Group-14 Project Report

Team:
Ian McIntosh
David Gong
Raffi Mannarelli
Anvesh Karangula
Raghav Vaidya

1. Description of Project Goals:

1.1 Description

The dataset we are investigating is a dataset of songs from the years 1960 - 2019. This data is taken from Spotify and we found this dataset on Kaggle. The dataset contains 19 columns with 1 column predicting popularity (0 or 1) and 18 columns for predictors. The predictors are many features of the song that relate to things such as the audio qualities or the duration of the songs. In the dataset, 1 is considered a popular song and 0 is considered a flop. A song is classified as popular if it was on the billboard 100 and it was in the US. The flops are also considered if it is from the US. This dataset has around 41 thousand rows of data and it has 50 percent 1s and 50 percent 0s.

The question we are trying to answer is what are the important features of a hit song and are we able to predict if a song will be a hit or not based on these features. Our goal is to be able to put in a song's parameters and get a return of if it is popular or not.

1.2 Importance of the problem

We are looking at this problem from the point of view of a third party music consulting firm. For years labels have been wondering what exactly it is that makes a song a hit. Many of the companies we work for have produced countless hits but also countless misses, and never know which category the song will fall into until it is released. They want a way to be able to tell for certain whether a song will be a hit before they release it to the public, thereby taking the guesswork out of selecting which songs to release.

That's where we come in. We want to produce a model that can predict how popular a song is going to be before anyone outside a label has heard it. This model can take in certain features of a song, such as how easy it is to dance to, the song length, whether it is explicit and much more, and compare it to songs in the past that have been similar to it. The model can then see how these similar songs have been doing in the past, before making a final prediction about the relative popularity of the song. From there, our work is done, and it is up to the label to decide whether the song will be popular enough to merit releasing it to the public.

We will be able to inform record label companies how popular and potentially profitable their songs will be. Using feature importance we can see what aspects of a song our model weighs the most in determining its

popularity. Therefore, even after companies make an informed decision on the songs they have, they will know what they should look for in the songs they produce going forward.

2. Exploratory Data Analysis

We have done a summary statistics of all the predictors to see if there are any outliers to be taken care of. The summary statistics are all shown in **Figure 1**.

We decided to drop the unimportant columns of artist name, song name, and URI - the unique resource identifier since these are redundant information.

The predictors in the model are as follows: time signature, number of sections, danceability (how suitable a track is for dancing based on elements such as tempo and regularity), energy (how intense the song is), key, loudness, mode (major is 1, minor is 0), speechiness (how many words are in the track), acousticness(is the track acoustic), duration (length of song in milliseconds), liveness (presence of an audience), valence (describes the music positiveness), tempo (beats per minute), chorus hit (when does chorus start). The data was already scaled so everything is from 0-1.

We also found out that there is no missing data. Hence, there is no requirement to manipulate the data. We also analyzed the distribution of various audio features like danceability, energy, and loudness to understand the characteristics of top hits. We constructed the correlation between popularity and different audio features to identify which features are strongly associated with higher popularity. We plotted the correlation matrix to better represent the correlations if it exists. The correlation matrix is shown in **Figure 2**.

From the correlation matrix, we can notice the correlation between danceability, loudness, instrumentalness and target variable which is popularity of the song. Even though the correlation coefficient are statistically significant, we cannot say that these cause the popularity since correlation ignores the interactions between the predictors. We can also notice few of the interactions between predictors from the correlation coefficients between these predictors. Some of these noticeable interactions are loudness vs energy, sections vs duration, acousticness vs energy.

We looked at a couple of relationships between a predictor and the response variable. In **Figure 3**, it shows that generally songs with low speechiness tend to more likely be a hit. In **Figure 4**, it shows that the louder a song is, the more likely it is to be a hit. In **Figure 5**, it shows that the more danceable a song is, the more likely it is to be a hit.

With these columns, we ran all the models we wanted to and got a decently high accuracy of around 80 percent. However, we looked at the results and realized that some of the top features were not relevant. For example, the top feature for all of them were instrumentalness, which is a number from 0-1 with 1 predicting that the song was all instrumentals.

This was not beneficial because most songs on the list were at 0 as they had lyrics in the song. This skewed our feature importance and led the models to believe that instrumentalness was the most important feature.

Since the data of instrumentalness is skewed (refer **Figure 6**), we decided to remove this from the prospective predictors to avoid skewing the results towards instrumentalness. After we removed this column, one of our top features became decades. This was another issue as decade does not help our goal of predicting if a new song is popular enough, as you cannot change the decade that a song gets released. With these, our final models dropped the decade and instrumentalness columns.

3. Solution and Insights

For this project, we framed this problem as a classification task. We aimed to predict whether the song is a hit or not based on the features of the song. We used various machine learning models, such as random forests, KNN, boosting, and logistic regression. We fit these models to find the model with the best accuracy available. We fitted a Knn model, but the model accuracy we received was very low with around 60 percent, so we ended up not using it.

We first split the data into training and testing data, with 80 percent train and 20 percent test. We then used grid search cross validation to find the best parameters for each of the models. After training and evaluating the models, we found that random forest gave the most accurate prediction for the test set of 76.23 percent. The boosting model was a close second with an accuracy of 75.97 percent. The final model, logistic regression, gave us an accuracy of 68.99 percent. We created a confusion matrix in **Figure 7** and then we calculated the accuracies from that.

The baseline accuracy is 50 percent as the problem is a classification one and there are 50 percent 1s and 50 percent 0s, so all of these 3 models are quite useful.

Looking at the important features, we can see that our 3 models all rank the features with different importances. In our top model, random forest(**Figure 8)**, it ranks accousticness as the most important, with danceabilty at a close second and loudness third. In our second best model, boosting (**Figure 9)**, it ranked danceability as the most important feature, with accousticness at a close second and mode third. In our last model, logistic regression (**Figure 10)**, it ranked loudness as the most important by far, and then energy second with danceability third close behind.

Our least useful predictors as per random forest model are time_signature, key, mode and sections. These offer identification, musical structure, and contextual information but their direct influence on popularity prediction could be limited. Instead, the analysis should be primarily focused on audio features and user behavior patterns to gain insights into the factors driving song popularity on the platform.

Our top two models, random forest and boosting, shows accousticness and danceability as the top 2, so we can say that these are the most important features. This was quite surprising as it did not seem like the most important feature to us. This can be explained with the fact that many people may like hearing music that sounds natural and not electrical. Danceability also makes sense as an important feature as songs that people can dance to in clubs or parties tend to become popular. Loudness is another top feature which also makes sense as people tend to listen to louder songs in parties.

4. Conclusion:

In this project, we explored the top hits dataset on Spotify from 1960 to 2019 and attempted to understand the factors that contribute to a song's popularity. By employing machine learning techniques such as random forest, boosting, and logistic regression, we identified key audio features that heavily influence a song's success on the platform. We decided that our best model was random forest with an accuracy of around 76 percent, and the most important features were the danceability and accousticness of the song.

There are a few limitations with our models. One is that there is oversampling due to the number of flops being equal to the number of hits. Preferences change overtime, and it may be hard to predict a song's popularity in a new time period without seeing how other songs perform in that period first. Another limitation is that we are just predicting whether the song is a hit or flop but not quantifying how popular the song is going to be.

As a consulting firm, we plan to use this model to fit in the songs that are going to be released by the companies we work with and predict the popularity of the song. We also plan to use data analysis done to find out the areas/ locations that the companies need to concentrate inorder to be successful depending on the genre and the type of music.

Further studies that can be done are gathering more dimensional information like where the marketing was done, where the song was produced, target audience etc., We can predict popularity at a granularity of artist, genre, and country to get more accurate predictions. We can have different models predicting popularity in different locations which can take in more predictors more related to the region where the song was produced.

This model can be extended to predict the amount of revenue and the number of streams the song can generate. This would lead to an enhanced business scope.

Figures:

	danceability	energy	key	loudness	mode	\
count	41106.000000	41106.000000	41106.000000	41106.000000	41106.000000	
mean	0.539695	0.579545	5.213594	-10.221525	0.693354	
std	0.177821	0.252628	3.534977	5.311626	0.461107	
min	0.000000	0.000251	0.000000	-49.253000	0.000000	
25%	0.420000	0.396000	2.000000	-12.816000	0.000000	
50%	0.552000	0.601000	5.000000	-9.257000	1.000000	
75%	0.669000	0.787000	8.000000	-6.374250	1.000000	
max	0.988000	1.000000	11.000000	3.744000	1.000000	

	speechiness	acousticness	instrumentalness	liveness	\
count	41106.000000	41106.000000	41106.000000	41106.000000	
mean	0.072960	0.364197	0.154416	0.201535	
std	0.086112	0.338913	0.303530	0.172959	
min	0.000000	0.000000	0.000000	0.013000	
25%	0.033700	0.039400	0.000000	0.094000	
50%	0.043400	0.258000	0.000120	0.132000	
75%	0.069800	0.676000	0.061250	0.261000	
max	0.960000	0.996000	1.000000	0.999000	

	valence	tempo	duration_ms	time_signature	chorus_hit	\
count	41106.000000	41106.000000	4.110600e+04	41106.000000	41106.000000	
mean	0.542440	119.338249	2.348776e+05	3.893689	40.106041	
std	0.267329	29.098845	1.189674e+05	0.423073	19.005515	
min	0.000000	0.000000	1.516800e+04	0.000000	0.000000	
25%	0.330000	97.397000	1.729278e+05	4.000000	27.599792	
50%	0.558000	117.565000	2.179070e+05	4.000000	35.850795	
75%	0.768000	136.494000	2.667730e+05	4.000000	47.625615	
max	0.996000	241.423000	4.170227e+06	5.000000	433.182000	

	sections	target	decade
count	41106.000000	41106.000000	41106.000000
mean	10.475673	0.500000	1982.775264
std	4.871850	0.500006	17.491234
min	0.000000	0.000000	1960.000000
25%	8.000000	0.000000	1970.000000
50%	10.000000	0.500000	1980.000000
75%	12.000000	1.000000	2000.000000
max	169.000000	1.000000	2010.000000

Figure 1: Summary statistics of the predictors

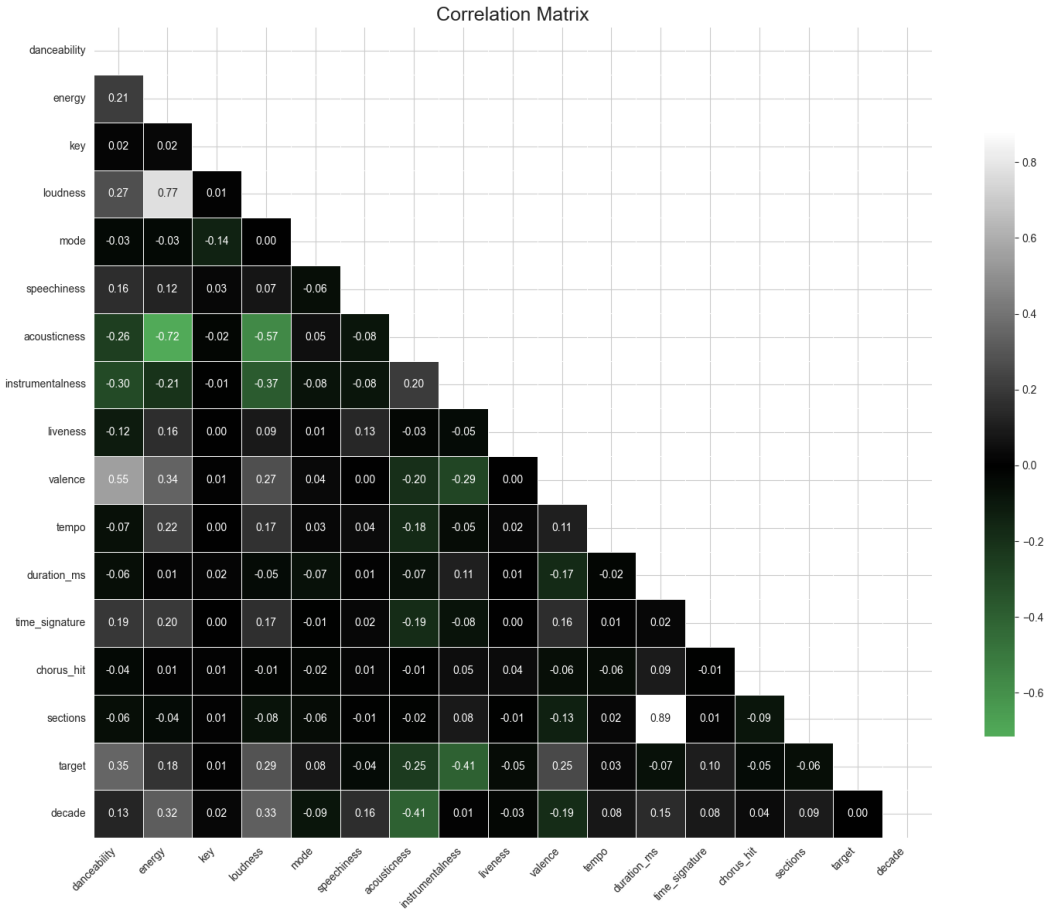

Figure 2: Correlation Matrix

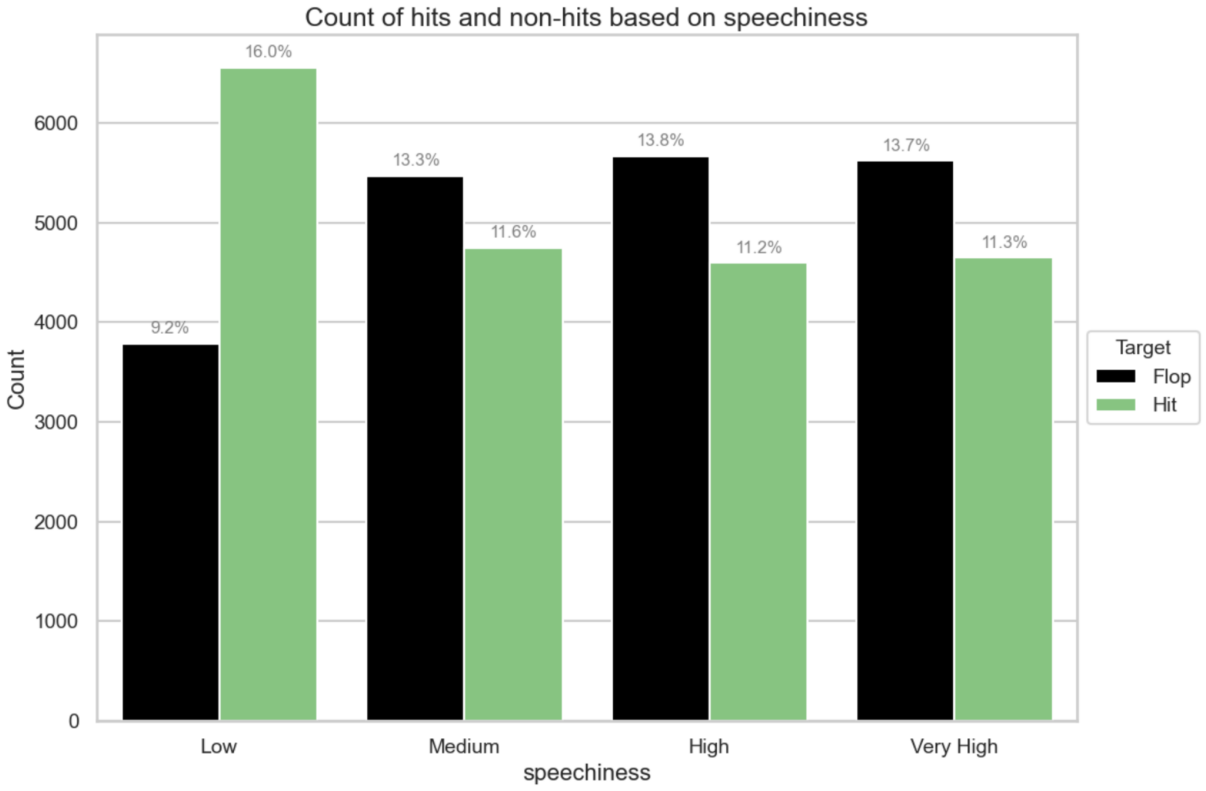

Figure 3: Count of hits and non-hits based on speechiness

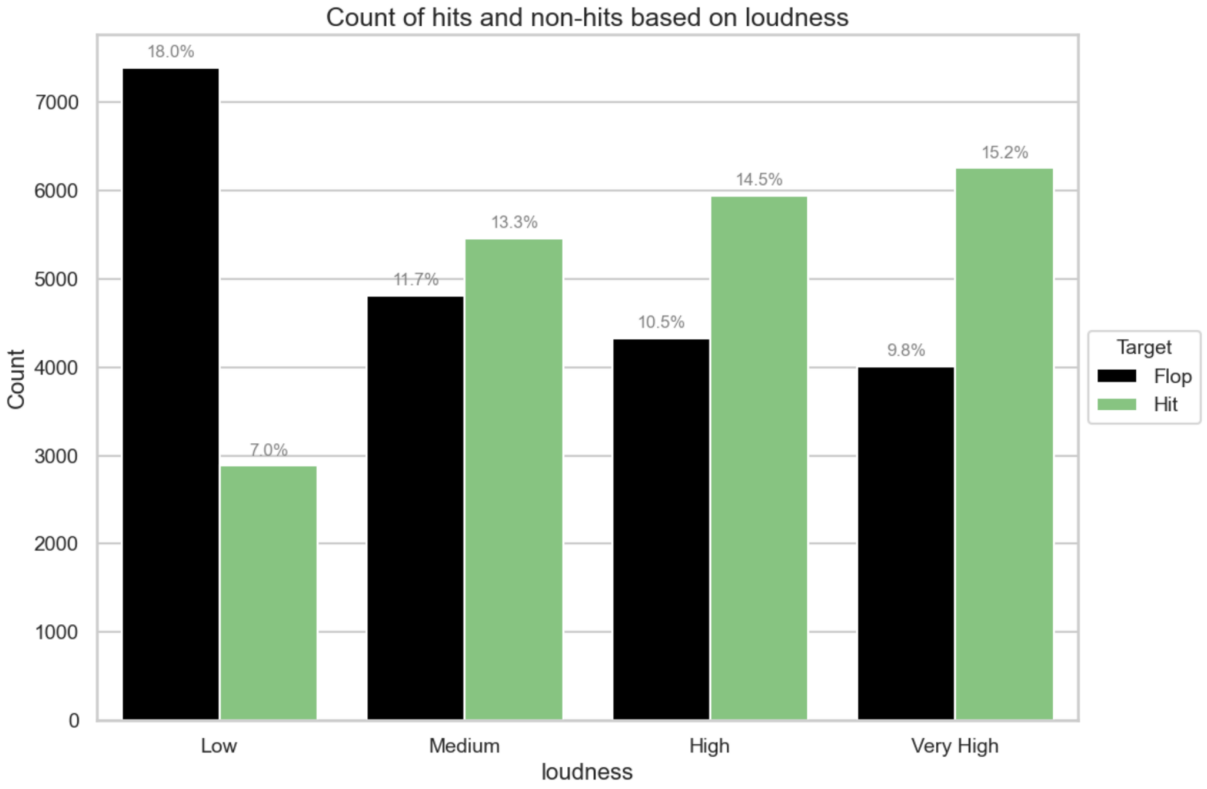

Figure 4: Count of hits and non-hits based on loudness

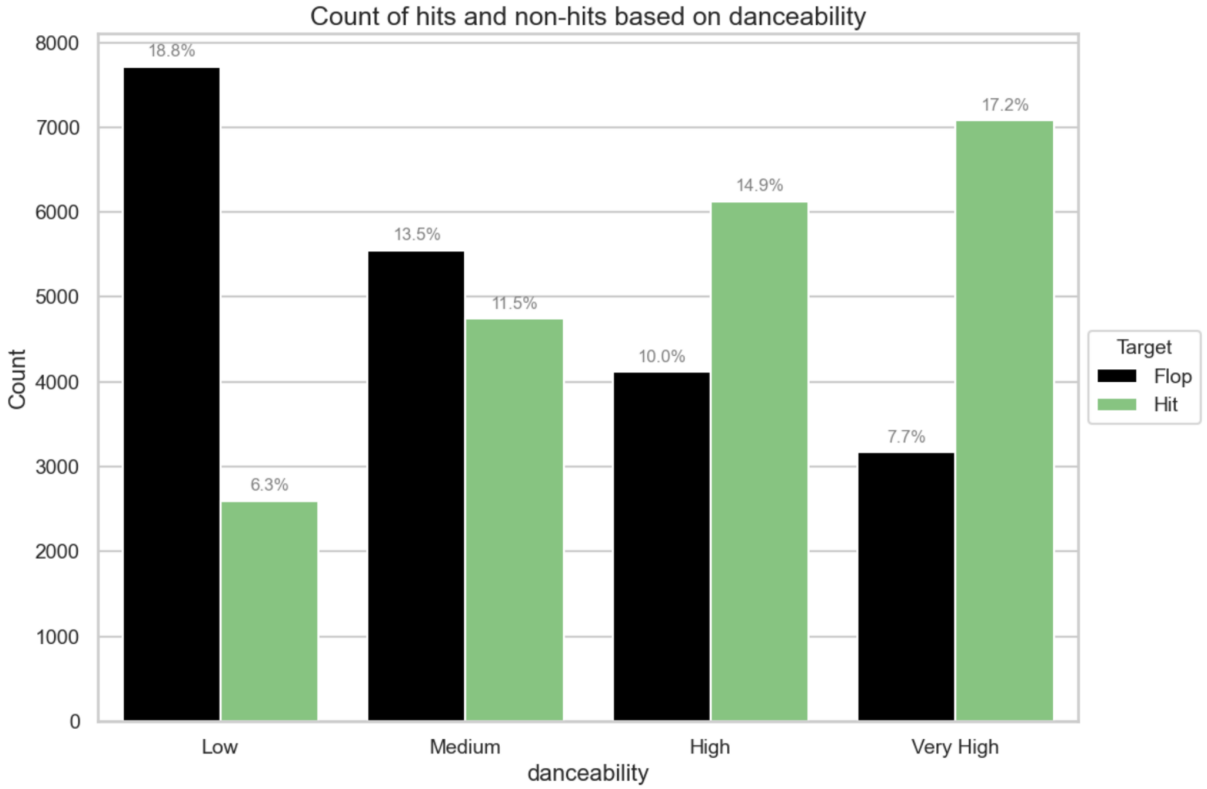

Figure 5: Count of hits and non-hits based on danceability

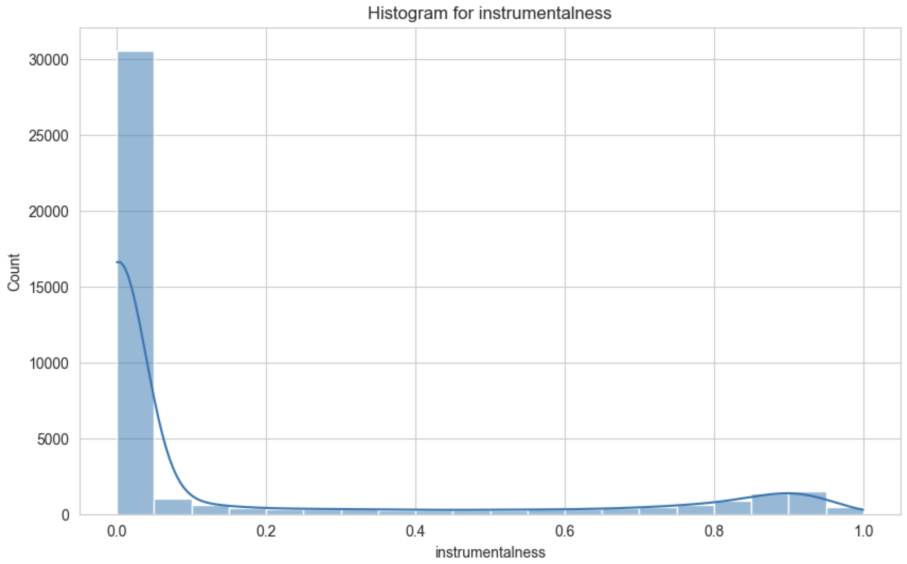

Figure 6: Histogram frequency distribution of Instrumentalness

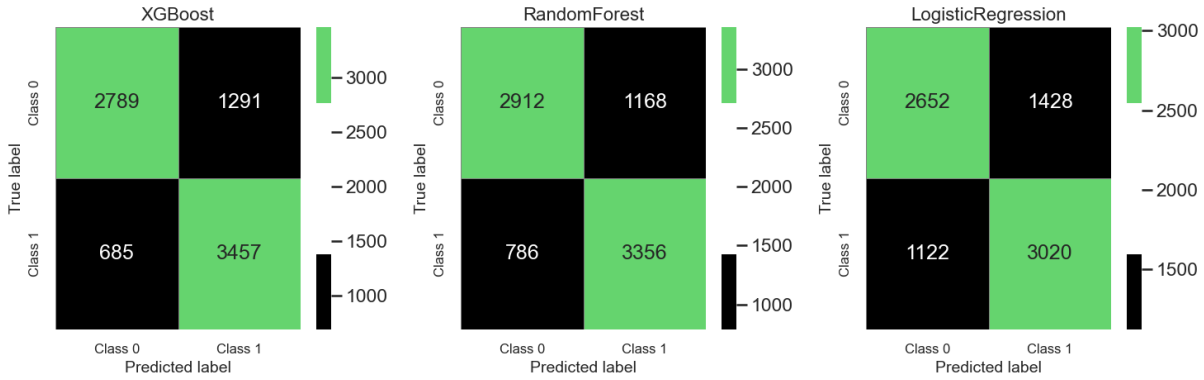

Figure 7: Confusion matrix comparison

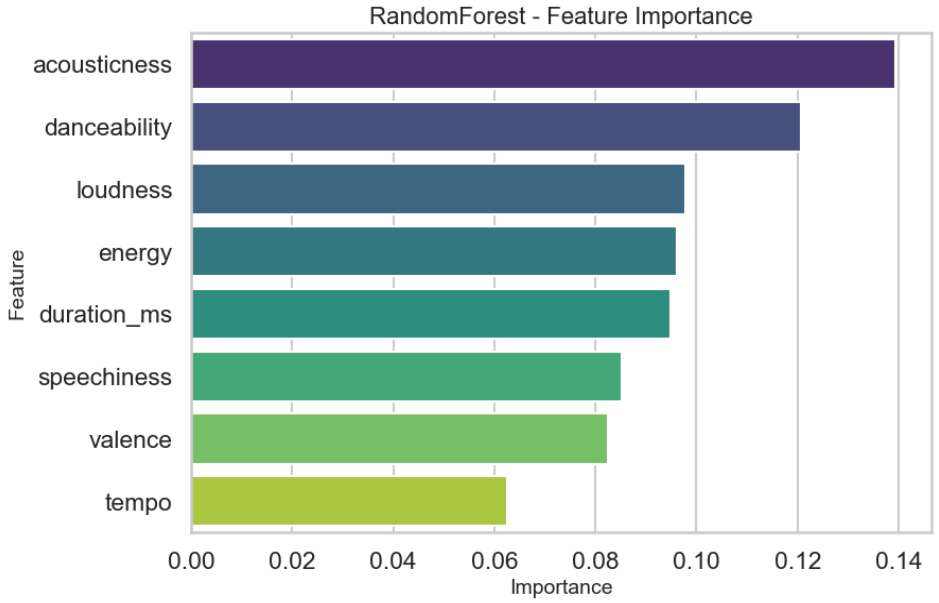

Figure 8: Random Forest - Feature Importance

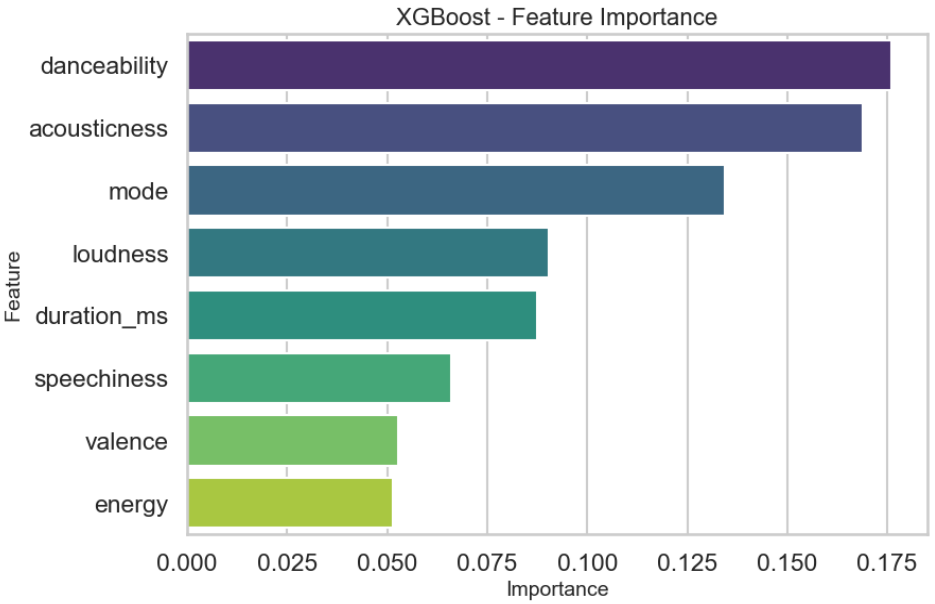

Figure 9: XGBoost - Feature Importance

Figure 10: Logistic Regression - Feature Importance