Business Analysis on Spotify

Katranitsiotis Panagiotis[1]

[1]pkatranitsiotis@gmail.com

Abstract. Nowadays more and more people tend to utilize technology for their music preferences. The most popular online music platform is Spotify. This paper investigates Spotify from the perspective of music producers. Due to the immense impact of Spotify in people's daily routine, it is highly important for music producers to predict whether a potential song will gain popularity or not. Moreover, a decisive role to the releasing of a song is also the existence of explicit content. This paper will provide a Management Information System to deal with these questions. More specifically, a menu will be designed in order for producers, who are not familiar with data processing and machine learning, to make these predictions. An extensive statistical analysis will be presented to investigate any relations among the song's features and several machine learning algorithms will be provided to predict the popularity and the explicit content of a song. Finally, an efficient programming will be proceeded in order to provide the least complex models with no unnecessary steps that will increase computer's time and memory for data processing.

Keywords: Business Analytics; SAS; MIS; UML; Spotify; Machine Learning; Artificial Intelligence; Efficient Programming;

Contents

1 Introduction

It is widely known that most people tend to use online platforms for their music preferences. Nowadays, Spotify, has become the most common streaming platform for songs, music and podcasts. It was founded in 2006 by Daniel Ek and Martin Lorenzon with the purpose of familiarizing people with listening to any kind and type of music whenever and wherever they prefer [1].

Prior to music platforms, the most common and the only existent ways for people to listen to music were either through available CDs or radio stations. Spotify, contributed to the breakthrough in music industry as it is a well-structured and innovative way to provide music to people. Consequently, the user has now the possibility of listening to any kind of music on throughout his daily routine.

Due to the evolution of technology and the internet's speed, Spotify users are increasing rapidly as most of the music industries provide their songs on this online streaming platform. More specifically, in 2021, Spotify has exceeded 70 million tracks and 356 million users [2]. Therefore, the necessity of the continuous provision of songs posted directly to Spotify, is occurred. The immense number of users and tracks leads to the huge amount of data gathered daily and need to be processed and analyzed for each industry's purposes.

The purpose of this paper, is to provide a Business Analytical Plan for music industries and producers to predict a profitable song. More specifically, it lies on the fact that songs with high popularity with no explicit content, contribute to the basic profit of the company. This paper will present a method in which the basic characteristics of a song such as accousticness, valence, danceability, etc will be analyzed to predict its popularity and its existence of explicit content. In particular, a Management Information System (MIS) will be synthesized with the purpose of any person, even if he lacks of technology and mathematical knowledge, to be able to give as an input a specific song and gets as an output either its probable popularity or the appearance of explicit content. Therefore, the user considering these predictions can decide whether it is worth releasing it or not. More specifically, several statistical tests will be used to find the most related song's characteristics to popularity and a variety of machine learning algorithms will be performed to build the most preferable and accurate predictions. Consequently, all the above research will contribute to an efficient programming and provision of the MIS. Finally, the last section will present our results.

2 MIS Design

MIS stands for Management Information System. More specifically, it is a system that was designed with the purpose of providing financial and quantitative information to all levels of management in a company [3]. In particular, it provides a system that its user regardless of his technological and mathematical knowledge will be able to find and produce helpful insights for the company.

As far as Spotify is concerned, an MIS will be designed to predict the popularity of a song and the existence of an explicit content on it. The purpose of this MIS is that a user working in the music industry can just provide the characteristics of a song and will have as a prediction its popularity and the potential existence of explicit content. Therefore, he will be able to decide whether he prefers this song to be released or not.

As this menu is referring to all kind of users regardless of their mathematical background, needs to be simple and easy to interfere. For this purpose, a UML design will be provided to simplify the internal mathematical procedures and strategies for the creation of this menu. Unified Modeling Language (UML) is a graphical language that provides through visualizations and graphs, the documentation and the structure of a system [4]. The UML regarding this paper's purpose is presented on a separate PDF file.

In particular, this menu will be a window in which the user will perform its preferable questions. As input the user will just provide the characteristics of the songs that he wishes to investigate. Subsequently, on the right side of the window there will be two buttons that represent the predictions. The first one will provide the popularity of the song and the second one the probable existence of explicit content. The design of this menu is presented on figure 14 (*Appendix E*).

3 Business Analysis

3.1 Data Import

The dataset used for this report, contains Spotify data from 1922 to 2021. This specific dataset was retrieved from Kaggle and it consists of 170,653 observations of 19 variables [5]. In particular, it contains a variety of audio features from a huge amount of songs released until 2021. These characteristics are valence, accousticness, artist, danceability, duration, energy, explicit content and popularity among others. An extensive explanation of all the variables is provided on Table 1 (*Appendix A*).

The questions occurred on the previous section, were the prediction of a song's popularity given specific audio characteristics and whether it contains explicit content or not. For this accomplishment, machine learning algorithms will be used and more specifically:

- k Nearest Neighbors for the prediction of popularity
- Logistic Regression for the prediction of explicit content

More specifically, this report will use SAS for handling the above concerns, which is a well-known and efficient programming language especially for business analytics. The necessity of each machine learning algorithm for answering each question is presented extensively on the next chapters.

3.2 Pre-processed steps

Performing the described methods, we will be able to supply the MIS with tools that will lead to the answer of the preferable question provided by the user. For the accurate programming for prediction models, a proper training of each model is essential. More specifically, the dataset needs to be divided into two subdatasets. The training

set which will contain the 70% of the data and will be used for the training procedure. The remaining 30% of the data will be used as a test set to evaluate the accuracy of the prediction model.
For efficient statistical results and insights, it is essential to investigate whether the dataset contains any missing values or not. As shown on Table 2 (*Appendix A*), it is concluded that no missing values are existed on the dataset, so any further analysis can be proceeded.
Prior to machine learning algorithms, a sufficient statistical analysis needs to be occurred. This will contribute to a better insight of the data and will provide relations among the characteristics. Therefore, audio features that most affect the popularity of a song and the explicit content, will be used as the variables for the modeling procedure.

3.3 Descriptive Analytics

Considering table 1 (*Appendix A*), it is observed that popularity is a number between 0 and 100 representing the percentage of popularity of a track. For the purpose of investigating the popularity of a song, another variable is created, called popular. In particular, it is a qualitative variable of three classes which describes if a song is not popular when popularity is less than 50 (classed as 0), slightly popular when popularity is between 50 and 70 (classed as 1) and highly popular when popularity is greater than 70 (classed as 2). Furthermore, the other concern of this paper is the existence of an explicit content on a song. As shown in table 3 (*Appendix B*), most of the songs (91.54%) do not contain explicit content and 77.28% are not popular songs. More specifically, only 3% of the songs are highly popular and the majority (67.40%) do not contain explicit content. However, a boxplot of these two variables, as shown in figure 1 (*Appendix B*), indicates that the mean value of popularity is higher for tracks having explicit content.

Both for the prediction of explicit content and popularity, numerous of variables such as accousticness, danceability, duration, instrumentalness, loudness and speechiness will be used. The descriptive statistics of these variables is shown on table 4 (*Appendix B*). It is indicated that the mean value of songs' popularity is approximately 31%.

Furthermore, prior to every analysis it is highly essential for the variables' distribution to be investigated. On figures 2 - 8 (*Appendix B*), where the normal distribution curve is drawn on each variable's one, it is concluded that most of the variables are normally distributed. However, the dataset contains 170,000 observations and thus due to the central limit theorem even for cases in which the normal distribution is not certain, the normality can be assumed [6].

3.4 Machine learning algorithms

In this section, an effective strategy to predict the popularity using k-NN will be presented as well as the prediction of the explicit content using logistic regression.

On table 5 (*Appendix C*), the Pearson correlation number among all the variables that will be used on the upcoming models, is presented. More specifically, it is concluded that popularity is highly correlated with accousticness and loudness and slightly with danceability. However, for an accurate prediction of the popularity an artificial intelligent model using the k-NN method will be created.

k Nearest Neighbors

K Nearest Neighbors (k-NN) is a common and effective method for both classification and regression but more efficient on the first one. It is based on the idea that observations that belongs on the same class have similar characteristics [7]. This approach will be used to predict whether a song will be slightly or highly popular or even not popular. The purpose of this, is to help a music producer to decide the release of a song or not.

From technical perspective, the independent variables will be accousticness, danceability, duration, instrumentalness, loudness and speechiness. The predictor will be the variable popular with no popular coded as 0, slightly popular as 1 and highly popular as 2. The model will use 5 nearest neighbors and will be fitted regarding the training data and use the test data for its evaluation.

The results of the fitting are shown extensively on table 6 (*Appendix C*). It is shown that most of the observations were classified correctly for its category. As a result, the provided model it appears to be highly efficient. Reviewing table 7 (*Appendix C*) it is shown that the error of this model is approximately 0.3 which is extremely low. Consequently, the provided model can predict highly accurate whether a song will gain popularity or not.

Logistic regression

Logistic regression is a machine learning algorithm which is used mostly for binary classification. In particular, it is used for 2-level outcomes and thus it can provide the possibility of a variable to belong in one class [8].

Considering this paper, this method will be helpful for the MIS as it can be used to predict whether a song will contain explicit content or not. From technical perspective it takes for input the variables accousticness, danceability, duration, instrumentalness, loudness and speechiness and outcomes whether a song will probably contain explicit content, as shown on figure 9 (*Appendix C*).

The results of the fitting procedure are presented on table 8 (*Appendix C*). It is shown that the probability of each variable's estimate to be 0 is below 0.001 and thus all the variables will be used by the model for predicting the explicit content. This argument is enhanced by reviewing the confidence level from table 9 (*Appendix C*). Therefore, the fitted logistic regression model will be used on MIS to predict efficiently the presence of explicit content on the provided song.

4 Efficient programming

This section will try to enhance the efficiency of the MIS. For this accomplishment, several changes need to be done to improve the time and the reserved space of all the above procedures. The reduction of the time and space needed for each analysis, will contribute to a more efficient programming.

At first, it is important, prior to each analysis, to create a library in SAS and to import the dataset through this library. As a result, it will be less complex for SAS to use the dataset as it will be saved on its repository. Moreover, to reduce the complexity for fitting the models it is essential to remove the non-missing values. Apart from the missing values is more efficient to keep only the variables that will be needed for the analysis. Taking all the above into consideration, a new dataset will be created through a SAS library and will contain no missing values and only the variables accousticness, danceability, duration, instrumentalness, loudness, speechiness, explicit and the new variable popular which was created for the paper's purposes. Afterwards, the training and test data will be created considering this new dataset and they will be used to perform the machine learning algorithms.

On figure 10 (Appendix D) it is shown that prior to efficient programming the time for fitting of the k-nn model was 17.09s and needed 14,361k memory. After the efficient programming, as shown on figure 11 (Appendix D) the time was reduced to 16.90s and the memory to 14,341k. As far as the logistic regression is concerned, before the efficient programming the model needed 0.88s and 24119.25k of memory, as shown on figure 12 (Appendix D). After the efficient programming both the time and memory were reduced. More specifically, reviewing figure 13 (Appendix D) the time needed for the model was 0.80s and 24090.68k of necessary space.

Therefore, the steps described in this chapter contributed to the efficient programming. Both k-nn and logistic regression are now need less time and memory for SAS to fit them.

5 Conclusion

The purpose of this paper was to provide an efficient and helpful Management Information System for music industries. More specifically, it provided an MIS in which a music producer, by giving the characteristics of a song, has the ability to predict the existence of explicit content on this song and whether it will be popular or not. More specifically, for popularity the user can predict if a track will be slightly popular, highly popular or not popular at all. Consequently, a music producer will be able to decide whether it is beneficial to release a specific song as its popularity and its content are highly related to its financial outcome.

From technical perspective, a dataset of Spotify was used as most of the songs nowadays are published directly to online music platforms. This paper used SAS for the creation of the appropriate models that lead to an efficient MIS. More specifically, for the prediction of the popularity, a k-NN was fitted and provided a model with high accuracy. For the prediction of the explicit content a logistic regression method was performed by giving numerous of songs' features. Last but not least, to achieve the best possible approach for the creation of both models, efficient programming was also performed. Due to the efficient programming both the time and the necessary memory for the modeling of each procedure were reduced.

Taking everything into consideration, this paper presented a sufficient strategy for the provision of an effective MIS. Therefore, a music producer can take advantage the artificial intelligent models to decide the releasing of a song without the necessity of mathematical knowledge. As a result, he can help its company to be developed financially.

References

1. Haupt, J. “Spotify”. Digital Media Reviews. Music Library Association. 69(1), pp.132-138, (2012).
2. Spotify, (2021) Company info. [Online] Available from https://newsroom.spotify.com/company-info/. [Accessed: 31st January 2022].
3. Oxford Reference. Management Information System. [Online] Available from https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100129791. [Accessed: 31st January 2022].
4. Booch, G. (1999). “UML in action”. Communications of the ACM. 42 (10), pp. 26–28.
5. Kaggle, (2020) Spotify Dataset 1922-2020. [Online] Available from https://www.kaggle.com/ektanegi/spotifydata-19212020. [Accessed: 31st January 2022].
6. Chang, H-J., “Determination of Sample Size in Using Central Limit Theorem for Weibull Distribution”, Information and Management Sciences. 17 (3), pp.31-46 (2006).
7. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). “KNN Model-Based Approach in Classification”. Lecture Notes in Computer Science. Springer. pp. 986–996.
8. LaValley, M. P. (2008). Logistic Regression. Circulation, 117(18), 2395–2399. doi:10.1161/circulationaha.106.682658

Appendices

Appendix A: Description of the variables

The dataset of Table 1 consists of 170,653 records with 19 variables about Spotify.

Table 1. Description of Variables

	Variable Name	Type	Description
1.	Valence	Numeric	Describes the positivity of each track (From range 0 to 1)
2.	Year	Integer	The year of each registration
3.	Accousticness	Numeric	The accousticness of each song from range 0 to 1
4.	Artists	Character	The artist of the registered song
5.	Danceability	Numeric	Describes how suitable the song is for dancing (from range 0 to 1)
6.	Duration_ms	Numeric	The duration of each song in milli-seconds
7.	Energy	Numeric	Describes how energetic the track is (from range 0 to 1)
8.	Explicit	Categorical	Whether the song contains explicit content (1) or not (0).
9.	Id	Character	The id of each track created by Spotify
10	Instrumentalness	Numeric	The ration of instrumental sounds
11.	Key	Categorical	The key of each song, described as categories from 1 to 11
12.	Liveness	Numeric	Describes the probability of the song to be recorded live (from range 0 to 1)
13.	Loudness	Numeric	The loudness of the song in dB ranged from -60 to 0
14.	Mode	Categorical	The scale of the track. 0: Minor 1: Major
15.	Name	Character	The name of each song
16.	Popularity	Integer	The popularity of each track.
17.	Release date	Integer	The released year of the album containing the specific track.
18.	Speechiness	Numeric	The ration of spoken words
19.	Tempo	Numeric	The tempo of each song in BPM

Variable	N	N Miss
valence	170653	0
year	170653	0
acousticness	170653	0
danceability	170653	0
duration_ms	170653	0
energy	170653	0
explicit	170653	0
instrumentalness	170653	0
key	170653	0
liveness	170653	0
loudness	170653	0
mode	170653	0
popularity	170653	0
speechiness	170653	0
tempo	170653	0

Table 2. Number of Missing Values on each variable

Appendix B: Descriptive Statistics

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of explicit by popular				
	popular			
explicit	0	1	2	Total
0	125038 73.27 80.04 94.81	27802 16.29 17.80 82.36	3380 1.98 2.16 67.40	156220 91.54
1	6844 4.01 47.42 5.19	5954 3.49 41.25 17.64	1635 0.96 11.33 32.60	14433 8.46
Total	131882 77.28	33756 19.78	5015 2.94	170653 100.00

Table 3. Crosstabulation matrix for explicit and popular

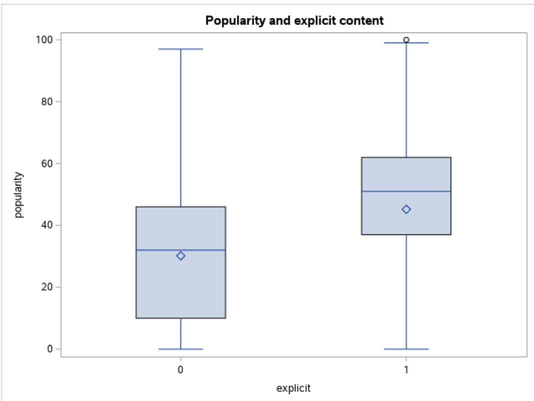

Fig. 1 Boxplot of popularity and explicit content

Variable	N	Mean	Std Dev	Minimum	Maximum
acousticness	170653	0.5021148	0.3760317	0	0.9960000
danceability	170653	0.5373955	0.1761377	0	0.9880000
duration_ms	170653	230948.31	126118.41	5108.00	5403500.00
instrumentalness	170653	0.1670096	0.3134747	0	1.0000000
loudness	170653	-11.4679900	5.6979429	-60.0000000	3.8550000
speechiness	170653	0.0983933	0.1627401	0	0.9700000
popularity	170653	31.4317943	21.8266151	0	100.0000000

Table 4. Descriptive Statistics

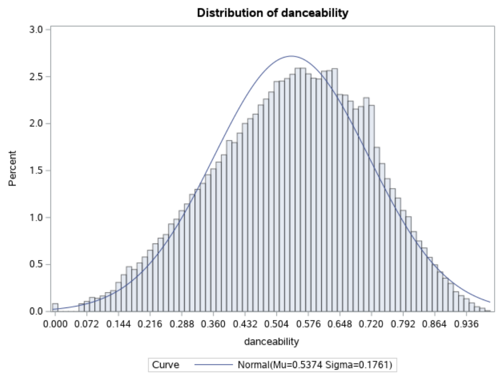

Fig. 2 Normality check for Danceability

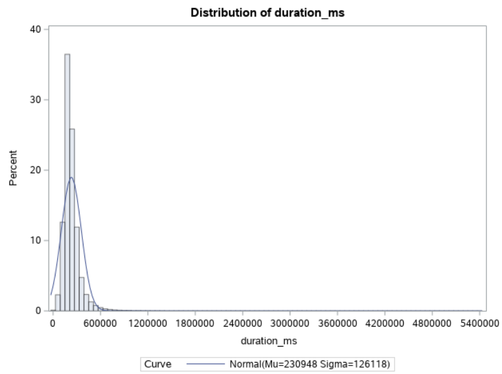

Fig. 3 Normality check for Track's Duration (in ms)

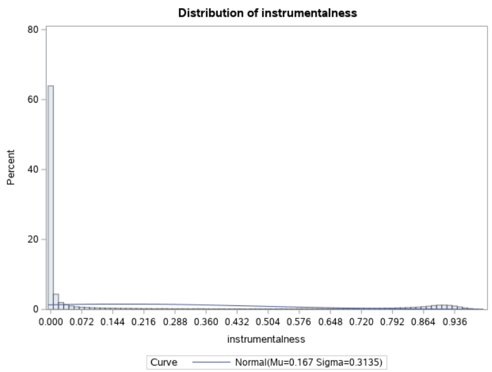

Fig. 4 Normality check for instrumentalness

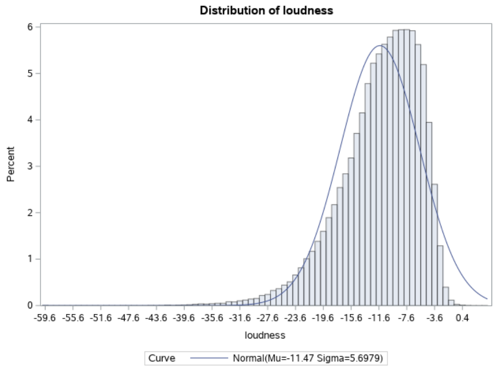

Fig. 5 Normality check for loudness

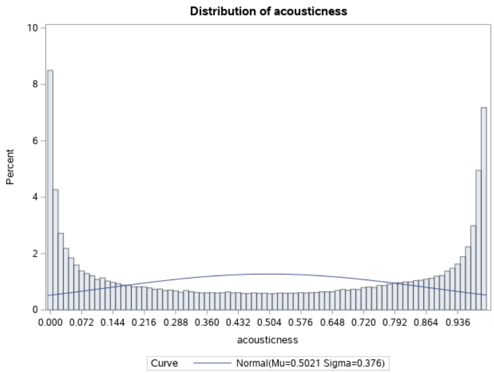

Fig. 6 Normality check for accousticness

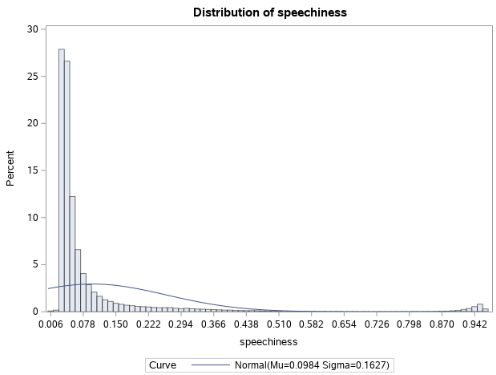

Fig. 7 Normality check for speechiness

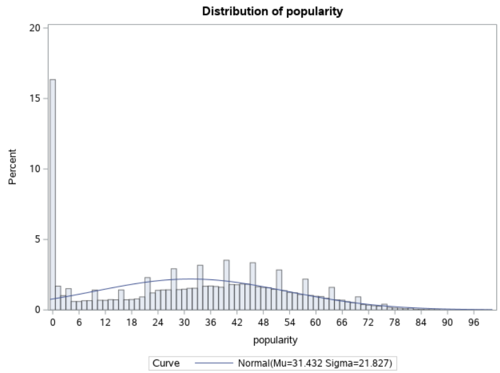

Fig. 8 Normality check for popularity

Appendix C: Machine Learning Algorithms

Pearson Correlation Coefficients, N = 170653 Prob > |r| under H0: Rho=0							
	acousticness	danceability	duration_ms	instrumentalness	loudness	speechiness	popularity
acousticness	1.00000	-0.26685 <.0001	-0.07637 <.0001	0.32982 <.0001	-0.56170 <.0001	-0.04398 <.0001	-0.57316 <.0001
danceability	-0.26685 <.0001	1.00000	-0.13994 <.0001	-0.27806 <.0001	0.28506 <.0001	0.23549 <.0001	0.19961 <.0001
duration_ms	-0.07637 <.0001	-0.13994 <.0001	1.00000	0.08477 <.0001	-0.00304 0.2096	-0.08460 <.0001	0.05960 <.0001
instrumentalness	0.32982 <.0001	-0.27806 <.0001	0.08477 <.0001	1.00000	-0.40861 <.0001	-0.12170 <.0001	-0.29675 <.0001
loudness	-0.56170 <.0001	0.28506 <.0001	-0.00304 0.2096	-0.40861 <.0001	1.00000	-0.13930 <.0001	0.45705 <.0001
speechiness	-0.04398 <.0001	0.23549 <.0001	-0.08460 <.0001	-0.12170 <.0001	-0.13930 <.0001	1.00000	-0.17198 <.0001
popularity	-0.57316 <.0001	0.19961 <.0001	0.05960 <.0001	-0.29675 <.0001	0.45705 <.0001	-0.17198 <.0001	1.00000

Table 5. Pearson Correlation among all variables

Number of Observations and Percent Classified into popular				
From popular	0	1	2	Total
0	30570 77.91	5776 14.72	2892 7.37	39238 100.00
1	2652 25.90	5307 51.83	2280 22.27	10239 100.00
2	109 7.16	184 12.08	1230 80.76	1523 100.00
Total	33331 65.35	11267 22.09	6402 12.55	51000 100.00
Priors	0.33333	0.33333	0.33333	

Table 6. k-NN for popular

Error Count Estimates for popular				
	0	1	2	Total
Rate	0.2209	0.4817	0.1924	0.2983
Priors	0.3333	0.3333	0.3333	

Table 7. k-NN error

Model Information	
Data Set	WORK.TRAIN_SPOTIFY
Response Variable	explicit
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	119000
Number of Observations Used	119000

Response Profile

Ordered Value	explicit	Total Frequency
1	1	10054
2	0	108946

Probability modeled is explicit='1'.

Fig. 9 Fitting of the logistic regression

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6395	0.0724	1330.5519	<.0001
acousticness	1	-2.9537	0.0599	2435.4205	<.0001
danceability	1	4.0292	0.0868	2152.4531	<.0001
duration_ms	1	-1.9E-6	1.714E-7	123.3347	<.0001
instrumentalness	1	-1.0182	0.1005	102.5906	<.0001
loudness	1	0.2296	0.00420	2988.2235	<.0001
speechiness	1	7.0890	0.0747	9011.5947	<.0001

Table 8. Estimates for each variable on logistic regression model

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
acousticness	0.052	0.046	0.059
danceability	56.217	47.418	66.649
duration_ms	1.000	1.000	1.000
instrumentalness	0.361	0.297	0.440
loudness	1.258	1.248	1.268
speechiness	>999.999	>999.999	>999.999

Table 9. Ratio Estimates on logistic regression

Appendix D: Efficient Programming

```
NOTE: There were 119000 observations read from the data set WORK.TRAIN_SPOTIFY.
NOTE: There were 51000 observations read from the data set WORK.TEST_SPOTIFY.
NOTE: The data set WORK.TESTOUT has 51000 observations and 24 variables.
NOTE: PROCEDURE DISCRIM used (Total process time):
      real time           17.09 seconds
      user cpu time       17.09 seconds
      system cpu time     0.01 seconds
      memory              14361.37k
      OS Memory           40912.00k
      Timestamp           31/01/2022 07:44:17 μ.μ.
      Step Count                        58  Switch Count  18
      Page Faults                       0
      Page Reclaims                     3133
      Page Swaps                        0
      Voluntary Context Switches        74
      Involuntary Context Switches      17
      Block Input Operations            0
      Block Output Operations           32048
```

Fig. 10 k-NN prior to efficient programming

```
NOTE: There were 119000 observations read from the data set WORK.TRAIN_SPOTIFY_EFF.
NOTE: There were 51000 observations read from the data set WORK.TEST_SPOTIFY_EFF.
NOTE: The data set WORK.TESTOUT_EFF has 51000 observations and 13 variables.
NOTE: PROCEDURE DISCRIM used (Total process time):
      real time           16.91 seconds
      user cpu time       16.90 seconds
      system cpu time     0.02 seconds
      memory              14341.84k
      OS Memory           41168.00k
      Timestamp           31/01/2022 07:58:00 μ.μ.
      Step Count                        107  Switch Count  7
      Page Faults                       0
      Page Reclaims                     3029
      Page Swaps                        0
      Voluntary Context Switches        45
      Involuntary Context Switches      26
      Block Input Operations            0
      Block Output Operations           9776
```

Fig. 11 k-NN after the efficient programming

```
NOTE: PROC LOGISTIC is modeling the probability that explicit='1'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 119000 observations read from the data set WORK.TRAIN_SPOTIFY.
NOTE: The data set WORK.OUTDATA has 119000 observations and 24 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.88 seconds
      user cpu time       0.82 seconds
      system cpu time     0.07 seconds
      memory              24119.25k
      OS Memory           52120.00k
      Timestamp           31/01/2022 07:45:59 μ.μ.
      Step Count                        64  Switch Count  16
      Page Faults                       0
      Page Reclaims                     5415
      Page Swaps                        0
      Voluntary Context Switches        57
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           107336
```

Fig. 12 Logistic regression prior to efficient programming

```
NOTE: PROC LOGISTIC is modeling the probability that explicit='1'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 119000 observations read from the data set WORK.TRAIN_SPOTIFY_EFF.
NOTE: The data set WORK.OUTDATA has 119000 observations and 13 variables.
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.80 seconds
      user cpu time       0.75 seconds
      system cpu time     0.06 seconds
      memory              24090.68k
      OS Memory           52632.00k
      Timestamp           31/01/2022 08:01:47 μ.μ.
      Step Count                        151  Switch Count  8
      Page Faults                       0
      Page Reclaims                     5443
      Page Swaps                        0
      Voluntary Context Switches        37
      Involuntary Context Switches      1
      Block Input Operations            0
      Block Output Operations           55120
```

Fig. 13 Logistic regression after efficient programming

Appendix E: Menu Design

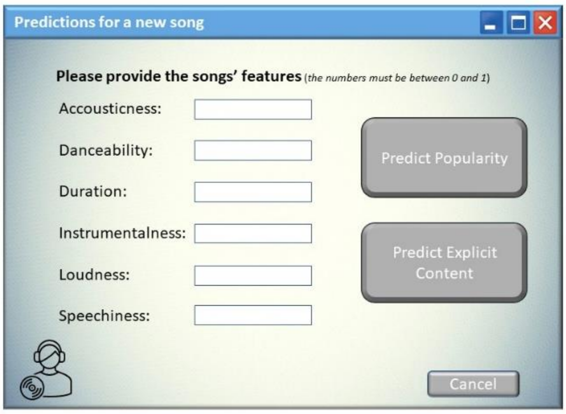

Fig. 14 Menu Design