Back to the past to charter the vinyl electronic market: a data mining approach

Sara Lousão[1], Pedro Ramos[2,3] and Sérgio Moro[2]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), Lisboa, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR-IUL, Lisboa, Portugal
[3] Instituto Universitário de Lisboa (ISCTE-IUL), IT-IUL, Lisboa, Portugal
Sara_Carolina_Lousao@iscte-iul.pt
pedro.ramos@iscte-iul.pt
scmoro@gmail.com

Abstract. This study focuses on perhaps the most iconic media format of all time, the vinyl record. By adopting a data mining approach, the goal is to understand which factors involved in the buying and selling of vinyl, influenced its price, with the initial hypothesis considering record labels and popular rankings to be some of the most contributing variables. To be able to evaluate it, four datasets were created in an endeavor to represent recent and past records of two different genres, Rock and Jazz, by extracting data from Discogs' marketplace and Billboard's Hot 100 chart. Such approach unveiled that an artist's presence in the charts and their labels belonging to one of the 'Big three' do not always dictate their records at highest prices. The results also showed that features which measure popularity become more relevant in the 'era' where the record's genre is more popular and that big record labels have been losing market share to an increasing number of independent labels.

Keywords: Vinyl record, pricing, data mining.

1 Introduction

Vinyl was one of the first formats for audio reproduction, created around 1920, managing to become increasingly more popular until the 80's, when the invention of the Compact Disc (CD) finally replaced the vinyl. This was mainly due to the lower costs of production of the CD, as well as requiring less space and maintenance, becoming easier to distribute. Interestingly, vinyl made a comeback, mainly due to avid collectors who unknowingly created a community that kept the format alive. Vinyl records reached, in 2008, its highest sales number since 1991 and reemerged as one of the preferred music formats. The sudden boom in vinyl sales could be explained by their intrinsic high-fidelity sound, its physicality, tactile and aesthetic appeal when compared to digital audio files and, last but not least, their vintage feel in a market that was and still is partially driven by old-fashion consumerism [1]

Despite all the changes that the physical product for the delivery of music has undertaken, the industries' division of labor and hierarchies have remained relatively

stable. Artists are the ones that create music whereas record labels help them produce, promote and distribute the final product, which will finally be consumed by the fans. In the USA, the music market in 2007 was mainly dominated by only five major record labels, so-called 'Big five', which were, namely, EMI, Sony, Universal-Vivendi, Time Warner and Bertelsmann BMG, controlling both production and distribution of music records. This market share division became even more favorable for the soon to be 'Big three', when, in 2008, Sony bought the remaining 50% stake held by BMG, thus acquiring them [2]. In the following years, EMI also met its demise, ultimately selling most of their branches to Time Warner, Universal [3] and Sony [4]. There is some believe that it is due to the existence of a monopoly capable of controlling the entire supply chain that artists have been prevented of independently producing and distributing their own material. This explains the fact that record labels are the ones who get to keep around 85 to 90% of the profit generated from music sales, using their influence to generate changes in technology to enhance their dominant position.

In the music industry, a popular music chart conveys the complexity of relationships among several factors such as business, musicians, music and consumer. It directly shows the state of the music business even to the most alienated of consumers. Music charts not only define what is popular but, more importantly, help shape the definition of popularity itself [5]. The most pertinent chart, that is constantly mentioned across the available literature, is the particular case of the Billboard magazine's charts which currently stands as a model of a universal ranking system [6].

With the evolution and development of the Internet, the vinyl marketplace, like many other markets that started taking some form online, enabling to share experiences, tastes and contributing to the evolution and preservation of vinyl records. Discogs leads the business of reselling vinyl records in electronic markets [7], since it is one of the most important web resources for anyone who wants to identify, locate, sell or buy any physically recorded media, due to its extensive database coverage of vinyl records and almost 140,000 contributors in 2016 [8].

One of the most valuable assets of any business is the information collected about how their costumers interact with one's products or services. Among the large amounts of data that result from these interactions, resides powerful information that can help shape a business or organization in their eternal quest for competitive advantages over the market. To uncover this hidden knowledge and draw the big picture of whatever insights a business needs to succeed, it is necessary to continuously collect, store, process and analyze the vast datasets that result from daily operation. To enable the classification, discovery of patterns and trends or prediction of possible outcomes the application of increasingly complex statistical tools combined with machine learning algorithms are used [9]. The usage of these automated processes has earned the designation of Data Mining and its results can be potentially applied to many different problems such as decision support, prediction, forecasting and estimation, providing crucial aid in important business decisions [10]. One of the many areas of application that has greatly benefited from the insights that Data Mining processes can yield is e-commerce, which translates to commercial trades made online through websites or mobile applications.

2 Materials and Methods

Web scraping is described by the automatic crawling of a website with the intention of collecting data, either by using scripts written for a specific task or by resorting to tools developed to extract information from the web [11]. For this research's web scraping process using the R programming language, the chosen data sources were two music genres from the vinyl marketplace Discogs (Rock and Jazz) and the Hot 100 chart from the ranking website Billboard (only chart with information from 1958 to the 1980's).

For the Discogs extractions regarding Rock and Jazz, the number of observations was 6,200 and 7,900 respectively, whereas for the Hot 100 chart, the gathered records for the time interval of 1958 and 1980 were 116,785 while, for the 2008 to 2019 interval, these were 58,200. Table 1 shows the features selected.

Table 1. List of features.

Feature name	Source	Description
label	Discogs	Label that published the release.
avg_vinyl_rating	Discogs	Average rate of the release.
nr_users_have	Discogs	Number of users that have the release.
nr_users_want	Discogs	Number of users that want the release.
media_condition	Discogs	Condition of the vinyl record.
sleeve_condition	Discogs	Condition of the vinyl's sleeve.
median_price	Discogs	Median price for the release.
release_year	Discogs	Year of the release.
artist	Billboard	Name of the artist in the chart.
artist	Computed	Extraction of artist from Discogs title feature.
min_peak_position	Computed	Best position in rank achieved by an artist with a song.
max_weeks_on_chart	Computed	Maximum number of weeks an artist stayed in rank with a song.
year_on_chart	Computed	Year extracted from week feature.

SAS Enterprise Guide was used to transform data and generate histograms, in order to help visualize which features are best related to price for each genre and each considered time interval.

The rows where missing values existed were removed to guarantee that the data is coherent. Once those were eliminated, it was possible to convert the nominal scale used to represent media and sleeve conditions into an ordinal scale from 0 to 9, where 0 and 1 stand for the values 'No Cover' and 'Generic' respectively (and are only

applicable to sleeve condition) and the remaining values from 2 to 9 range from 'Poor (P)' to 'Mint (M)' condition.

Regarding the datasets extracted from Billboard's Hot 100, the missing values for 'min_peak_position' were attributed the value of 9999, since a lower number represents a better position. The opposite was done with 'max_weeks_on_chart' where a larger number means more time on charts. As such, the missing values were replaced by 0. The two datasets were merged into a single dataset to determine which was the peak position each artist had been able to achieve in this rank and, at the same time, check the maximum number of weeks they had been in the rank. It is important to point out that both peak position and number of weeks on chart are connected to a specific song by an artist and the calculated result was applied to the artist alone. This means that a given artist might have reached its peak position with one song while beating the record of weeks on chart with another. In order to be able to remove the repeated artist occurrences, while keeping the best position in chart for an artist and the longest time that artist remained in the chart with a specific song, the resulting dataset from Billboard was filtered by selecting only the rows where the number of weeks on chart matched the maximum number of weeks on chart.

This finally enabled the merge between these results and each of the generated datasets for Discogs, attributing minimum peak position, maximum weeks on chart and the year on chart to the artists in Discogs Rock and Jazz data that also appeared in the resulting Billboard's Hot 100 records. For the remaining artists, that were not among the Billboard's Hot 100 artists, these three new columns acquired dummy values of 9999 (for year on chart and minimum peak position) and 0 (for maximum weeks on chart).

In the end, both Rock and Jazz datasets were divided by 'release_year' to consider the golden era of vinyl between 1958 and 1980 and its reemergence post 2008. This resulted in two datasets for each genre with 2,642 entries for Jazz's past records and 100 for more recent releases while the datasets for Rock's past and present records comprised 1,832 and 831 entries respectively.

To obtain a better knowledge of, not only what conditions price, but also what changes between old releases and new releases for these two music genres, data modeling was applied with SAS Enterprise Miner to each of the considered genres and 'eras'. To better understand the datasets generated and what really influences the price of a record, it was necessary to first choose the target variable and then try to discover which of the remaining features had the most direct relation with the target. The variable chosen was 'median_price', which results from a calculation made by Discogs for the average price of a release.

3 Results and Discussion

The plots contributed to understanding what some of the important factors for price on the Discogs marketplace were, such as 'avg_vinyl_rating' and 'nr_users_want', as can be expected from any market, since the quality and popularity of the product are always the main contributors for price. However, the goal of this research is more

focused in trying to discover if the popularity of the artist and the influence of its record label contributed to a higher pricing of their records. As such, each of the four datasets was filtered to create a smaller sample where only the entries that presented a 'median_price' above 30 were considered. This value was chosen as the lower bound for the 'expensive' records filter, since according to Palm [12], the average price of a record is around 25 dollars, and the goal of this step was to single out the records priced well above average. This showed that, although there were a few extremely popular artists, where some even registered a 'min_peak_position' for the time intervals considered, these did not 'dominate' the samples.

Regarding the 'label' diversity, such different labels would easily lead to the assumption that 'the big three' lobby was not obvious in the gathered data. Nevertheless, by researching these recording companies and looking at their histories, their huge market share becomes very clear, as a lot of these businesses (Table 2) were eventually purchased by either one of the three. Although these major record labels are also represented for 'present' records, it is a lot more common in the 'past' datasets, since the music industry has changed drastically throughout the years and, many of the first labels did not grow big or fast enough to resist being acquired by the few ones that did. In addition, it is also important to report that the decrease in the presence of 'big' recording companies is also clearly contrasted by a much larger number of independent labels, self-released records and a select few that achieved such success, managing to create their own companies and retrieve all the rights to their records.

Table 2. Labels ownership and dataset distribution examples.

Owner	Nr Labels	Rock Past	Jazz Past	Rock Present	Jazz Present
Sony Music Entertainment	6	1	5	-	-
Warner Music Group	6	4	-	2	-
Universal Music Group	7	3	3	1	-
Independent	14	2	-	10	2

One of the main disadvantages of using black box models (computer generated mathematical models that cannot be directly interpreted) is the inherent difficulty in understanding what or why the model was generated in a certain way and how the target's relationship with its characterizing features has shaped that same model. Techniques such as sensitivity analysis or rule induction from networks were developed, enabling a better importance assessment of the input variables [9]. In accordance with the data presented in Figure 1, it is possible to observe some patterns in what variables play the largest and the smallest role in the evaluated datasets. It is important to mention that the variables 'nr_users_want' and 'avg_vinyl_rating' are not present in this plot since they were overshadowing the remaining contributing factors.

For the case of past Rock records, the variables 'sleeve_condition' and 'media_condition' present an interesting difference among the remaining features. In contrast, for its present counterpart, these sleeve and media conditions were the least significant, with 'min_peak_position', 'max_weeks_on_chart' and 'year_on_chart' appearing as more relevant, which shows some interesting differences between 'eras'. As for the Jazz datasets, for the one representing past releases, the most interesting pattern is in how similarly to the present Rock records, 'min_peak_position', 'year_on_chart' and 'max_weeks_on_chart' also represent important factors. Oppositely, in the set of data representing Jazz present, it is possible to, just like in Rock past, see 'media_condition' and 'sleeve_condition' maintaining some relevance. These similarities across genres and eras motivate a comparison between the Rock and Jazz datasets. As mentioned, it is possible to observe in Figure 1, by focusing on Jazz past and present or Jazz and Rock past, how the variables 'min_peak_position', 'year_on_chart' and 'max_weeks_on_chart' tend to switch places in importance with 'sleeve_condition' and 'media_condition' across genres and eras.

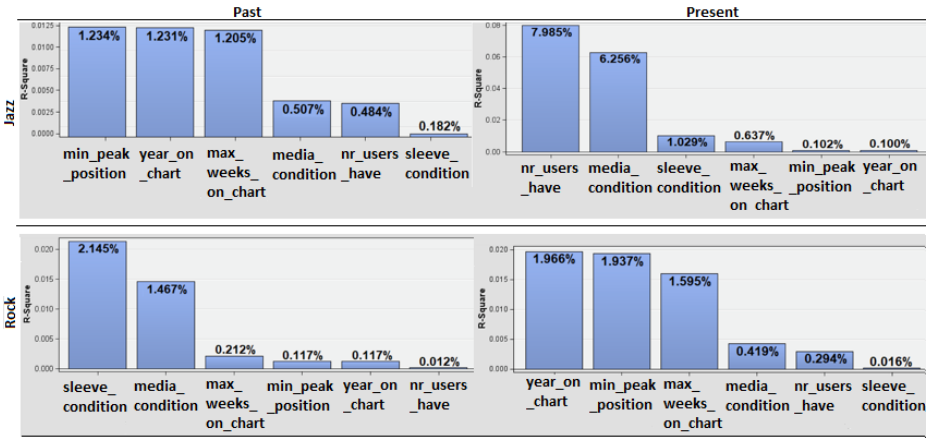

Fig. 1. Variable importance.

Given the achieved results presented above, it is possible to observe some patterns and theorize on which factors may or may not be plausible explanations for what was reported while providing useful insights in characterizing what influences a vinyl record market value. Regarding an artist's popularity and his presence in the charts, and the record label behind the success, it is possible to conclude that both these factors can definitely be relevant to price, in some cases with the examples of Deep Purple or Led Zeppelin, which show a good 'min_peak_position' in Billboard's Hot 100 as well as labels that eventually became part of one of the 'big three'. Nevertheless, they do not always dictate which artists will have their records at highest prices. This could be explained by a record release with a reduced number of produced records that becomes a collectible, making it a lot more valuable without necessarily having an immensely famous artist or a very powerful label behind it.
Focusing more deeply on the results presented on the record labels, it is interesting to observe the ongoing change that becomes obvious once one realizes which labels belong to which other labels. This change shows how the big record labels have been

losing market share to an increasing number of independent labels, implying a paradigm shift on the industry going from a more label-centered to a more artist-centered market, where new alternatives to music production and distribution are made available every day and where more and more musicians become successful enough to build their own recording companies, giving artists in general more rights in the managing part business.

Besides these more goal-oriented results, the data mining process also uncovered some patterns that are worth discussing. In the variable selection process, it is also possible to see these differences (as shown in Figure 1), pointing to the changes that have been occurring in the music industry and, consequently, at the generalized change in perception, from the artists to the audience, regarding the vinyl records. The variables 'min_peak_position', 'year_on_chart' and 'max_weeks_on_chart' appear to be more relevant in Jazz past and Rock present. Then, in their respective counterparts, Jazz present and Rock past, where 'media_condition' and 'sleeve_condition' tend to take the place of the first three. Such is worth considering, as the most obvious thing in common between Jazz past and Rock present is that, in the past, Jazz was one of the most popular genres, just like Rock is nowadays, which suggests that the features extracted from Billboard to measure popularity become more relevant in the 'eras' where the record's genre is more popular. It is also interesting to see what happens when the genres are not as appealing to the masses, with 'sleeve_condition' and 'media_condition' becoming more relevant, indicating that the records might become more perceived as collectibles.

4 Conclusions

This research, like many others, results from an endeavor to better understand some part of the present reality, gathering information about the subject and trying to extract knowledge from it. The initial objective of this study started as an effort to determine what factors influenced the price of vinyl records. The initial hypotheses consisted on two strong candidates as big contributors, the first being their popularity and the second being the label that originally released the record. For this, four datasets were built by scraping data from Discogs, extracting a sample of old and new records for both Rock and Jazz genres, and from Billboard's Hot 100 chart, where the ranking positions extracted were matched by artist with the entries from Discogs. The variables obtained from Billboard were used to assess the likely effect of popularity while the labels were extracted from Discogs along with several other features, including the target 'median_price'.

With these datasets, it was possible to generate plots to visualize the distribution of the data and existing visible correlations, showing that the record labels and the ranking positions, although present among the most expensive records, did not seem to be the majority of the cases, indicating that many other factors might also influence the price of a record.

Finally, the variable selection process, used to filter the inputs for three of the models generated for each dataset, suggests that variables like 'avg_vinyl_rating', the ranking on Disgocs for the record's release, and 'nr_users_want', the number of users

wanting the same release, seem to be the more relevant factors influencing price. Besides those obvious features, there is also another observable pattern regarding 'max_weeks_on_chart' and 'min_peak_position'. These variables seem to be more relevant in the periods of more popularity for Jazz, with the past 'era', and Rock, with the 'present' era. This means that when Jazz was one of the most popular genres, the artist's popularity would have a larger contribution to the price of his records. The same can be said about Rock and the relevance of its more recent popularity to price.

In summary, this research suggests a possible set of variables that can be relevant to a vinyl record's price and a data mining approach to analyzing the datasets, creating an overview of what the main factors to be considered are and showing how the genre's popularity impacts the importance of an artist's popularity. For example, if nowadays an artist becomes popular within the Jazz genre, which has a relatively smaller fan base when compared to Rock, then the influence of his fame on the price of his records will actually be less than the influence of the record's condition. However, for more popular genres, with larger communities of followers and wider media coverage, the opposite becomes true. As such, the work presented in this research along with the knowledge yielded from the accomplished results, opens up several paths for further investigations, as well as providing companies and interested people some tools to better understand how to evaluate and attribute a price to a record they might want to sell.

References

1. Sarpong, D., Dong, S., & Appiah, G.: 'Vinyl never say die': The re-incarnation, adoption and diffusion of retro-technologies. Technological Forecasting and Social Change, 103, 109-118 (2016).
2. Kreps, D.: Sony buys out Bertelsmann, Ending Sony BMG, https://www.rollingstone.com/music/music-news/sony-buys-out-bertelsmann-ending-sony-bmg-101255/, last accessed 2019/09/20 (2008).
3. Perpetua, M.: Universal Music Group Purchases EMI Music, https://www.rollingstone.com/music/music-news/universal-music-group-purchases-emi-music-233091/, last accessed 2019/09/22 (2011).
4. Wang, A.: Sony Doubles Its Song Catalog – and Its Control of the Music Industry, https://www.rollingstone.com/music/music-news/sony-doubles-its-song-catalog-and-its-control-of-the-music-industry-629865/, last accessed 2019/09/22 (2018).
5. Attali, J.: Noise: The political economy of music (Vol. 16). Manchester University Press (1985).
6. Bhattacharjee, S., Gopal, R. D., Lertwachara, K., Marsden, J. R., Telang, R.: The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. Management Science, 53(9), 1359-1374 (2007).
7. Rosenblatt, B.: Vinyl Is Bigger Than We Thought. Much Bigger. Retrieved from https://www.forbes.com/sites/billrosenblatt/2018/09/18/vinyl-is-bigger-than-we-thought-much-bigger/#59e38e371c9c, last accessed 2019/09/24 (2018).
8. Diggin' Into Discogs Data., https://blog.discogs.com/en/diggin-into-discogs-album-releases/, last accessed 2019/09/26 (2016).

9. Silva, A. T., Moro, S., Rita, P., Cortez, P.: Unveiling the features of successful eBay smartphone sellers. Journal of Retailing and Consumer Services, 43, 311-324 (2018).
10. Moro, S., Cortez, P., Rita, P.: A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31 (2014).
11. Moro, S., Ramos, P., Esmerado, J., Jalali, S. M. J.: Can we trace back hotel online reviews' characteristics using gamification features?. International Journal of Information Management, 44, 88-95 (2019).
12. Palm, M.: Vinyl Records after the Internet. The Dialectic of Digital Culture, 149 (2019).