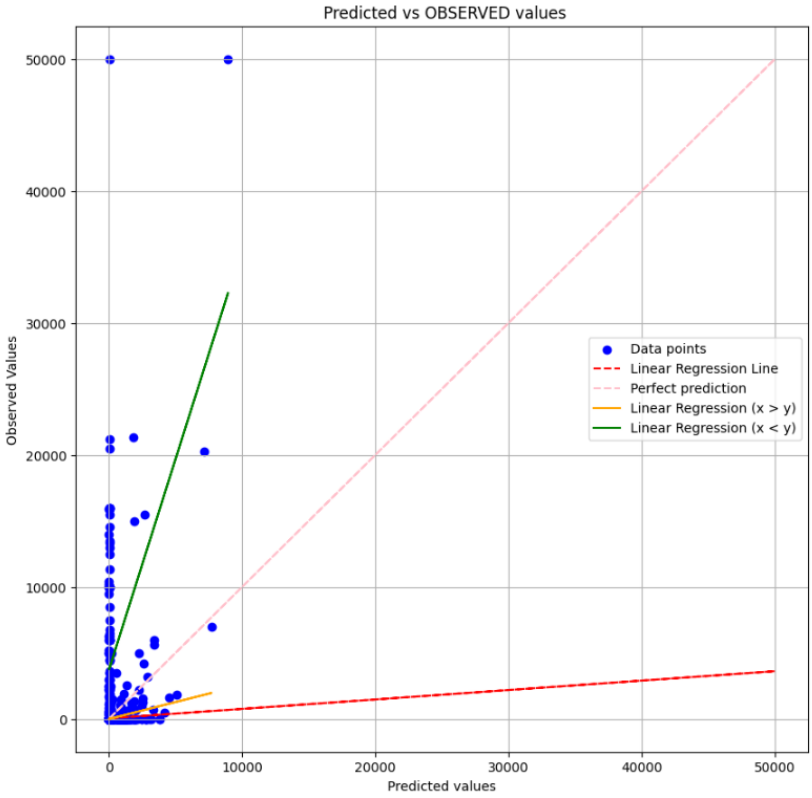

Figure 5- Result of Financed Amount model after response with temporal train/test split

In the context of a temporal train/test split, the importance of visually inspecting the model's performance becomes evident. Simply relying on quantitative metrics such as MAE and RMSE might lead to the assumption that the model is performing well. However, a closer examination reveals nuances that these metrics alone cannot capture. For instance, when evaluating the models graphically, it becomes apparent that the model with the lowest MSE (RF24) always predicts zero. This can be attributed to the high prevalence of target values equal to zero in the dataset. On the other hand, model RF25 which had an under-sampling of the majority class, exhibits better visual results, but it still encounters challenges as can be seen in Figure 5. It predicts values greater than zero for observations that are zero, and vice versa, indicating areas where the model struggles to capture the underlying patterns in the data.

4.1.3. Financed Amount Category after Response

Evaluating regression problems presents additional complexity due to the need to consider both quantitative metrics and graphical analyses. To address this challenge, an exploration was conducted to determine whether translating a regression problem into a classification problem could maintain or enhance the predictive quality.

Multiple models were developed using various binning techniques where each observation point was assigned a class. Initially, an experiment employing five bins was pursued; however, this method was deemed overly intricate for decision-making due to the presence of five distinct categories.

Consequently, a four-bin approach was adopted, incorporating feature selection and hyperparameter tuning. Upon evaluation, it was observed that all categories achieved an AUC of 0.50. This result indicated that the binning process was suboptimal, resulting in random classification.

Other approach was tested in which the observations were divided into two categories, with the first-class comprising observations up to the 50th percentile, and the second class including the remaining ones. Additionally, a three-bin split was tested, assigning class 1 to observations up to the 50th percentile, class 2 to those up to the 95th percentile, and class 3 to the remaining ones, in an attempt to capture customers with very high values.

The summaries of the alternative bin-splitting techniques are provided in Table 11 for reference, in this case the evaluation metrics were calculated per class in each model.

Table 11 - Results of Financed Amount category model

Train/ Test Split	Nº of quantiles	Algorithm	Model	Category	Precision	AUC	Recall	Feature Selection	Details
Random	3	RF	RF26	1	0.6008	0.6360	0.6360	All Features	Baseline
				2	0.5569	0.5945	0.5746		
				3	1	0.5125	0.0250		
			RF27	1	0.5938	0.5973	0.6045		Baseline with SMOTE
				2	0.5502	0.5915	0.5892		
				3	1	0.5375	0.0750		
	2	RF	RF28	1	0.6026	0.6074	0.6202		
				2	0.6106	0.6052	0.5902		
		Xgboost	X6	1	0.6042	0.5914	0.5213		
				2	0.5824	0.5914	0.6615		
		NN	NN8	1	0.5446	0.5529	0.6180		
				2	0.5630	0.5529	0.4878		
	3	Xgboost	X7	1	0.5903	0.5894	0.5730	RFE -30 features	
				2	0.5256	0.5689	0.5770		
				3	0.1538	0.5186	0.0500		
			XG8	1	0.5407	0.5425	0.5371	SFS -30 features	
				2	0.4813	0.5129	0.2836		
				3	0.0521	0.5204	0.2750		
			X9	1	0.5994	0.5824	0.4876	Random Forest	
				2	0.5233	0.5669	0.5770		
				3	0.1111	0.5703	0.2250		
			X10	1	0.5933	0.5962	0.6000	SFS -30 features	Tuned with parameter search
				2	0.5403	0.5718	0.5086		
				3	0.1017	0.5440	0.1500		
Temporal			X11	1	0.9527	0.9413	0.9283	SFS -30 features	Best random model
				2	0.8555	0.7384	0.5547		
				3	0.1312	0.7015	0.6039		

As seen in Table 11, initially, a test was conducted using three bins, and subsequently, the impact of employing the SMOTE approach on the minority class was evaluated. This adjustment resulted in an increase in the recall of the third category, which represents loans with the highest financed amount. To further assess model performance, a comparison was made between three algorithms: NN, RF, and Xgboost, using two bins.

Notably, Xgboost exhibited the highest recall for category 2, although it demonstrated slightly inferior results in the other categories when compared to NN and RF.

Since it was challenging to directly determine which approach, either 2 bins or 3 bins, was superior based solely on model performance metrics, a cumulative gain plot of revenue was generated to compare the two approaches using their best models in terms of recall (RF27 and X6) as it can be seen on Figure 6. In both cases, the client data was ordered based on prediction values for the categories in descending order. This ordering allowed the customers with category 3 to appear at the top, and a similar approach was applied for the result with 2 bins. The plot in Figure 6 illustrates the model's performance in terms of the potential revenue generated when selecting a sample of customers.

The model utilizing 3 bins demonstrated higher revenue for some given samples, highlighting the potential advantages of this approach.

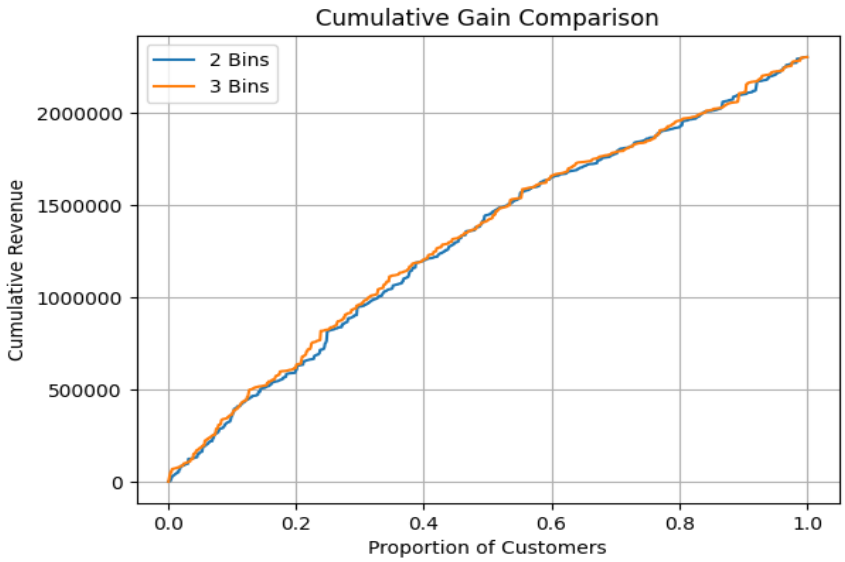

Figure 6 Cumulative gain comparison of different financed amount bin approaches

Subsequent efforts were directed toward improving the model's performance through feature selection and hyperparameter tuning. It was observed that Model X8, despite demonstrating the highest recall for Category 3, exhibited inferior results for Category 2 when compared to Model X10. This discrepancy was significant, considering that Category 2 had a substantially larger sample size than Category 3. Consequently, Model X10 seemed to provide better overall compensation.

Regarding the model from temporal train-test split validation (X11), the application of temporal validation after the response results led to an extremely imbalanced dataset. Using percentiles, as done for the random train-test split, resulted in non-meaningful class splits, with categories 1 and 2 exclusively comprising zeros and category 3 containing all non-zero observations. As a consequence, the results lack meaningful comparability. To address this issue, a new method for category splitting, such as defining values according to euros intervals, is suggested. Although this approach may be less dependent on good response results, it introduces new challenges, such as imbalanced categories. Due to the unreliable nature of the results, the model using this method will not be compared to the others.

4.1.4. Financed Amount

For the models that forecasted the financed amount without initially predicting the response variable, the analysis involved the utilization of the Random Forest algorithm. The detailed outcomes and corresponding metrics for these models are explicitly presented in Table 12.

Table 12 Results of Financed Amount model without predicting response before

Train/Test Split	Model	MAE	RMSE	R2	Feature Selection	Details
Random	RF29	1054	3472	-0.0400	All features	Baseline
	RF30	1062	3509	-0.0622	Random forest features	
	RF31	1058	3484	-0.0471	Random forest features	Tuned with parameter search space
Temporal	RF32	77	344	-0.6370	Random forest features	Best Random model

For the models predicting the financed amount, without first predicting the target variable, the Random Forest algorithm was employed. The baseline model, which incorporated all available features, exhibited marginally superior performance compared to the model that underwent feature selection and hyperparameter tuning. It's important to note that due to time constraints, limited experimentation was conducted on this model. As such, there may have been untapped potential for achieving better results had more extensive optimization efforts been pursued, which bias the results.

In the temporal train/test split, upon analysing the Figure 7, it becomes evident that several observations tend to overestimate values that are equal to 0, while the underestimations are more prominent for the highest observed values.

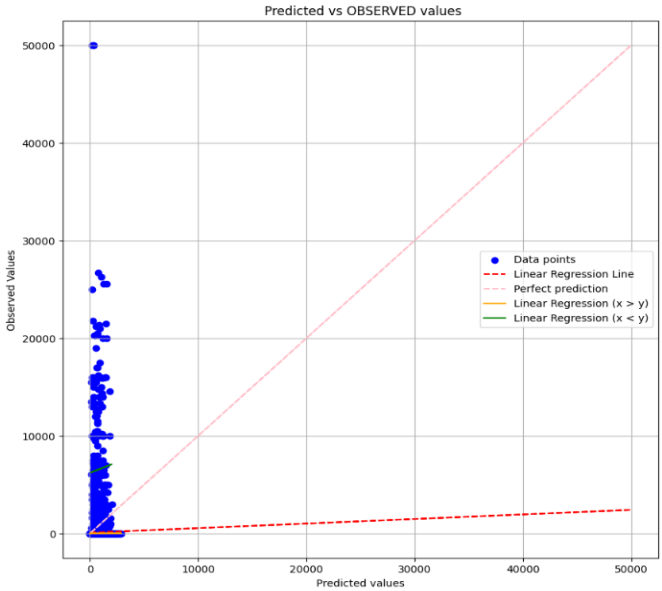

Figure 7 Result of Financed Amount model without response with temporal

4.2. Comparison of best models in terms of Cumulative Gain of Financed Amount

Analysing the cumulative gain comparison among the best models in Figure 8 – including the response, financed amount after response, financed amount category after response, financed amount without predicting response, and random sorting – reveals valuable insights.

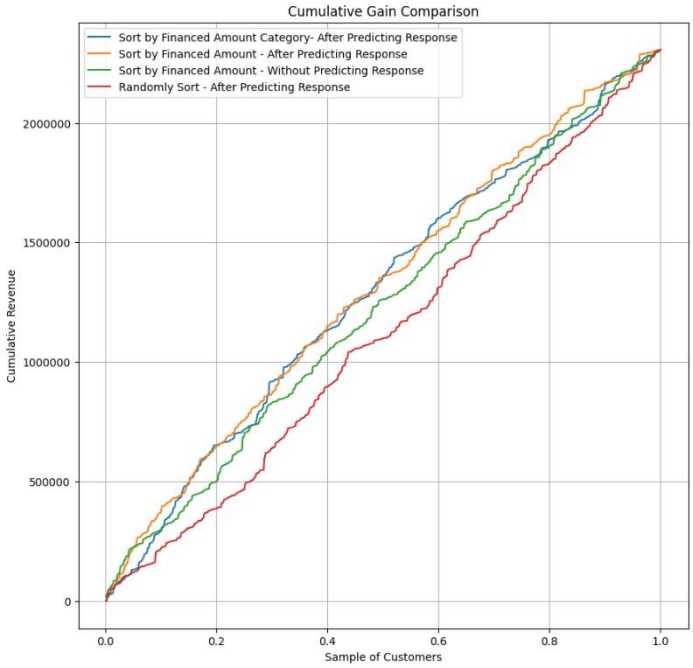

Figure 8 Cumulative Gain comparison of Financed Amount between the best model of each random train/test split approach

In Figure 8, the impact of each approach on the revenue gained from the customer sample is demonstrated. The comparison of model outcomes was based on the test sample population of the Financed Amount approach, specifically after predicting positive responses. All models converge to the same revenue value since they are evaluated on the same population. Thus, with 100% of the population, the revenue remains consistent across all models.

Observing the graphical representation, it's evident that the response approach results in comparatively less revenue when the communication count is reduced. This approach consistently remains below the other lines in the graph, except when reaching maximum revenue. The prediction of financed amount categories and financed amount itself yields superior results, with the former slightly outperforming the latter. This improvement may be attributed to a more profound analysis of customer values, enabling a more precise customer ranking.

It's important to note that the model "financed amount without predicting response" was not as extensively fine-tuned as the others. As a result, the cumulative gain comparison in this context does not definitively indicate which approach is better. For a clearer analysis, ensuring that the test sample population remains the same across all models would provide a more accurate assessment of their performance. This observation pertains to the performance of the predictions in the financed amount model without predicting responses, instead of ranking customers based on these predictions, thus the Cumulative Gain present in Figure 9 can provide some more insights.

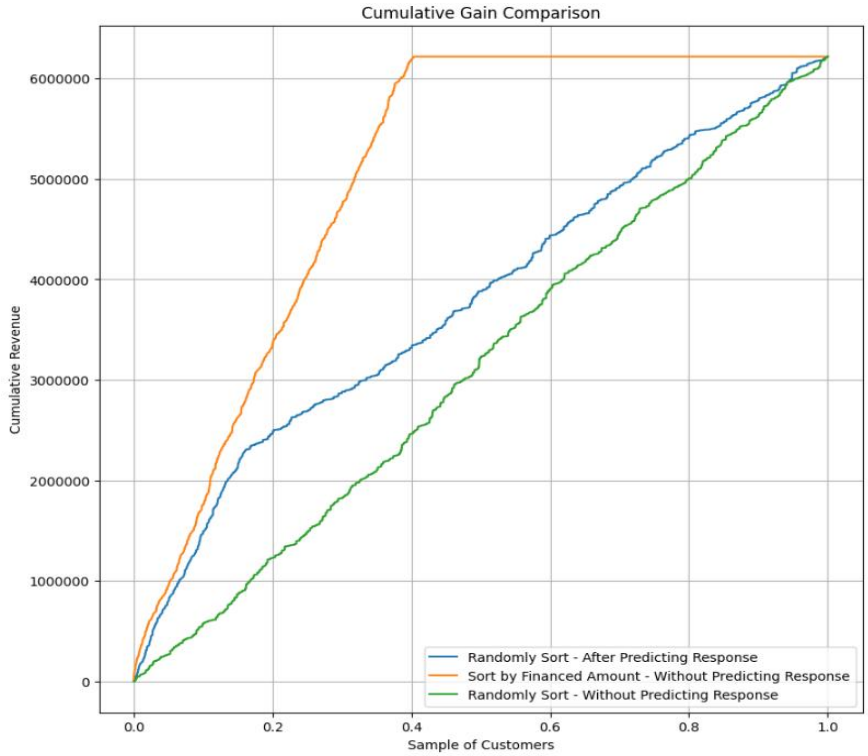

Figure 9 Cumulative Gain Comparison on test population for random test/split

In Figure 9, a direct performance comparison can be made between ranking customers based on Financed Amount and Response, using the same test population for ranking purposes. Initially, up to 20% of the test size population, both methods—ranking customers based on financed amount and response—displayed similar observations. Notably, the Financed Amount Ranking approach allowed capturing the maximum potential revenue by contacting only 40% of the population, while the Response method required reaching 100% of the population for similar results.

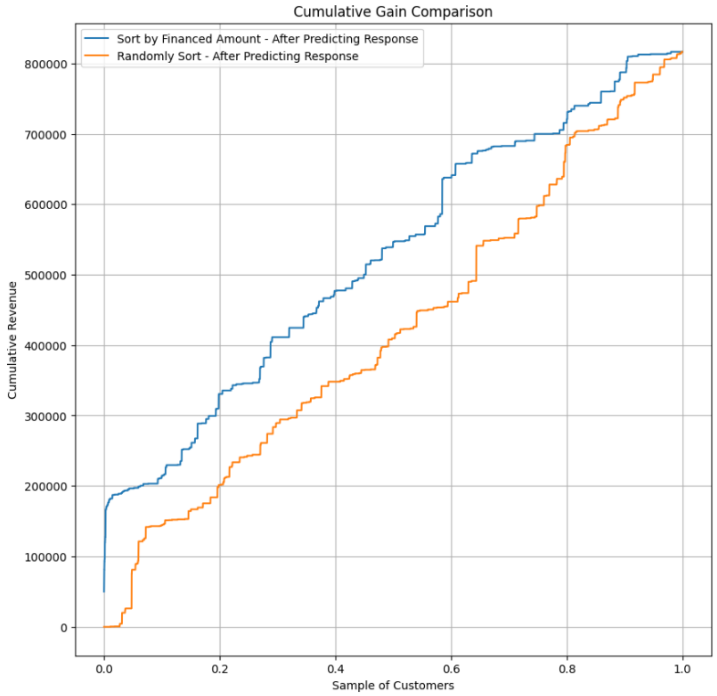

Figure 10 Result of best models for temporal train/test split

The cumulative gain comparison was also made for the temporal split as observed in Figure 10. In this, the best performing was predicting the financed amount, thus reinforcing the positive outcome of predicting the financed amount. An observation from the graphical data reveals that by selecting 20% of customers predicted with a positive response and ranking them by the financed amount, there's an increase of over 150 thousand in revenue compared to randomly selecting 20% of customers.

CHAPTER 5

Conclusion and Future Work

5.1. Summary

Optimizing telemarketing targeting is a crucial subject under the growing pressure to increase profits and reduce expenses in the banking sector (Moro *et al*., 2014).

In this context, the utilization of predictive modelling for identifying positive respondents in direct marketing campaigns proves to be invaluable. Not only does it enable the identification of potential respondents, but if there is an opportunity to discern not just the response likelihood but also the value associated with customers who are likely to respond, it allows for more informed decision-making. This knowledge empowers a more strategic approach, facilitating significant cost reductions in communication. The ability to select the top-value customers among those likely to respond positively offers a targeted and efficient method for campaign optimization.

Throughout this study, the analysis compared the revenue outcomes of predicting the value of customers after a positive response, specifically the financed amount in loans, against the conventional approach of predicting only the response of customers. This comparison was conducted for the selection process of a personal finance company, particularly in scenarios aiming to reduce the targeted population. The dataset utilized in this analysis pertained to a French Bank's personal finance company and encompassed over 4 million observations related to marketing campaigns conducted between 2019 and 2021.

The primary objective was to model the financed amount of a loan following a marketing communication and to compare these results with modeling the response alone. The research also aimed to identify the optimal approach for modeling the financed amount. This involved exploring whether it would be more effective to predict the response first and then the financed amount, considering whether the financed amount should be predicted as a category or a numerical variable, or if it would be more advantageous to solely predict the financed amount without considering the response. The modeling task involved utilizing features related to the customer's existing loans, revolving card details like card quantity and card usage, as well as sociodemographic features.

Additionally, new features were created, inspired by the RFM analysis. The training and testing of the models incorporated two distinct approaches: a random split and a temporal test. The random split method aimed to predict these targets across various months, capturing trends within the dataset. In contrast, the temporal test was based on a holdout for the four most recent months of available data, accounting for potential leaks that random testing might introduce. This approach considered the evolution of time-sensitive features such as the number of previous communications.

Moreover, the temporal test aimed to simulate a more real-life scenario, offering added value when presenting the results to non-technical stakeholders. A concentrated effort was made to reduce the initial set of more than 140 features to 30 during the modeling phase. This reduction was made in an effort to be consistent with the patterns seen in the literature review of related publications, which indicated that using a reduced set of features could potentially improve the model's performance. Three different feature selection techniques were attempted in order to accomplish this: Sequential Forward Selection (SFS), Recursive Feature Elimination (RFE), and taking into account the most significant features found through random forest's output analysis. Also, four algorithms were compared: random forest (RF), Xgboost and neural networks (NN).

For the comparison of models predicting categorical targets, three key metrics were utilized: the area under the receiver operating characteristic curve (AUC), precision, and recall. Emphasis was placed on the recall metric due to the context, where sending a communication to a customer likely to convert holds greater value than the cost associated with communicating to a non-respondent customer. In the case of numerical features, the primary evaluation metrics were Mean Absolute Error (MAE) and R-squared. Additionally, a graphical analysis was performed, plotting observed versus predicted values and examining linear regression when the model was overestimating and underestimating, offering further insights into the model's performance.

Data imbalance presented a significant challenge due to the rarity of positive responses, accounting for less than 1% of the communication observations. To address this issue, the study explored various approaches, including using Synthetic Minority Over-sampling Technique (SMOTE) combined with undersampling techniques. Furthermore, the study applied a bagging approach during the undersampling process, allowing for the integration of different models trained on various samples of negative observations. This technique effectively addressed sample bias and significantly improved the model's predictions by aggregating distinct models. The analysis compared this approach against other methods, including hyperparameter tuning for class weights and custom loss functions. The custom loss function utilized the dataset's mean financed amount to represent the cost of misclassified positive responses. Ultimately, the results revealed that the hyperparameter underperformed compared to undersampling and SMOTE, while the custom loss function achieved comparable results. Moreover, the Box Cox transformation was applied to the numerical financed amount, which demonstrated skewness since the higher financed amount loans are rarer.

For the final comparison of modeling approaches and to assess their impact on revenue gain, customers were ranked based on the predicted feature using the best model from each approach.

Subsequently, cumulative gain plots were generated for all approaches against randomly selected customers from a sample. In both the train/test split methods, the strategy that consistently generated the highest revenue among most sample sizes of positively predicted customers involved predicting the financed amount subsequent to forecasting the response. Furthermore, the holdout approach in predicting the response demonstrated a potential reduction of up to 91% in communications. However, focusing on predicting the financed amount allowed for even greater reductions while maximizing revenue, surpassing the approach of merely sampling customers identified as positive. Specifically, selecting the top 20% of predicted positive respondents and ranking them based on the financed amount could result in over €150,000 difference in revenue compared to randomly selecting 20% of the customers.

5.2. Research Questions

To what extent does predicting the customer value in direct marketing within the domain of personal finance contribute to enhancing Return on Investment and reducing overall marketing costs?

Answering the research question regarding the added value of predicting customer value in direct marketing, rather than solely focusing on response, yields a positive response. This research demonstrates that maximizing ROI is possible by predicting the value of the customer, in this case, the financed amount. Such an approach captures more revenue while contacting a smaller sample of customers, assuming communication costs remain constant. The suitability of this approach depends on various factors. If time and resources are constrained, a binary classification model for response is a simpler, more interpretable, and quicker solution since predicting the financed amount, being a numerical feature, involves a more complex process. Transforming the regression problem into a classification one also requires decision-making and analysis to determine the best approach and number of categories. The choice also depends on the company's strategy. This is advantageous when the company must curtail the budget for direct marketing. However, there are certain constraints to consider. If the company has a minimum volume requirement for communications or faces pressure to spend a specific budget on direct marketing, and the number of positively predicted respondents for the response is needed to meet these constraints, predicting the financed amount after the response won't be as useful, as both approaches will ultimately yield the same revenue value.

Nonetheless, if time is not a constraint, predicting the value gives more detailed information on expected customer value, which can be of added value even if there is no intention to proceed with the cut-off of communications sent.

What approach or model offers the most effective means to predict customer value in the context of direct marketing for personal loans?

Direct comparison among the three approaches—Financed Amount, Financed Amount after Response, and Financed Amount Category after Response—highlighted improved performance in the latter two. However, as noted, the discrepancy in test samples used for both models prevent a conclusive determination, indicating the promising nature of the alternative approaches.

5.3. Research Limitations

This study encountered several limitations despite its promising outcomes. The summary below highlights these restrictions and offers recommendations for further research:

- **Data Availability Constraint:** The study faced a significant limitation due to internal company policies, resulting in data availability only a few months before the project deadline. This imposed a stricter time constraint and limited the depth of exploration in model development.
- **Varying Model Depth:** Owing to the time constraints, not all models were explored with the same depth. This could introduce bias to the results as models were compared, even if some were not tuned to the same extent as others.
- **Complex and Large Datasets:** Initial data tables were notably vast, demanding extensive time for analysis, identification of content-related issues, and preprocessing. Additionally, the interconnection between tables resulted in a substantial time investment in reconstructing the final database if any issues were identified.
- **Infrastructure Limitations:** As the data was private and processed on the company's server, limited RAM often caused notebooks to crash, prolonging the time required to achieve results.
- **Cost Analysis Oversight:** The study did not consider the real costs associated with different types of communication. Incorporating this information could yield more insightful results in terms of ROI, factoring not only revenue but also the associated costs.
- **Uplift Testing Missed Opportunity:** At the study's onset, controls—individuals with identical characteristics to those who received communication but were not contacted—were excluded. This decision missed the chance to conduct uplift testing, which could have significantly impacted the ROI analysis by neglecting the aspect of campaign efficiency.

5.4. Future Work Recommendations

The study proposes the following recommendations for further research:

- **Customer Segmentation Approach and Consistent Positive Predictions:** Further investigations might benefit from a more nuanced customer segmentation approach. Kane *et al.* (2014) proposed four distinct customer types, categorizing those unlikely to convert due to negative perceptions of marketing campaigns, those who convert regardless of communication, individuals influenced to convert if contacted, and those unlikely to convert when contacted but might do so if left uncontacted. Not deeply considering these customer types in the study might have resulted in incomplete addressing of consistently predicted positive respondents across different algorithms and hyperparameter tuning. Not differentiating between these customer types could have partially contributed to the consistent prediction of certain customers as positive respondents despite changes in modeling techniques. Inclusion of these customer typologies into the response model could significantly improve the study's outcomes, representing a promising area for future investigations.
- **Addressing Multi-Communication Impact Within the Conversion Time Frame:** Future studies could address scenarios where customers received multiple communications within the timeframe leading to a positive response prediction. For instance, in this study, if a customer converted within 60 days, understanding which specific communication or combination of communications within that time frame had the actual positive impact on customer response is crucial. Creating a feature to differentiate the impact of multiple communications as a single or distinct influence within the conversion period can enhance analysis accuracy and contribute to a more comprehensive ROI assessment.
- **Refinement of Testing Approaches in Response-Based Models for Consistency in Results:** For a more thorough evaluation, future research should explore alternative testing methods for response-based approaches. Ensuring that the test sample is the same for all approaches and enabling direct comparisons between different models would enhance the study's validity and comprehensiveness. Specifically, employing the same test sample across various approaches, potentially training the model for financed amount after a response on the same train set as other models, could provide more robust insights.

References

Agarwal, K., Jain, M., & Kumawat, A. (2021, September 5). Comparing Classification Algorithms on Predicting Loans. *Information Systems and Management Science*, 240–249. https://doi.org/10.1007/978-3-030-86223-7_21

Asare-Frempong, J., & Jayabalan, M. (2017, September). Predicting customer response to bank direct telemarketing campaign. *2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*. https://doi.org/10.1109/ice2t.2017.8215961

Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. (2002, April). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*(1), 191–211. https://doi.org/10.1016/s0377-2217(01)00129-1

Bose, I., & Chen, X. (2009, May). Quantitative models for direct marketing: A review from systems perspective. *European Journal of Operational Research*, *195*(1), 1–16. https://doi.org/10.1016/j.ejor.2008.04.006

Chlebus, & Osika. (2020). *Comparison of tree-based models performance in prediction of marketing campaign results using Explainable Artificial Intelligence tools*. University of Warsaw, Faculty of Economic Sciences.

Durango-Gutiérrez, M. P., Lara-Rubio, J., & Navarro-Galera, A. (2021, January 26). Analysis of default risk in microfinance institutions under the Basel III framework. *International Journal of Finance & Economics*, *28*(2), 1261–1278. https://doi.org/10.1002/ijfe.2475

Elsalamony, H. A., & Elsayad, A. M. (2013). Bank direct marketing based on neural network and C5. 0 Models. *International Journal of Engineering and Advanced Technology* (IJEAT), 2(6).

Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y., & Harb, A. (2020, April 9). Business Analytics in Telemarketing: Cost-Sensitive Analysis of Bank Campaigns Using Artificial Neural Networks. *Applied Sciences*, *10*(7), 2581. https://doi.org/10.3390/app10072581

Gu, J., Na, J., Park, J., & Kim, H. (2021, August 2). Predicting Success of Outbound Telemarketing in Insurance Policy Loans Using an Explainable Multiple-Filter Convolutional Neural Network. *Applied Sciences*, *11*(15), 7147. https://doi.org/10.3390/app11157147

Guido, G., Prete, M. I., Miraglia, S., & De Mare, I. (2011, August). Targeting direct marketing campaigns by neural networks. *Journal of Marketing Management*, *27*(9–10), 992–1006. https://doi.org/10.1080/0267257x.2010.543018

Han, X., Wang, L., & Huang, H. (2017). Deep Investment Behavior Profiling by Recurrent Neural Network in P2P Lending. ICIS 2017 Proceedings, 1–11. http://aisel.aisnet.org/icis2017/Peer-toPeer/Presentations/11

Jin, Y., & Zhu, Y. (2015, April). A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending. *2015 Fifth International Conference on Communication Systems and Network Technologies*. https://doi.org/10.1109/csnt.2015.25

Kane, K., Lo, V. S., & Zheng, J. (2014, December 19). Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods - Journal of Marketing Analytics. SpringerLink. https://doi.org/10.1057/jma.2014.18

Kim, K. H., Lee, C. S., Jo, S. M., & Cho, S. B. (2015, November). Predicting the success of bank telemarketing using deep convolutional neural network. In 2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR) (pp. 314-317). IEEE.

Koumétio, C. S. T., Cherif, W., & Hassan, S. (2018, October). Optimizing the prediction of telemarketing target calls by a classification technique. In *2018 6th International conference on wireless networks and mobile communications (WINCOM)* (pp. 1-6). IEEE.

Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017, June 28). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, *9*(1). https://doi.org/10.1186/s13321-017-0226-y

Ładyżyński, P., Żbikowski, K., & Gawrysiak, P. (2019, November). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems With Applications*, *134*, 28–35. https://doi.org/10.1016/j.eswa.2019.05.020

Liu, X., & Han, L. (2022, June 17). Artificial Intelligence Enterprise Management Using Deep Learning. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/2422434

Moro, S., Cortez, P., & Rita, P. (2014, June). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001

Moro, S., Cortez, P., & Rita, P. (2015, February). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems With Applications*, *42*(3), 1314–1324. https://doi.org/10.1016/j.eswa.2014.09.024

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology.

Piatetsky-Shapiro, G., & Masand, B. (1999). Estimating campaign benefits and modeling lift. In Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 185-193).

Piñeiro, G., Perelman, S., Guerschman, J. P., & Paruelo, J. M. (2008, September). How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling*, *216*(3–4), 316–322. https://doi.org/10.1016/j.ecolmodel.2008.05.006

Raschka, S. (2018). Mlxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. The Journal of Open Source Software, 3(24), 638.

Turkmen, E. (2021, February 4). Deep Learning Based Methods for Processing Data in Telemarketing-Success Prediction. *2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*. https://doi.org/10.1109/icicv50876.2021.9388467

Youqin Pan, & Zaiyong Tang. (2014, June). Ensemble methods in bank direct marketing. *2014 11th International Conference on Service Systems and Service Management (ICSSSM)*. https://doi.org/10.1109/icsssm.2014.6874056

Zhang, M. (2022, May 12). Research on Precision Marketing Based on Consumer Portrait from the Perspective of Machine Learning. *Wireless Communications and Mobile Computing*, *2022*, 1–10. https://doi.org/10.1155/2022/9408690

Zhang, M., Sun, J., & Wang, J. (2022, April 9). Which neural network makes more explainable decisions? An approach towards measuring explainability. *Automated Software Engineering*, *29*(2). https://doi.org/10.1007/s10515-022-00338-w

Zhang, T., & Yang, B. (2017, April 3). Box–Cox Transformation in Big Data. *Technometrics*, *59*(2), 189–201. https://doi.org/10.1080/00401706.2016.1156025

Appendix

Table 13 Hyperparameters used in the non-Baseline models.

Model	Hyperparameters
RF6, RF7	max_depth= 8, min_samples_leaf= 10, min_samples_split=2, n_estimators=500
NN6	activation='relu', dropout_rate=0.4571049918680663, hidden_layers=2, neurons= 32, batch_size= 428, epochs=50, learning_rate=0.009997926691408476
X2	base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, eval_metric='mlogloss', gamma=0, gpu_id=-1, importance_type='gain', interaction_constraints='', learning_rate=0.3, max_delta_step=0, max_depth=6, min_child_weight=1, monotone_constraints='()', n_estimators=100, n_jobs=16, num_parallel_tree=1, objective='binary:logistic', random_state=0, reg_alpha=0, reg_lambda=1, scale_pos_weight=None, subsample=1, tree_method='exact', use_label_encoder=False, validate_parameters=1, verbosity=None
X3	base_score=0.5, booster='gbtree', colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1, eval_metric='mlogloss', gamma=0, gpu_id=-1, importance_type='gain', interaction_constraints='', learning_rate=0.3, max_delta_step=0, max_depth=6, min_child_weight=1, monotone_constraints='()', n_estimators=100, n_jobs=16, num_parallel_tree=1, objective='binary:logistic',

	random_state=24, reg_alpha=0, reg_lambda=1, scale_pos_weight=None, subsample=1, tree_method='exact', use_label_encoder=False, validate_parameters=1, verbosity=None
RF14, RF15, RF16, RF17	n_estimators=264, min_samples_split = 2, min_samples_leaf=4, max_depth=32
RF19, RF20, RF24, RF25	n_estimators=100, min_samples_split = 2, min_samples_leaf=4, max_depth=22
RF22	max_depth= 6, min_samples_leaf= 9, min_samples_split= 29, n_estimators = 101
RF23	max_depth= 4, min_samples_leaf= 9, min_samples_split= 15, n_estimators = 365
X10, X11	learning_rate= 0.18322330494908368, max_depth=5, n_estimators=131
RF31	max_depth= 31, min_samples_leaf=3, min_samples_split=2, n_estimators= 500