A Work Project, presented as part of the requirements for the Award of a Master's degree in Management from the Nova School of Business and Economics.

The Music Industry in the Streaming Age: Predicting the Success of a Song on Spotify

MATTEO MATERA

Work project carried out under the supervision of:

Qiwei Han

21-05-2021

Abstract

The digitization of information goods has fundamentally changed the consumption patterns of music, such that the music popularity has been redefined in the streaming era. Still, the production of hit music that captures the lion's share of music consumption remains the central focus of business operations in the music industry. This paper aims at building a machine learning model capable of predicting the success of songs on Spotify. The created dataset contains 14,303 songs some appeared in Spotify's Global Top 200 chart and others never entered in the chart. The problem was approached as a classification task and the best results were obtained by the Random Forest classifier with an F1 score of 85,6% on the validation set.

Keywords:

Music Industry, Spotify, Data Science, Machine Learning

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

Table of Contents

1. Contextual background

The story of the music industry over the past 20 years is one of radical transformation, with the shift from physical to digital at its heart. In 2020, global recorded music industry revenues reached $21.6 billion (IFPI Global Music Report 2021), 8% more than 2019 and, more importantly, a figure that had not been reachead for exactly eighteen years. The trend in music industry revenues over the past two decades was characterised by an inexorable decline until 2014 and a subsequent upturn. In 2001, at the dawn of the internet, revenues were $23.6 billion, more than 97% of which came from physical record sales. The proliferation of new technologies, such as the mp3 format for digitizing music, and the increased use of the internet have transformed the way consumers enjoy music products. Initially, the decline in record sales at the expense of digital piracy implemented by services such as Napster was not offset by any new revenue streams. The lowest figure was recorded in 2014 with $14 billion. Since then, however, the trend has reversed, thanks mainly to a market that has been able to innovate and has stopped concentrating its efforts on the sale of physical products and instead emphasised the most innovative digital services. The factor that had the biggest impact on revenue growth was streaming services, which in 2017 generated more revenue than physical album sales for the first time ($6.5 billion versus $5.2 billion) and thus legitimised the definitive transformation of the music industry. In 2020, streaming services generated 68% of global recorded music industry revenues, while physical album sales only accounted for 20%.

As a leading online platform, Spotify dominates the streaming services market with 356 million monthly active users in 178 markets, giving it more than 30% market share. Users have 70 million songs to choose from and can take advantage of the services offered for free or by paying a monthly subscription.

2. Related works

The current dominance of streaming services within the music industry has led scholars to analyse the specificities and changes in the sector. According to Waldfogel (2017), digitisation has increased the production of new songs, making them available to a wider audience, and has transformed traditional distribution and promotion channels. An obvious example is the ease with which an artist can release their music on a streaming platform and make it available to consumers without the need to physically release a record.

Changes in consumer buying behaviour in the digital age were studied by Aguiar et al. (2016). The authors analysed clickstream data from 16,500 European consumers and found two main effects of digital music platforms: a stimulus effect on digital music sales and a change in consumer purchasing habits since the 2000s. These studies highlight how the increased availability of licensed songs has changed individuals' music consumption alternatives. Indeed, the possibility to download any song available on the platforms has facilitated the purchase of music products, which would only have been available in physical format in the past.

Léveillé Gauvin (2018) also points out that the change in consumption patterns is attributable to the immediate access to a larger collection of songs and the possibility of skipping them. The newly created ecosystem is highly saturated and the competition within it can be explained in terms of the attention economy, as artists compete to acquire listeners' attention. Specifically, the author highlights changes in four parameters in the musical composition from U.S. top-10 singles over the last 30 years: a decrease in the number of words in songs' titles, an increase in the average tempo, a reduction in time before the entry of the vocals and the title being mentioned.

However, Aguiar and Waldfogel (2018) point out that the availability of increasingly large music catalogues causes a product discovery problem that platforms try to address in two ways: by using recommendation systems based on individual user preferences or by creating compilations or Charts. With reference to the creation of charts in the article "*Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market",* Salganik et al. (2006) highlight the impact that charts have on the ultimate success of songs. In detail, the authors divide the participants of the study into two groups who were asked to arrange the songs in order of preference. Those belonging to the treatment group were provided with the number of times a song has been downloaded, as a proxy for popularity. The conclusion they come to is that one of the main determinants of a song's success is the social influence derived from additional information, such as the number of downloads, thus overshadowing the song's musical characteristics. In a subsequent study, Aguiar and Waldfogel (2018) manage to quantify the impact that a song's presence in the charts has on its success. Specifically, if a song is added to Today's Top Hits, a famous spotify's playlist which has 18.5 million followers, the streams will increase by almost 20 million and the value added to the song between $116,000 and $163,000. Therefore, given the prominence achieved by streaming services in the music industry, it is even more urgent for record companies and artists to understand which are the determinants for create popular songs and access playlists or charts.

With this in mind, many papers have attempted to predict or explain song success using machine learning. Herremans et al. (2014) attempted to predict whether a dance song will enter the top 10 using songs in the archives of Billboard and the Official Charts Company. The variables used were audio features provided by The Echo Nest and results obtained provided valuable insights into song prediction. The classification models used were very simple but nevertheless the accuracy obtained with a logistic regression was 80%.

A more comprehensive study was performed by Interiano et al. (2018) who collected 500,000 songs produced from 1985 to 2015 to analyze the dynamics of success, "*defined as making it into the top charts"*. Noteworthy is the use of a variable called "Superstar" that represents the past presence in the charts by an artist, this approach allowed them to increase the accuracy of their model by 10% attaining a prediction accuracy of 86%.

Finally, Araujo et al. (2019) turned their interest towards streaming platforms and specifically Spotify, trying to predict whether or not a song would enter the Top 50 Global ranking. The authors delineated the problem as a classification task and the data used contained past information and acoustic features of the songs. The best results were obtained using a Support Vector Machine with KBF kernel which achieved an AUC score higher than 80%.

3. Research Objective

The following research is within the context of the *Hit Song Science* defined by Pachet as "*an emerging field of science that aims at predicting the success of songs before they are released on the market*"(Pachet, 2008).

The primary objective is to build a classification model that can predict the success of a song in the context of the Spotify. In this specific case, success is defined as whether the song would appear on Spotify's Global Top 200 chart. In order to achieve the desired results, the following research started from the consideration of pre-existing works, extrapolating the most effective methodologies regarding analysis and creation of variables. Finally, for the first time, specific models will be created for certain music markets on Spotify using the same methodologies applied for Spotify's Global Top 200 chart. This contribution will allow to assess the effectiveness of generalising the predictive models and evaluate the differences in the results obtained in each countries.

The data was mainly collected from the Spotify platform as streaming services represent the main source of revenue for the music industry and therefore it is crucial to analyse and understand the underlying dynamics of predicting a song's success. In order to increase the variance of the data and make the classification task more similar to reality, songs released in the same timeframe but which were not successful were collected. The time interval considered is that between the years 2017 and 2021. The importance of the contemporaneity of the analysis period allows this model to be used to predict the success of a song at this point in history as it contains the most up-to-date information.

In this work, specific models were created for some music markets on Spotify using the same methodologies applied for Spotify's Global Top 200 ranking. this contribution makes it possible to assess whether the results obtained can be replicated to predict the success of songs also in specific markets with different characteristics from each other.

4. Data

No readily available datasets were found that met the aims of this research. Therefore, the data was collected from scratch and curated using the following tools: *Python*, as programming language, *Pandas*, for data manipulation, and interacting with the api (Application Programming Interface) made available by Spotify.

The following section will list the procedures performed to obtain the dataset feed into the predictive model. A visual explanation of the procedure is provided in *Appendix 1.*

4.1 Global Top 200 songs

The first dataset created contains information about the songs on Spotify's Global Top 200 chart found at spotifycharts.com. The data was scraped using a web-crawler programmed in Python. The daily charts obtained using this method range from 1st january 2017, the first available date on the website, to 1st February 2021 and contain information related to: the chart position, the number of times the song has been streamed, the url, the title and the author of the song. Afterwards, two variables have been added to complete the dataset. The first one, called *id*, has been obtained from the previously mentioned url variable. It represents the 22 digit alphanumeric code that identifies each song and it is used to interact with spotify's api. The second one, named *date*, represents the date of the daily charts,

After saving the charts in memory in .csv format, the files were concatenated to create *charts_df* dataset, consisting of 297,800 rows and 7 columns corresponding to 1489 daily charts. In this research, which was not aimed at analyzing the trend of the songs in the charts, only the information regarding the first appearance of each song was taken into account. In the end, the dataset includes data about 6,266 songs. *Table 1* shows the variables name, description and type.

Table 1: *charts_df* variables explanation

Variable	**Description**	**Type**
position	chart position of the song	Int64
track_name	title of the song	Object
artist	name of the artist	Object
streams	number of times the song has been streamed	Int64
url	url of the song	Object
Date	date of first appearance	Datetime
Id	song identification code	Object

The second dataset created contains the audio features of each song. In order to collect this information it was necessary to extensively use the api provided by spotify, which allows access to the archive of information, metadata and musical content held by the company. In order to use this technology, an app named *success_songs* was created on the spotify developers' website, which allowed to obtain the access credentials *Client ID* and *Client Secret*. Then, for each song's id in the charts_df dataset, a GET request has been made to the api’s endpoint for audio features. The results obtained in json format were saved in *audio_features_df* dataset composed of 6.266 rows and 14 columns. The variables obtained are of two types: the first ones extracted directly from the tracks containing intrinsic and objective characteristics (e.g. duration_ms, key, etc.) and the second ones calculated by an algorithm owned by Spotify (e.g liveliness, acousticness, etc.). *Table 2* provides an explanation of the variables.

Table 2: *audio_features_df* variables explanation

Variable	Description	Scale
Acousticness	Measures the likelihood a track is acoustic.	0 - 1
Danceability	how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.	0-1
Duration_ms	length of the track in milliseconds.	30133 - 943529
Energy	Perceptual measure of intensity and activity.	0-1
Instrumentalness	likelihood that a track has no vocals.	0-1
Key	Estimated overall key of the track using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D	0-11
Id	song identification code	N/a
Liveness	Presence of an audience in the recording	0-1
Loudness	average loudness of a track in decibels (dB).	-60 – 0 (dB)
Mode	whether a song is Major (1) or minor (0).	Binary {0,1}
Speechiness	Measure of presence of spoken words in a song.	0-1
Tempo	Estimated tempo of a track in beats per minute (BPM).	45.78 - 216.334
Time_signature	Estimated overall time signature	1-5
valence	Measure of musical positiveness.	0-1

Lastly, *additional_info_df* contains information about the artist and album of each song. Requests made towards the tracks endpoint of the api, provided information regarding: release date, number of artists, explicit, and the album id of the song. The dataset was then completed with some variables that would provide more album details as described in *Table 3*. A further step of the work was to merge the previously obtained datasets thus creating the *onchart_df.*

Table 3: additional_info_df variables explanation

Variable	Description	Type
Id	Song's identification code	Object

Album_type	Whether a song is a single or is part of an album of compilation	Object
Release_date	Song’s release date	Object
Explicit	whether the content of the song is explicit	Bool
Total_tracks	Number of songs in the album	Int64
N_artists	number of artists taking part in the song	Int64
Label	record label that produced the song	Object
Artists_id	Identification code for each artist	Object
Album_id	Album’s identification code	Object

4.2 Out of chart songs

Information about songs that are not on the chart allowed for an increase in the variation of musical features so that a model with greater predictive power could be obtained. The collection of the following dataset was more difficult than the previous one since Spotify's api does not provide any endpoint that allows searching for published songs by specifying the desired time interval.

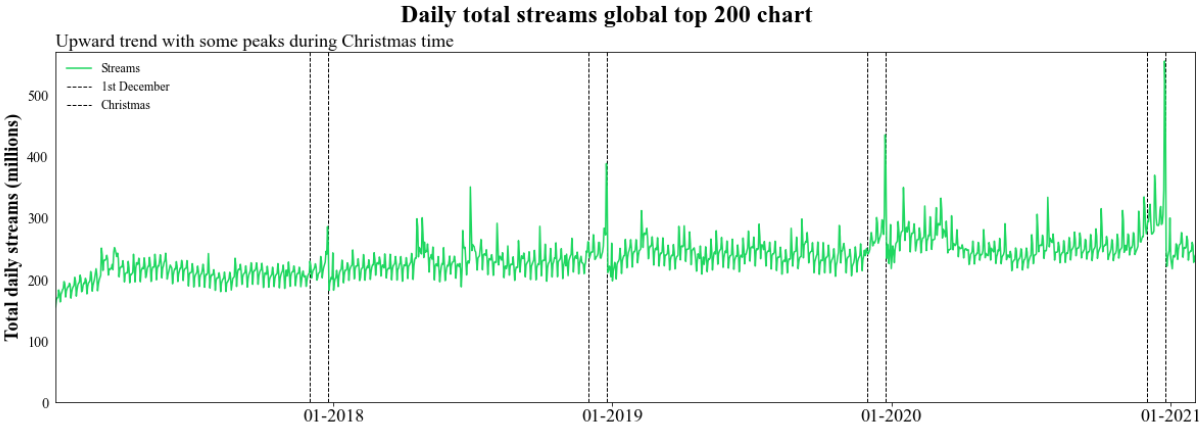
The solution adopted to overcome the limitations encountered was to collect information about albums released from 2017 to 2021 from two websites: metacritic and wikipedia. The procedures to collect and curate the data were identical for both sources used. The first step was scrape information related to album title and artist name using a web-crawler programmed in Python. Then, through the search endpoint of the spotify api it was possible to obtain the identification code related to the collected albums to allow access to the tracks contained in each of them. Often the songs contained in an album are very similar to each other either because they are sung by the same artist or because they are the expression of a defined musical project. In order not to add too much similar information, it was decided to randomly select only one for each album. The collected data were checked so that there were no duplicates and none of the tracks were also present in *onchart_df*. Finally, the procedures adopted previously to obtain audio features and additional information were applied thus creating *outchart_df*, consisting of 10,130 songs.

4.3 Data analysis and Feature engineering

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work (Shekar 2018). Some of the new variables created were obtained after performing an analysis of some features and others were integrated from pre-existing research work.

Analyzing the Streams variable in *onchart_df* shows a growth in daily listening during the period considered. In 2017 the daily average was 1 million while in 2020 it was 1.3 million marking a 25% increase. *Figure 1* shows the sum of the daily streams of the Global Top 200 chart. The main takeaway appears to be a marked seasonality in the streaming with notable peaks near the Christmas period marked by the listening of famous Christmas hits. In fact, the song that had the most listens in one day was "All I Want for Christmas Is You" by Mariah Carey with 17.2 million recorded on 24th December 2024 and in the top 10 of the number of daily streams there are 8 Christmas themed songs. These songs made it onto the chart because of the specific theme covered and the time of year during which they are streamed.

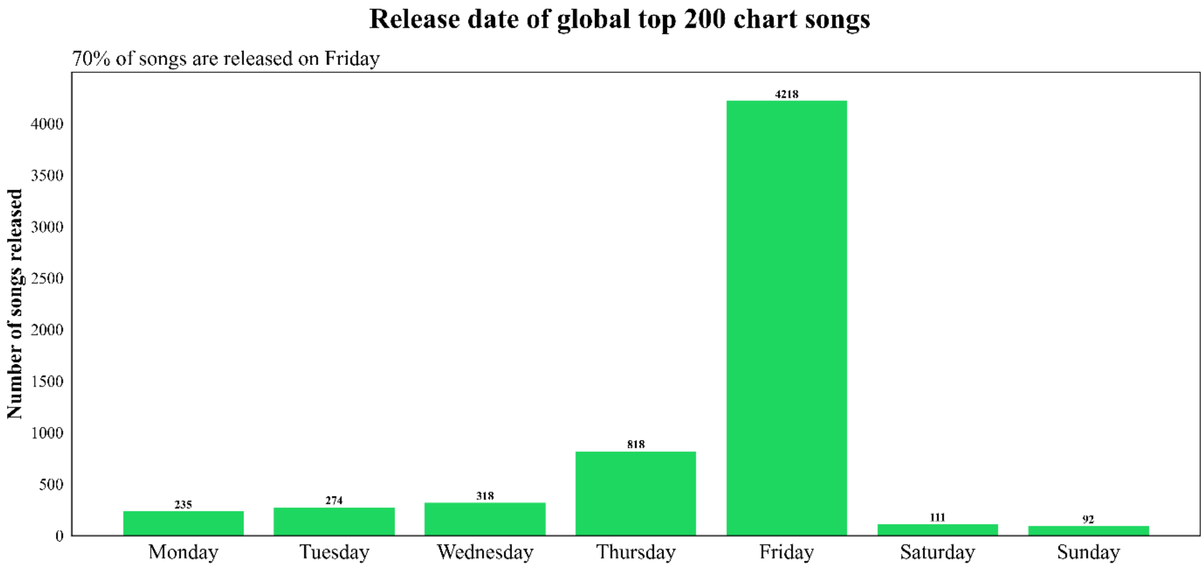
Figure 1: Number of daily streams for the Global Top 200 chart.



Another characteristic that emerged is that some tracks were released much earlier than the time frame considered. This could have influenced the model since their presence is not solely due to musical characteristics but to the occurrence of determining circumstances such as the Christmas season or the death of a singer. Following this reasoning 524 songs were eliminated because they were released before 2016.

Analysis of the data showed that songs are not released evenly throughout the week and that 70% occur on Fridays as shown in *Figure 2*. To use this information, the variable *day* was created.

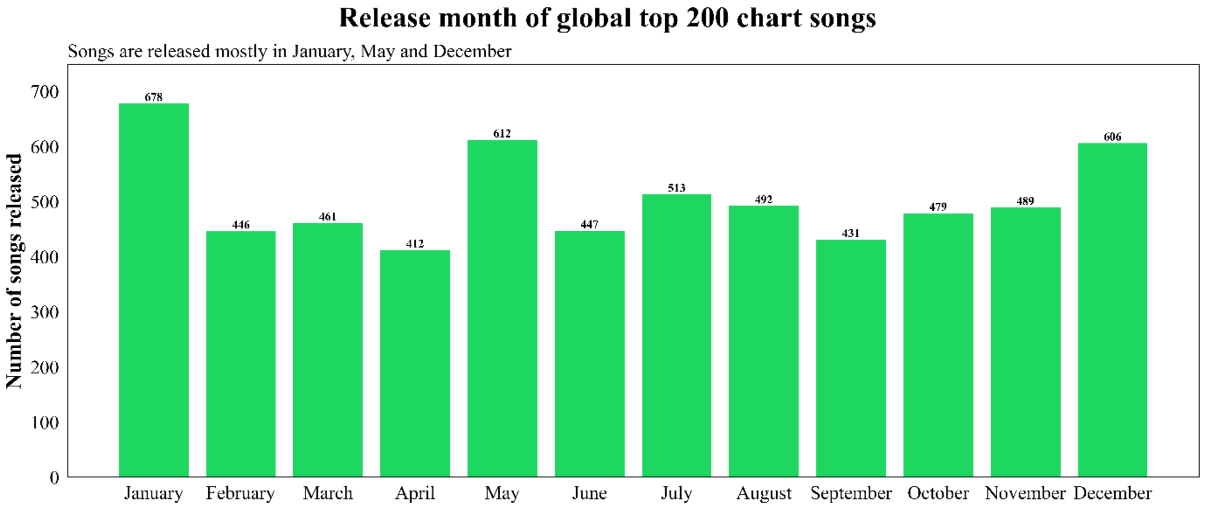
Figure 2: Daily release distribution for songs



In terms of release months, the distribution is less skewed as shown in *Figure 3*. There are two peaks at the beginning and end of the year. another one in May, which may be caused by the release of hit summer songs.

The two datasets created in section 2 were concatenated so as to compose the dataframe *model_df*. A next step was to identify the dependent variable called *onchart* that the model should try to predict. It can assume two values according to the dataset of origin, if a song was present in the Global Top 200 chart it assumes a value of True, on the contrary if the song was not successful it has a value of False.

Figure 3: Monthly release distribution for songs



For another variable, was considered the fact that in a highly competitive market such as the music market, only a small number of artists manage to reach a remarkable success as consumers tend to listen to songs of already famous artists to minimize the cost of research. In fact, there are only 1,026 artists in *onchart_df* over a 3-year period. In this regard, Interiano et al. (2018) introduced the concept of "Superstar" that allows mapping the celebrity of an artist thus improving the predictive power of the model used. Following the reasoning set forth by the authors, the variable *main_artist_famous* was created which takes on a value of True if the artist was already on the charts prior to the release of the song.

An additional aspect was the observation that over the past decade, musical collaborations between two artists have increased across all musical genres (Seekhao,2020). The author observes that the growth of this phenomenon is closely related to that of streaming service revenues, which are in addition to professional motivations (new compositional techniques, sharing ideas, acquiring skills). Moreover, collaborations between artists allow to expand the position of each artist and get in touch with new fans. For this reason two variables have been created: *has_featuring* that shows if the song is the result of a musical collaboration and *featuring_artist_famous* that is the transposition of the superstar effect towards other artists.

In order to try to add information from the artists and quantify numerically the size of their celebrity, the variable *previous_songs* has been created because it is assumed that the more songs produced by an artist, the greater will be the fanbase and the probability of entering the charts. Despite the changes that have occurred to the music industry during the digital age, Kaimann et al. (2020) highlight that songs produced by the three major record labels (Sony, Universal, and Warner) are more likely to be successful than songs produced by independent labels. *major_label* variable represents whether or not a song belongs to one of them.

4.4 Final Preprocessing

Before feeding the final dataset into the predictive model it is necessary to quality check and curate the data. The variables *key*, *time_signature* and *mode*, even if they are numerical, actually indicate categories. For this reason the values are transformed into string-type. The temporal dependency between tracks in *model_df* is crucial to maintain as it does not make sense to use future songs to predict songs in the past. For this reason, the data was sorted in ascending order using the *date* variable.

Next, a procedural problem was encountered for some values of the *main_artist_famous* variable. Specifically, the first 200 songs entered, which correspond to January 1, 2017, will definitely have incorrect values since they cannot be compared to any previous songs in the absence of data. Therefore, the information that would be passed into the model would be incorrect and could decrease performance. To minimize this error, it was decided not to consider songs that entered the charts throughout January 2017. Therefore, the dataset was filtered and the new time range covers a period of exactly 4 years from 1st February 2017 to 1st February 2021.

At a later stage, variables that were previously used to interact with the api or those that contain values not needed by the model were removed. *Model_df* consists of 5,308 songs that were successful and 8,995 songs that never made the chart. This distribution creates an unbalanced class problem. However, no undersampling procedure will be implemented in this research since in reality, the songs that will not be successful are far more than the others. Therefore, all tracks have been kept so as to keep the dataset as similar as possible to reality.

In the last step a Pipeline is used to encode the categorical variables using sklearn's *OneHotEncoder*. The numerical variables instead have not been transformed because the models used are tree classifiers that are not affected by different scales of values.

5. Machine Learning

5.1 Evaluation metrics

The metrics used to evaluate the performance of the classification model in this research are shown in *Table 4*. The formulas are based on the confusion matrix which is a summary of the results of the model predictions (*Appendix 2*). As pointed out previously the label onchart is imbalanced, so Accuracy was not used because it would have produced misleading results.

Table 4: Evaluations metrics

Evaluation Metric	**Formula**
Precision	$\frac{TP}{TP+FP}$ (1)
Recall	$\frac{TP}{TP+FN}$ (2)
F1 score	$2 \times \frac{Precision \times Recall}{Precision+Recall}$ (3)

Precision (1) measures the proportion of true positive instances out of all the predictions predicted as positive. In this research, it denotes how many of the songs predicted as positive are actually positive. From a business perspective, achieving high accuracy is critical for record labels in order to reduce the costs associated with false positive instances. Investing in artists and songs that will not be successful could result in financial losses.

Recall (2) measures the proportion of true positive instances out of all positive instances. It denotes how many successful songs have been predicted successfully. In this case, a high recall allows record companies not to miss the opportunity to invest in songs which might turn out to be new hits.

F1 score (3) is the harmonic mean between recall and precision and shows a tradeoff that can minimize the differences between the two values. This measure is often used to select the best classifier in a group when the costs associated with precision and recall are similar. The model that will offer the best F1 score will allow simultaneously to obtain safer investments and not to lose investments that will be successful.

5.2 Model selection

A common practice used to select a model and obtain less biased results is called k-fold Cross Validation. In short, this procedure randomly splits the train set into k folds and iteratively trains the model on k-1 folds and validates the results on the remaining one until all the data are used. The final score is the average of those obtained on each fold. Since the samples of each fold are randomly selected,, this method is not suitable for time series because it does not take into account the temporal dependency and could cause data leakage. Therefore, a forward chaining technique called Nested Cross Validation will be used to maintain the chronological order of the data. The folds created using sklearn's *TimeSeriesSplit* can be seen in *Appendix 3.*

The models evaluated in this research are as follows: *LogisticRegression*, *DecisionTreeClassifier*, *RandomForestClassifier*, *LGBMClassifier*, *AdaBoostClassifier*, *GradientBoostingClassifier*, *XGBClassifier.*

The results obtained are shown in *Appendix* 4. With the exception of *LogisticRegression* and *DecisionTreeClassifier*, all classifiers performed satisfactorily, above 0.8 in all metrics. Recall tends to be higher than Precision in all estimators.

In order to choose the models to be subsequently optimised, it was necessary to analyse, from a business perspective, the underlying implications of the 3 evaluation metrics as highlighted in Section 5.1. Finally, F1 score was chosen in order to obtain a model with more balanced results, trying to obtain high Precision and Recall values at the same time. In fact it is beyond the scope of this research to analyse the costs associated with incorrectly predicted successful songs and those associated with not predicting future hits.

Therefore the two models which will be optimised are *RandomForestClassifier* and *XGBClassifier*. Sklearn's *RandomizedSearchCV* was used for Hyperparameters tuning procedures since this technique is very effective in finding the set of hyperparameters that can guarantee optimum performance. Specifically, for each estimator a fixed number (*n_iter* = 25) of parameters are randomly sampled from the distributions specified in *Appendix 5*.The results show that the best model turns out to be RandomForestClassifier with an F1 score of 86.2%.

6. Results

6.1 Results for Global Top 200 chart

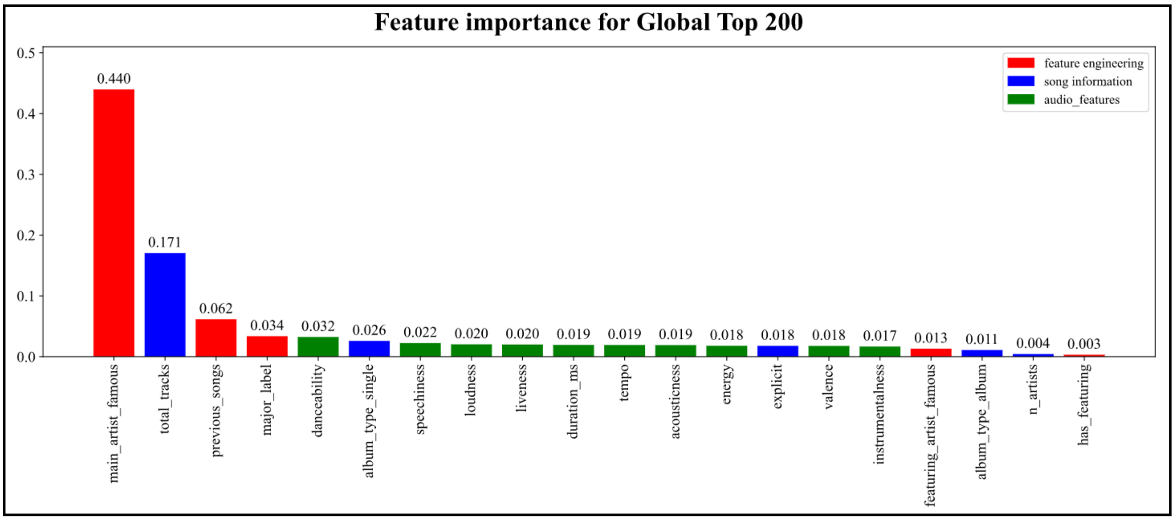
The results for predicting the songs contained in the test set are satisfactory and show that the optimized model can accurately distinguish the characteristics that determine the success of a song. RandomForestClassifier obtained an F1 score of 85.6% which represents a 1% improvement over the baseline model.

Figure 4 allows us to better understand the solved classification task and provides indications on the relevance of the variables used, guaranteeing greater interpretability. For simplicity, the 20 most important variables are shown, given that the values obtained from the remaining variables are negligible. It can be seen that *main_artist_famous* is clearly the feature that obtained the highest value and highlights how the fame of an artist is fundamental to identify success.

The second variable in importance is *total_tracks*, it indicates the number of tracks contained in an album. This result was not anticipated and can be traced to the possible cannibalization that occurs between tracks released simultaneously by an artist (Kaimann, et al. 2020). Later, *previous_song* and *major_label* turn out to be relevant, these variables were created in section 4.3 and testify to the quality of the procedures performed. Finally, audio features do not turn out to be particularly important with *danceability* getting the best results.

The performance of the model is also visible in the confusion matrix in *Appendix 7*. Specifically, 2,572 songs are correctly predicted, while 223 are false positives and 66 are false negatives. In order to shed light on the errors made by the classifier, the differences between the hit songs that were predicted correctly and those that were misclassified will be analyzed.

Figure 4: Model's feature importance



First of all, a marked difference in the variables *main_artist_famous* and *major_label* is noticed. True positive songs have a famous artist in 91% of cases and 46% of them are produced by a major record label while false negatives have values respectively of 37% (*main_artist_famous*) and 20% (*major_label*).

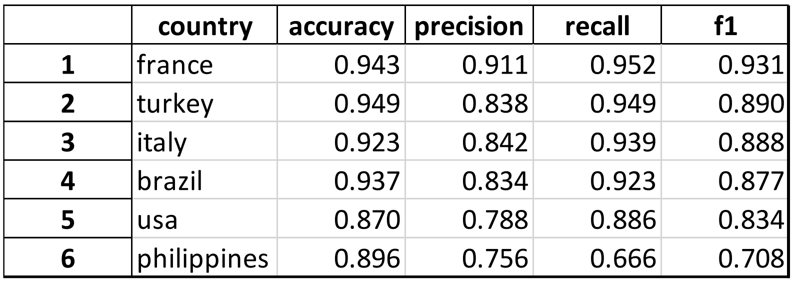
The distributions of numerical variables are provided in A*ppendix 8*. In general, the model struggles to correctly predict songs produced by emerging artists and smaller record companies and that are less likely to dance.

6.2 Results for countries

In this section we report the results for the predictions of hit songs for the top 200 charts of 6 countries: Italy, France, United States of America, Philippines, Turkey and Brazil. The countries were chosen because for their geographical location and it is assumed that the dynamics of success may be different to the global top 200 charts. The 6 datasets were created through the steps listed above and for the dataset providing off-charts songs, offcharts_df was used. For simplicity, the classifier used was the *RandomForestClassifier* given the results previously obtained, and hyperparameter tuning was performed for each of the datasets.

The values contained in *Table 5* provide evidence that the methodology devised for this research is generalizable to other markets. Specifically, all of the models' f1 scores are satisfactory and quite high, all above 0.8 with the exception of the Philippines. The best results were recorded for French songs with a Precision of 0.911 and Recall of 0.952, which together produce an F1 score of 0.931.

Table 5: Performance summary for each country



	country	accuracy	precision	recall	f1
1	france	0.943	0.911	0.952	0.931
2	turkey	0.949	0.838	0.949	0.890
3	italy	0.923	0.842	0.939	0.888
4	brazil	0.937	0.834	0.923	0.877
5	usa	0.870	0.788	0.886	0.834
6	philippines	0.896	0.756	0.666	0.708

Again, the variable *main_artist_famous* is the most important in all models (*Appendix 9*). In general, the variable *total_tracks* has the second highest score except in France where it is *explicit*. Among the audio features, the most important are *danceability* and *instrumentalness*.

As for the remaining variables, there are no similarities in the distribution and their order is specific to each dataset.

7. Conclusion

Motivated by the changes caused by digitisation and the subsequent dominance of streaming services for the music industry, the aim of this research is to build a model capable of predicting the success of a song on Spotify. A song is defined successful if it appeared in the chart. The data used was collected on the one hand from Spotify's global top 200 for a period from 2017 to 2021 and on the other hand from lists of albums produced in those years found on Metacritic and Wikipedia. The collected information was expanded using Spotify's API to add audio features of the songs and other characteristics related to the albums and artists. New variables were then created by analysing the dataset and integrating existing studies. Then, after testing multiple models, the one that proved to be best was the RandomForestClassifier with a Precision of 79.4%, a Recall of 92.8% and consequently an F1 score on the test set of 85.6%. The results are satisfactory and show that predicting the success of a song is possible.

The analysis of the importance of the variables showed that audio features are less important than predicted and therefore the musical characteristics of the songs are not sufficient to guarantee success as assumed by Salganik et al. (2006). In fact, factors such as the level of celebrity of an artist, the number of tracks on an album and production by a major record label are much more decisive.

Finally, datasets with songs collected from the top 200 charts of 6 different countries were created applying the methodology used in this research. The results obtained are satisfactory and show that the dynamics of success are similar in different markets. In fact, even in these cases the most important determinant was found to be the level of celebrity of an artist.

7.1 Limitations

While doing this Work Project, some limitations arose. First, the off-chart songs were obtained from album titles contained on two websites and not directly from the spotify platform. In addition, the number of these songs is limited and therefore does not allow for the construction of a dataset similar to the reality of the music industry where the songs that are successful are very minor compared to all those produced. Second, the use of the spotify api was a major constraint on the range of data that could be collected. For example, there is a lack of information regarding the genre and language of the artists. Also, the audio features provided are not comprehensive.

8. Bibliography

Shekar, Amit. 2018. What Is Feature Engineering for Machine Learning? Accessed April 2021.https://medium.com/mindorks/what-is-feature-engineering-for-machine-learning-d8ba3158d97a

IFPI Global Music Report 2021. Accessed April 2021. https://gmr2021.ifpi.org/report

Seekhao, Nuttiiya. 2020. How Music Collaborations Evolved in the Digital Era: A Decade in Review. Accessed April 2021. https://blog.chartmetric.com/the-evolving-role-of-music-artist-collaborations

Kaimann, Daniel & Tanneberg, Ilka & Cox, Joe. 2020. “I will survive”: Online streaming and the chart survival of music tracks. *Managerial and Decision Economics*. 42. 10.1002/mde.3226.

Aguiar, L. & Joel Waldfogel. 2018. Platforms, Promotion, and Product Discovery: Evidence from Spotify Playlists; *JRC Digital Economy Working Paper* 2018-04.

Aguiar, L. & Martens, Bertin. 2016. Digital music consumption on the Internet: Evidence from clickstream data. *Information Economics and Policy*, 34. 27-43. 10.1016/j.infoecopol.2016.01.003.

F. Pachet and P. Roy. 2008. Hit song science is not yet a science. In *Proc. of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pages 355

Herremans, Dorien & Martens, David & Sörensen, Kenneth. 2014. Dance Hit Song Prediction, *Journal of Musical Research*, 43(3):291-302.

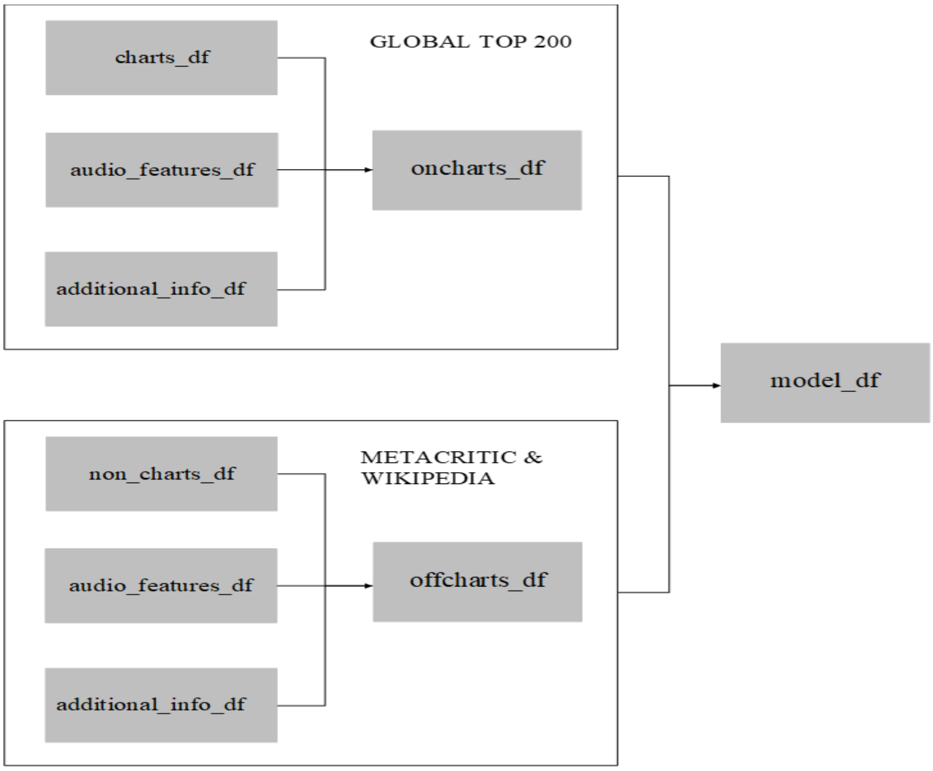
Myra Interiano, Kamyar Kazemi, Lijia Wang, Jienian Yang, Zhaoxia Yu and Natalia L. Komarova, 2018. Musical trends and predictability of success in contemporary songs in and out of the top charts, *Royal Society Open Science*, 5(5):171274.

Salganik, M & Dodds, Peter & Watts, Duncan. 2006. Experimental Study of Inequality and Unpredicatbility in an Artificial Cutlural Market. *Science*. 311. 854-856.

Waldfogel, J. 2017. How Digitization Has Created a Golden Age of Music, Movies, Books, and Television," *Journal of Economic Perspectives*, 31, 195-214.

Appendix

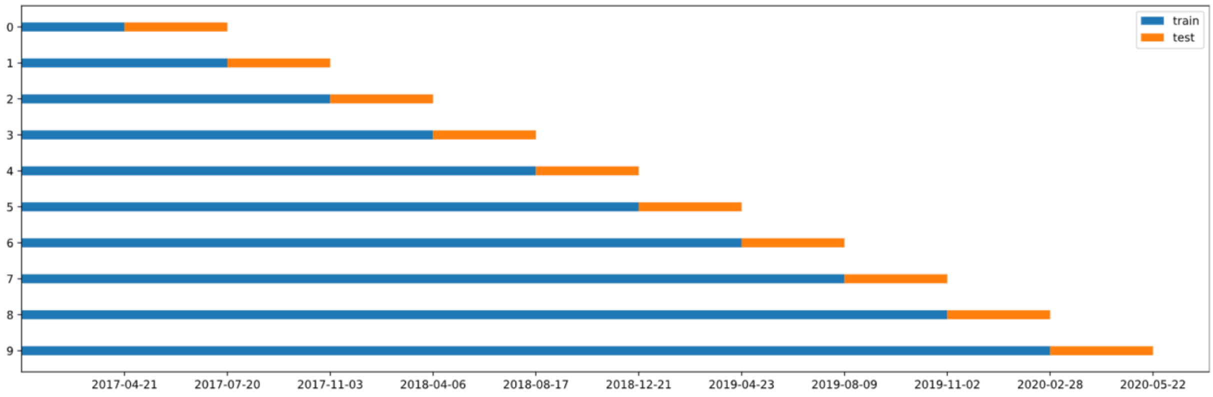
Appendix 1: Schema of dataset creation



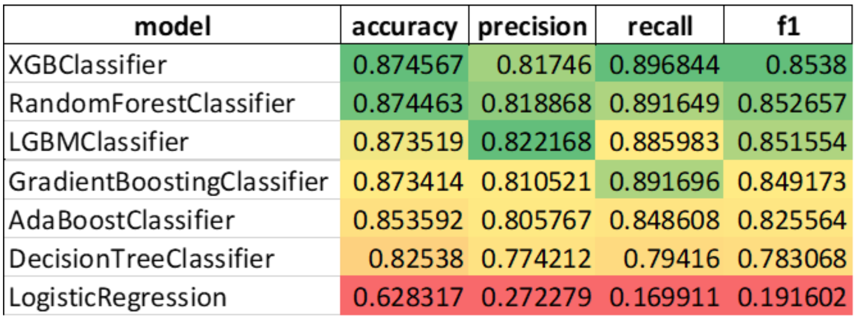
Appendix 2: Confusion matrix

	Predicted Positive	Predictive Negative
Actual Positive	True positive (TP)	False Negative (FN)
Actual Negative	False positive (FP)	True Negative (TN)

Appendix 3: TimeSeriesSplit folds



Appendix 5: Nested cross validation results



model	accuracy	precision	recall	f1
XGBClassifier	0.874567	0.81746	0.896844	0.8538
RandomForestClassifier	0.874463	0.818868	0.891649	0.852657
LGBMClassifier	0.873519	0.822168	0.885983	0.851554
GradientBoostingClassifier	0.873414	0.810521	0.891696	0.849173
AdaBoostClassifier	0.853592	0.805767	0.848608	0.825564
DecisionTreeClassifier	0.82538	0.774212	0.79416	0.783068
LogisticRegression	0.628317	0.272279	0.169911	0.191602

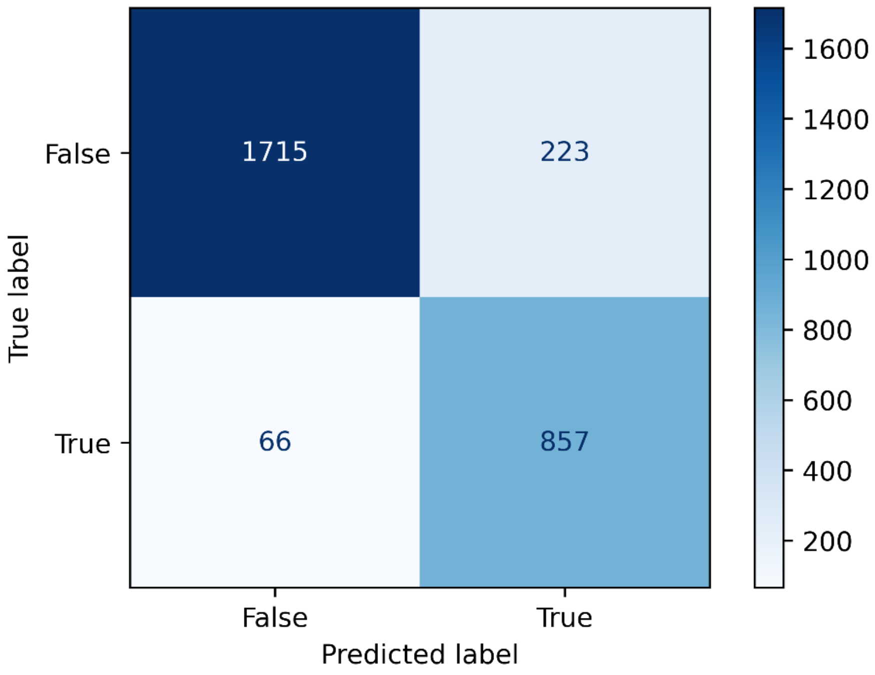
Appendix 6: Hyperparameters tuning

Model	Parameters dictionary	Best parameters	Best score

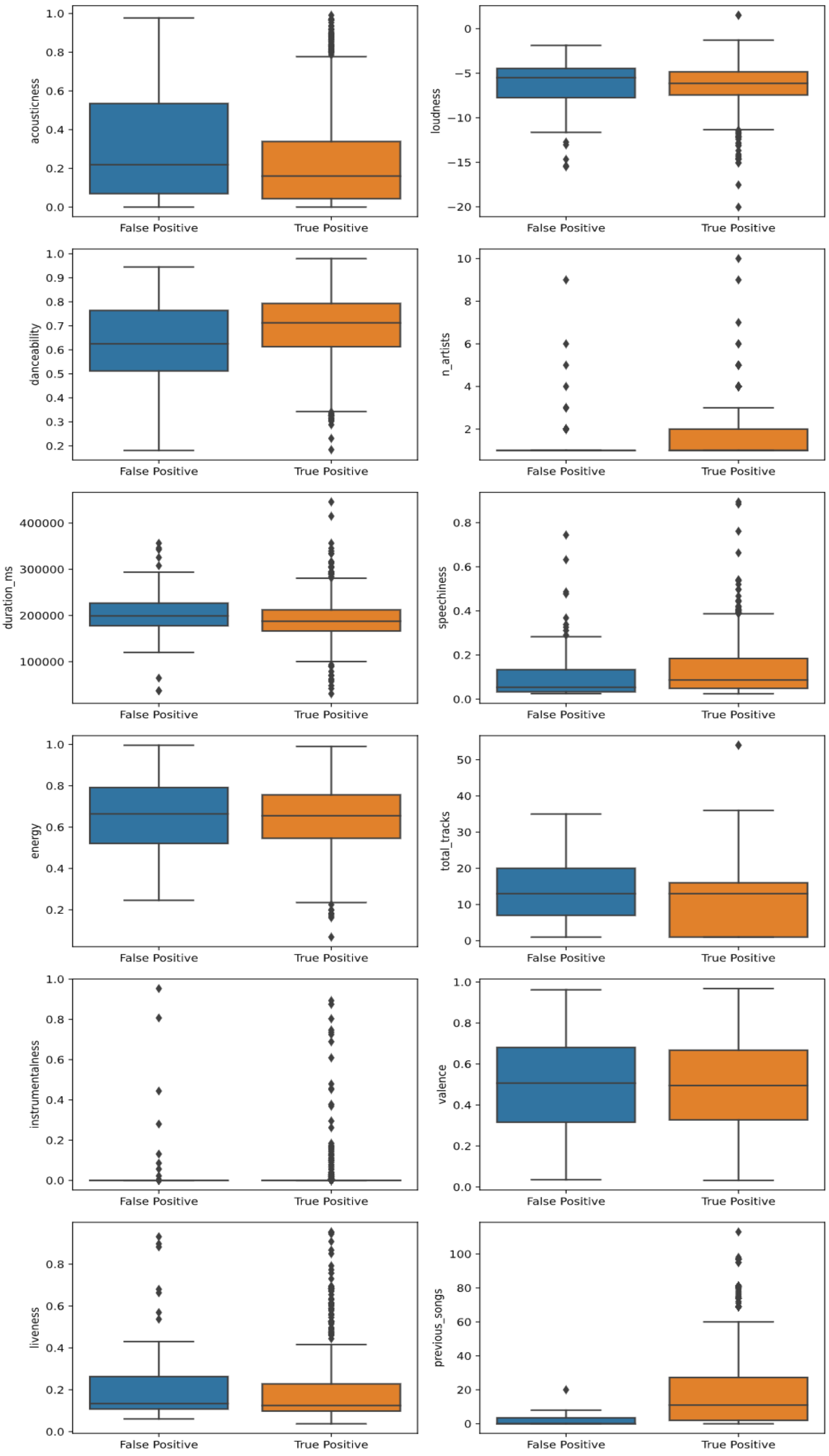
Random ForestClassifier	n_estimators = np.arange(1000,*step*=50)
	max_depth = np.arange(10,100,*step*=5)
	min_samples_split = np.arange(1,15)
	min_samples_leaf = np.arange(1,10)
	max_features= np.arange(1,x_train.shape[1]))
		n_estimators=400
		min_samples_split= 9
		min_samples_leaf=3
		max_features=20
		max_depth= 30
			0.862

XGBClassifier	min_child_weight= np.arange(1,10)
	max_depth = np.arange(10,100,*step*=5)
	gamma= [i/10.0 *for* i *in* range(0,10)]
	subsample= [i/10.0 *for* i *in* range(6,10)]
	colsample_bytree= [i/10.0 *for* i *in* range(6, 10)]
	reg_alpha=[1e-5, 1e-2, 0.1, 1, 100]
		subsample= 0.7
		reg_alpha=1
		min_child_weight=8
		max_depth= 75
		gamma= 0.2
		colsample_bytree= 0.8
			0.858

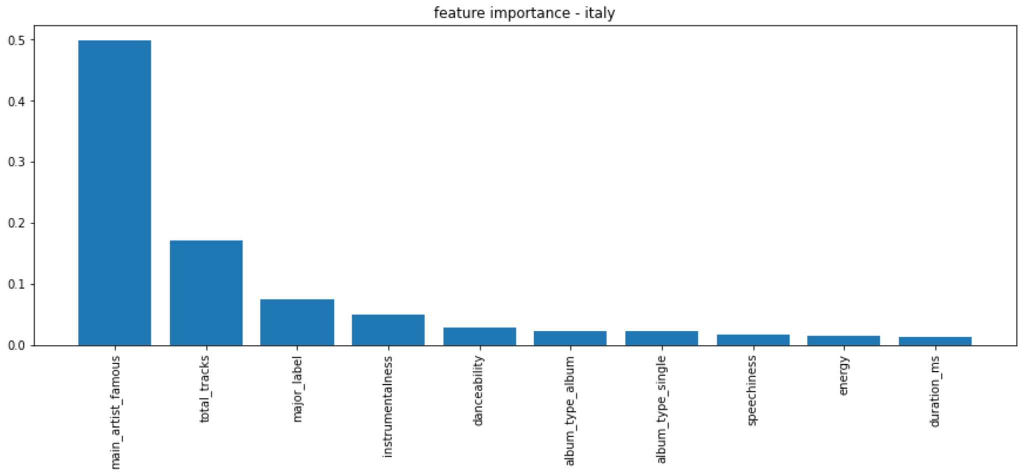
Appendix 7: Confusion matrix for global top 200 chart

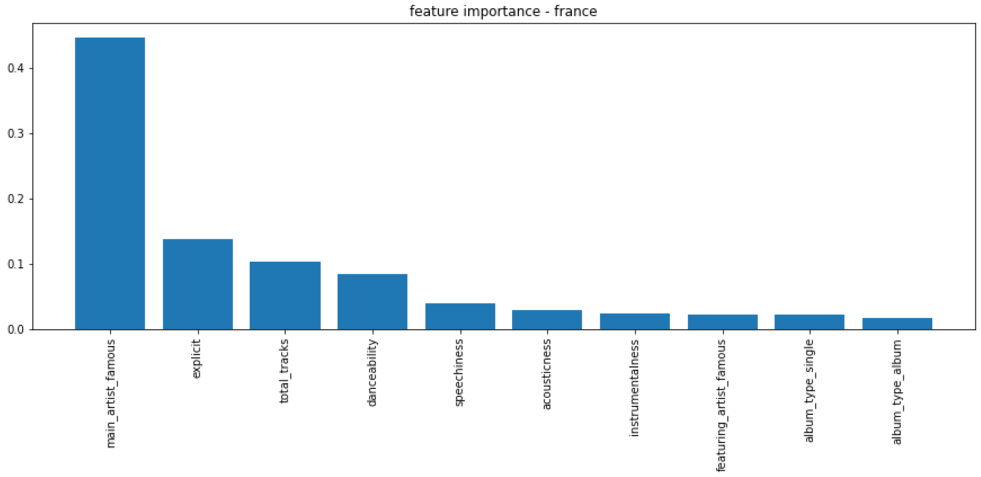


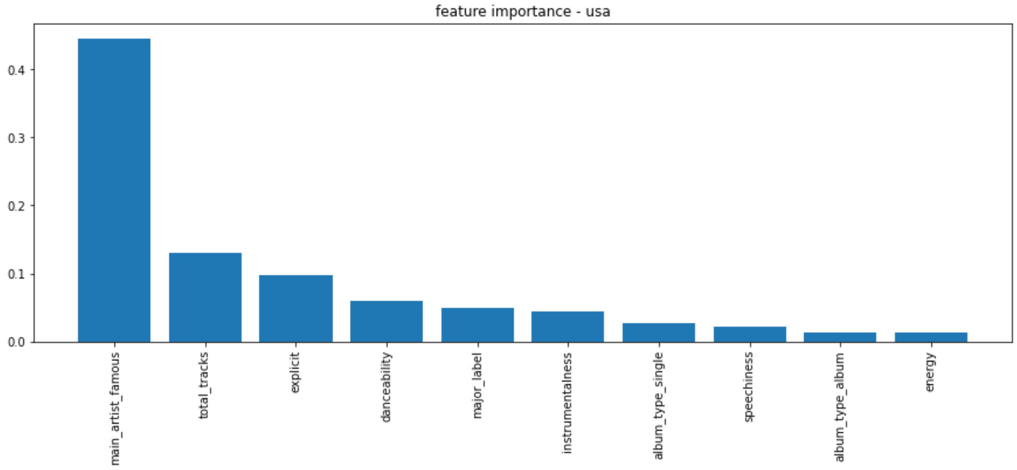
Appendix 8: Variables distribution of false positive and true positive.

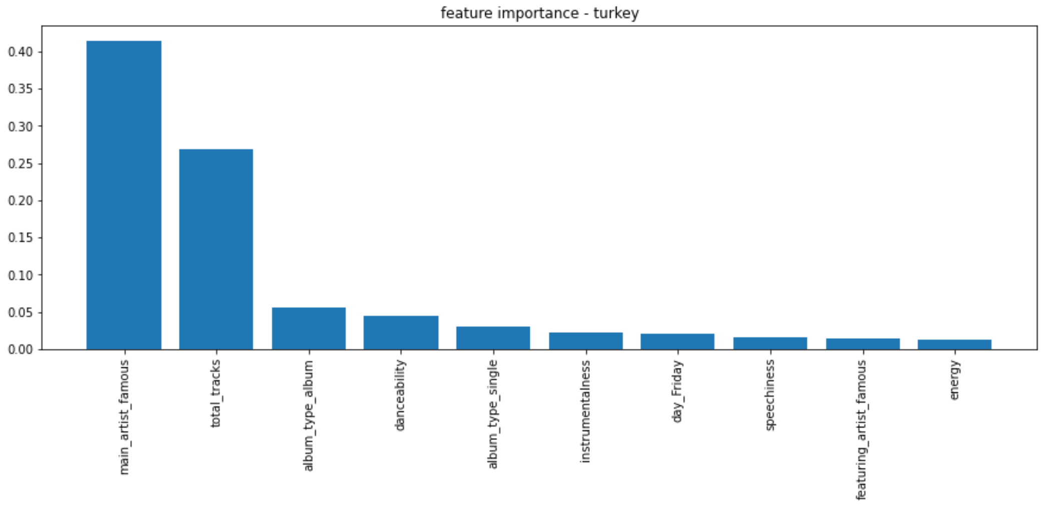


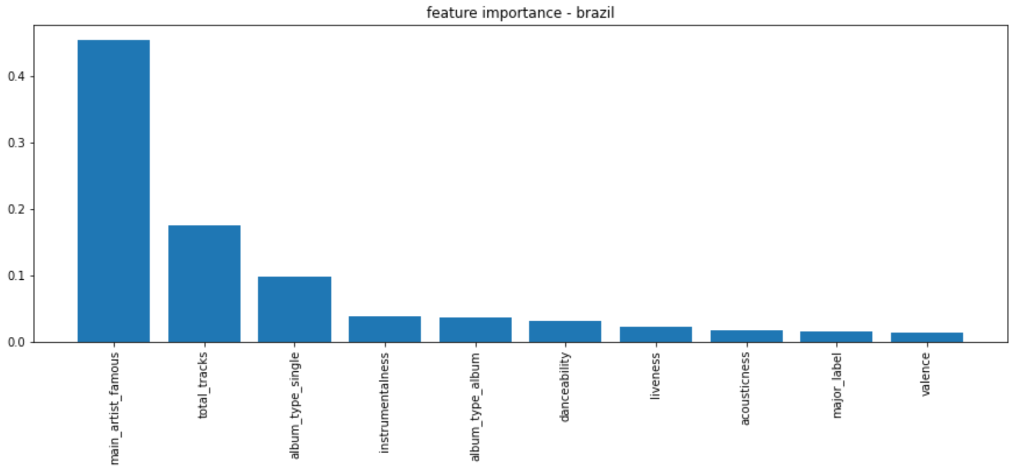
Appendix 9: Feature importance for each country

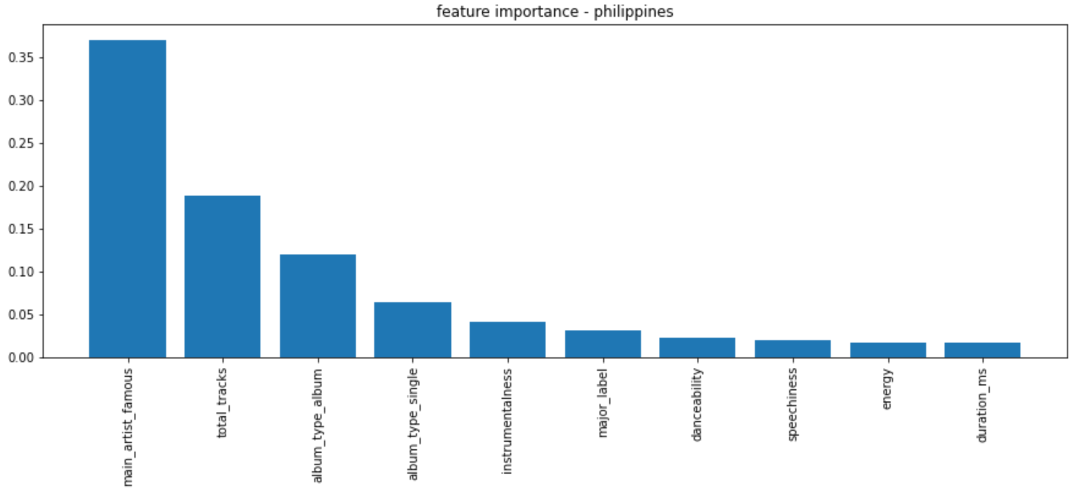

feature importance - italy


feature importance - france


feature importance - usa


feature importance - turkey
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
main_artist_famous
total_tracks
album_type_album
danceability
album_type_single
instrumentalness
day_Friday
speechiness
featuring_artist_famous
energy


feature importance - brazil
0.4
0.3
0.2
0.1
0.0
main_artist_famous
total_tracks
album_type_single
instrumentalness
album_type_album
danceability
liveness
acousticness
major_label
valence


feature importance - philippines
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
main_artist_famous
total_tracks
album_type_album
album_type_single
instrumentalness
major_label
danceability
speechiness
energy
duration_ms