UCS1625 - FOUNDATIONS OF DATA SCIENCE

PROJECT WORK

ANALYSIS OF TECHNIQUES FOR POPULARITY PREDICTION OF SPOTIFY TRACKS

Aarthi.V.S

185001003

CSE-A

Problem statement

Spotify is a popular music streaming service. It has a huge dataset of songs and the service also collects data about the musical features, the release details and the artists that release those tracks. Spotify also tracks the popularity of each track based on a lot of factors like number of streams, frequency of streams, number of playlists that a track has been added to, how recently the track has been streamed etc.

For an artist or record label a higher popularity of their track on Spotify indicates higher revenue generation. For any music enthusiast it is certainly interesting to analyse why songs become popular and what are the factors affecting the popularity of songs.

This project attempts to build a Spotify popularity predictor and analyse various methods to achieve the same. The project also attempts to identify which factors are most important in determining the popularity of a track. Clustering-then-Regression analysis has also been done.

Literature Survey

A few popular approaches to the problem of music success prediction, despite it having received growing attention for years can be split into three general strategies: the first uses social network data to assess current public perception and extrapolate how successful a song or album will be. The second relies on past data from charts to predict whether a song will be featured in that same chart in the future. The third relies on acoustic information of songs to predict their success.

An example of the first strategy is the work of Dhar and Chang [2]. The authors gathered comments related to 108 albums, before and after they were released, from sources that included social networks and blogs. Their objective was to verify if the sales were reflected in the comments. They also took into consideration data from the songs and artists: the recording label, the time between announcement and release of the album and the reception of the artists' previous work—the number and positiveness of the reviews published in specialized venues, such as the Rolling Stone and Entertainment Weekly magazines, and the scores given by users in online platforms. The authors concluded that the factors that best correlated with sales were the number of posts in social networks and blogs—without checking if those comments were positive or negative—, and the average rating of the artist's previous works by users. At the same time, they also remark that traditional factors, such as traditional media coverage, are still important, and that albums released by larger musical labels attain results 12 times greater.

Herremans, Martens and Sorensen [1] is an example of the second approach. The authors extracted data from the OCC Top 40 Dance Music between 2009 and 2013 and considered a song to be successful if it was featured up to a certain position in the ranking. The authors obtained the better results when the Top as success, the ones in positions #11 to #30 were not used in the experiment and

the last 10 were considered as flop. Using a SVM classifier with polynomial kernel they reached an accuracy of 85% in average.

Among works in the third group, the ones based on features, Lee and Lee [3] collected data on 16; 686 songs that appeared for more than two weeks on Billboard's Hot 100 ranking between 1970 and 2014. For each song, the authors extracted features like chroma, rhythm, timbre, and MFCC. They also employed non acoustic information, including the number of weeks each song was featured in the ranking and the average of their weekly ranks. They trained a Support Vector Machine (SVM) model with RBF kernel and achieved 70% accuracy when predicting which would be the best rank a song would achieve.

Scope

This project is limited to analysing the tracks of a particular music streaming service and as with any streaming service the users of the service are primarily young people. This might affect the model's accuracy in predicting the popularity of tracks across the entire population.

The dataset only contains information about the tracks themselves but not any relevant information about the artists producing the tracks. The dataset also does not contain any historical data about popular songs that might be useful.

The dataset does however contain relevant data from a business use-case standpoint where in the popularity of a track on a service like Spotify generally indicates a good reach and overall "hit" level.

Dataset Description

The dataset was obtained from Kaggle[4].The dataset contains the data about the features of tracks on Spotify, the popular music streaming platform. The data was collected using the Spotify Web API. There are 174389 tracks in the dataset with each track being described by 19 features.

- Primary
    - ID: The primary identifier for the track, generated by Spotify
- Numerical
    - Valence: The positiveness of the track. Higher values mean, the track evokes positive emotions (like joy) otherwise means, it evokes negative emotions (like anger, fear). Ranges between 0-1.
    - Year: The release year of the track.
    - Acousticness: The value that describes how acoustic a song is. Higher values mean that the song is most likely to be an acoustic one. Ranges between 0-1.
    - Danceability: The relative measurement of the track being danceable. Higher values mean that the song is more danceable. Ranges between 0-1.
    - Duration: The length of the track. In milliseconds.
    - Energy: The energy value of the track. Higher values mean that the song is more energetic. Ranges between 0-1.

- Instrumentalness: The relative ratio of the track being instrumental. Higher values mean that the song contains more instrumental sounds. Ranges between 0-1.
- Liveness: Detects the presence of an audience in the recording. Higher values represent an increased probability that the track was performed live. A value above 0.8 provides a strong likelihood that the track is live. Ranges between 0-1.
- Loudness: The overall loudness of a track in decibels (dB). Loudness values are averaged across the entire track and are useful for comparing the relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typical range between -60 and 0 dB.
- Popularity: The popularity of the song. Ranges between 0 and 100. (dependent variable)
- Speechiness: The relative length of the track containing any kind of human voice. Ranged between 0-1.
- Tempo: The tempo of the track in Beat Per Minute (BPM).

- Categorical
  - Key: All keys on octave encoded as values ranging from 0 to 11, starting on C as 0, C# as 1, etc.
  - Mode: Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.
  - Explicit: The binary value whether the track contains explicit content or not.
- Nominal
  - Artists: The list of artists credited for the production of the track.
  - Release Date: The date of release of the track in yyyy-mm-dd, yyyy-mm, or even yyyy format
  - Name: The name of the song.

Proposed solution

The strategies used for popularity prediction are:

1. EDA and Feature Analysis

   Corelation Analysis was done to identify important features and hypothesis Testing for Independence was done using chi-square tests to test for dependencies of the popularity on categorical variables.

2. Linear Regression – Only selected acoustic features

   The acoustic features that had the strongest correlations were used to perform linear regression. The features selected were:

   a. Year
   b. Acousticness
   c. Energy
   d. Loudness
   e. Instrumentalness

3. Linear Regression – Only selected acoustic features and artist information

   The acoustic features that had the strongest correlations along with the encoded artist information were used to perform linear regression. The features selected were:

   a. Year
   b. Acousticness
   c. Energy
   d. Loudness
   e. Instrumentalness
   f. Artists that produced the track

4. Linear Regression – All available information

   All available information was used to perform linear regression after appropriately encoding the artist information and the categorical variables.

5. Decision Tree – only selected acoustic features and artist information

   The acoustic features that had the strongest correlations along with the encoded artist information were used to build a decision tree. The features selected were:

   a. Year
   b. Acousticness
   c. Energy
   d. Loudness
   e. Instrumentalness
   f. Artists that produced the track

   The maximum number of leaf nodes was treated as a hyperparameter and was tuned for least RMSE value.

6. Logistic Regression – All available information

   The tracks were binned into two categories: popular (popularity>=55) and not popular (popularity<55).
   This binned data was used as the dependent variable to train a logistic regression model using all available information as independent variables.

7. Clustering-then-Regression: K-Means Clustering and Decision Tree using only selected acoustic features and artist information

   K-means Clustering was used to add extra genre information to the dataset. The newly added feature along with selected acoustic features and encoded artist information was used to build a decision tree model.

   For the K-means algorithm the optimal k value was obtained by plotting it against the SSE value.

   The decision tree was built using the above mentioned process.

Tools used

1. Python
2. Jupyter Notebook
3. Python Libraries Used:
    a. Numpy
    b. Scipy
    c. Pandas
    d. Scikit-learn
    e. matplotlib

Other Tools/Methodologies

Some other algorithms that can be used are:

1. K Nearest Algorithm
2. Support Vector Machines
3. Lasso Regression
4. Artificial Neural Networks

Some other tools that can be used include :

1. R
2. TensorFlow
3. TensorBoard
4. Tableau

Performance metrics

For the linear regression models (Linear Regression and Decision Tree) the metrics used were:

RMSE: The RMSE is the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. RMSE is an absolute measure of fit. RMSE has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.

R-squared: It signifies the "percentage variation in dependent that is explained by independent variables".

Adj. R-squared: This is the modified version of R-squared which is adjusted for the number of variables in the regression. It increases only when an additional variable adds to the explanatory power to the regression.

Prob(F-Statistic): This tells the overall significance of the regression. This is to assess the significance level of all the variables together unlike the t-statistic that measures it for individual variables. The null hypothesis under this is "all the regression coefficients are equal to zero". Prob(F-statistics) depicts the probability of null hypothesis being true

AIC/BIC: It stands for *Akaike's Information Criteria* and is used for model selection. It penalizes the errors mode in case a new variable is added to the regression equation. It is calculated as number of parameters minus the likelihood of the overall

model. A lower AIC implies a better model. Whereas, BIC stands for *Bayesian information criteria* and is a variant of AIC where penalties are made more severe.

Durbin-Watson: Another assumption of OLS is of homoscedasticity. This implies that the variance of errors is constant. A value between 1 to 2 is preferred.

For the classification model (Logistic Regression) the performance metrics used were:

Accuracy: It is the number of correct predictions made as a ratio of all predictions made.

Precision: It is the ratio between true positives and all actual positive samples.

Recall: It is the ratio between true positives and all samples that the model predicted as positive.

F1-Score: It is the harmonic mean of precision and recall.

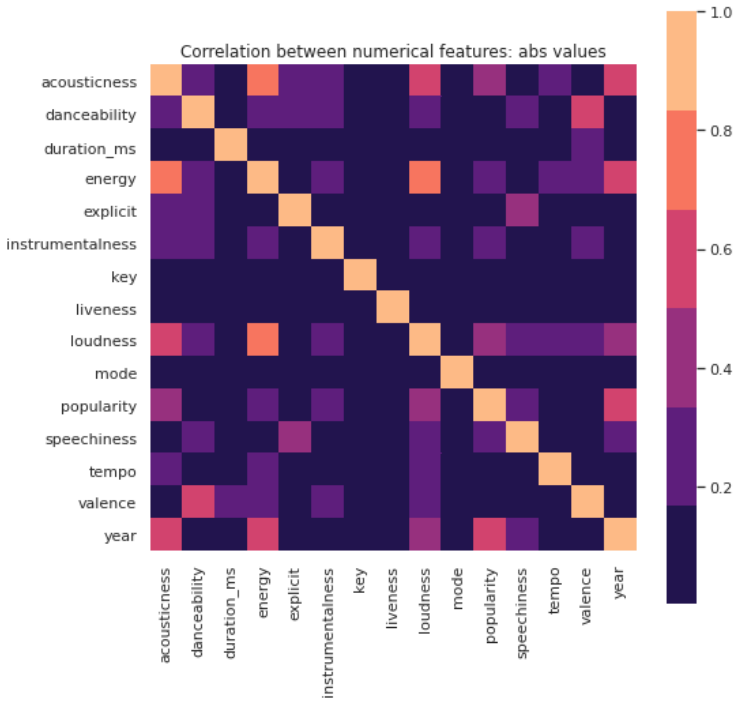
AUC: A curve is plotted between True Positive Rate and False Positive rate. The area with the curve and the axes as the boundaries is called the Area Under Curve(AUC). It is this area which is considered as a metric of a good model. With this metric ranging from 0 to 1, we should aim for a high value of AUC.

Results

1. EDA and Feature Analysis

As a pre-processing step the unnecessary attributes: id, name, and release date, were removed from the dataset.

The correlation of the various features with the dependent variable: popularity was measured and the following Pearson Correlation Matrix was obtained.



The strongest linearly correlated numeric features to popularity are:

Year: 0.51

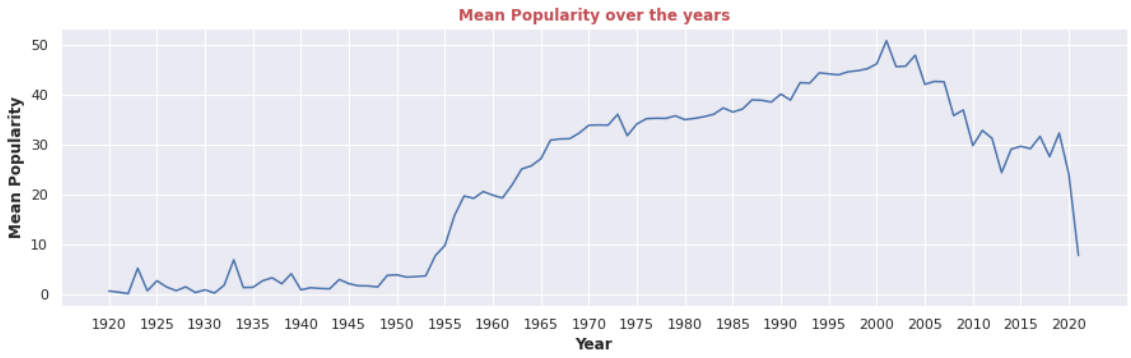
Acousticness: 0.40

Loudness: 0.34

Energy: 0.33

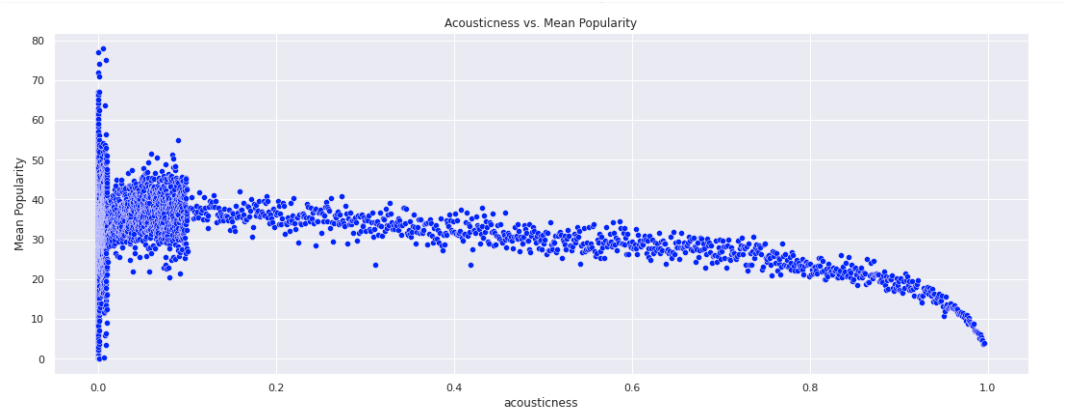
Instrumentalness : 0.30

1) Year:



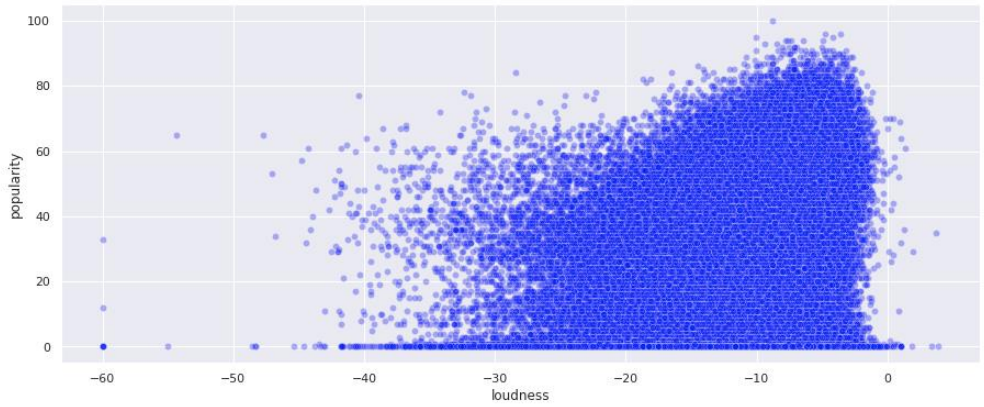
This indicates a steady increase in popularity of songs that are released more recently.

2) Acousticness:



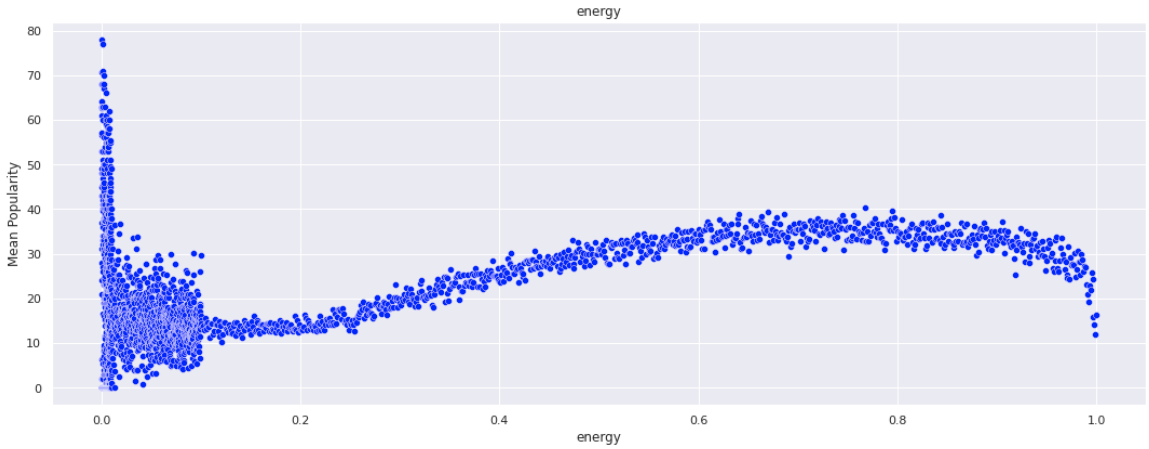
This shows a negative correlation between popularity and Acousticness.

3) Loudness:



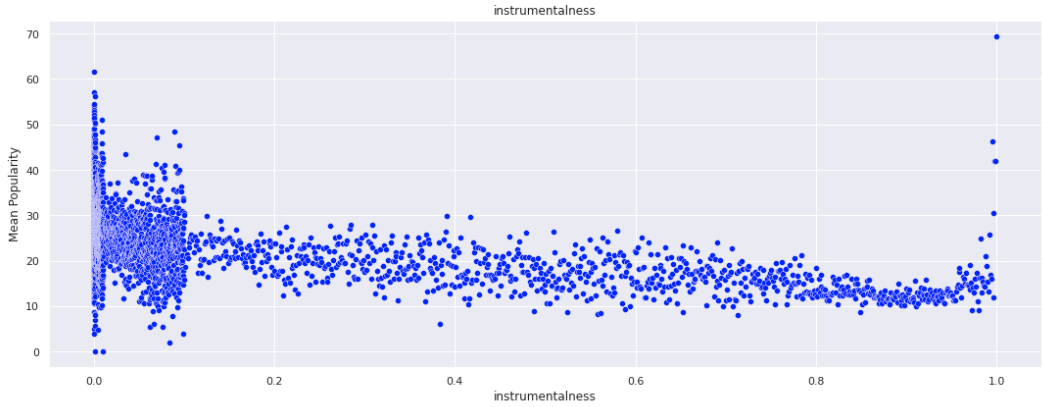
This shows a positive correlation between loudness and popularity.

4) Energy:



This shows a positive correlation between energy and popularity.

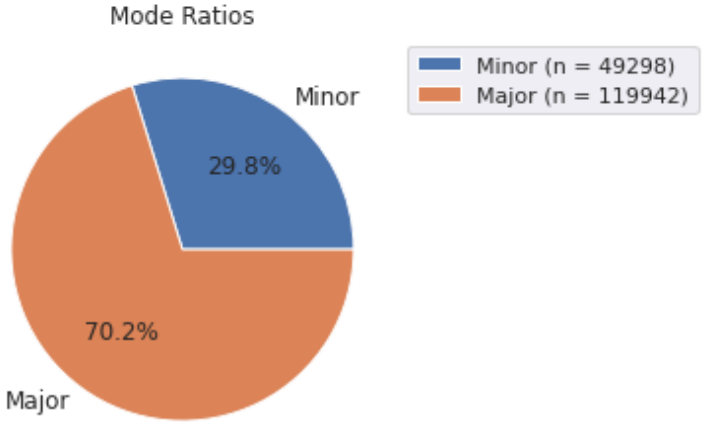
5) Instrumentalness:



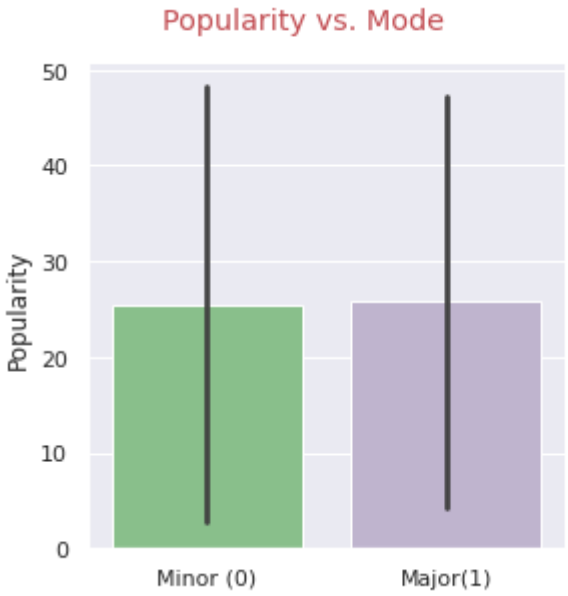
This shows a negative correlation between popularity and Instrumentalness.

Analysis of Categorical features

1) Mode



Major mode songs are much more common in the dataset.



The above bar graph shows that major mode songs are slightly more popular than minor mode songs.

A chi-square test for independence between the mode and popularity of tracks in the dataset gave the following results:

Null Hypothesis : The mode and popularity of songs in the dataset are not related.

Alternate Hypothesis: The mode and popularity of songs in the dataset are related.

A Chi-Square test of independence was performed, and the following results were o obtained:

Contingency Table:

Popularity / Mode	<25	25-50	50-75	>75
Major	58927	43980	18455	1126
Minor	25904	16879	8428	690

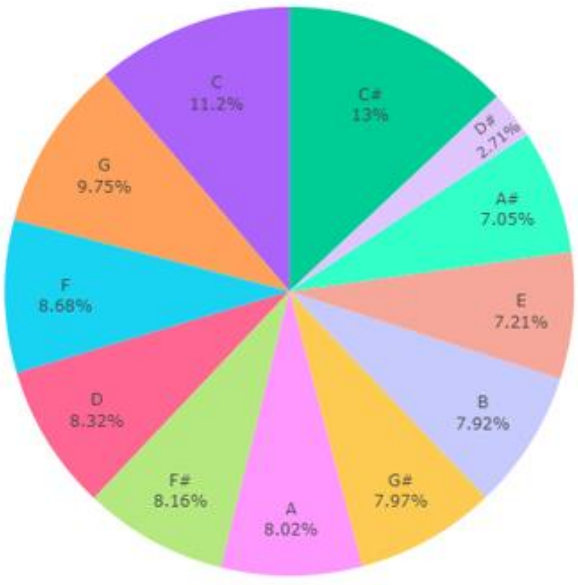
$\chi 2$: 235.32844529966354

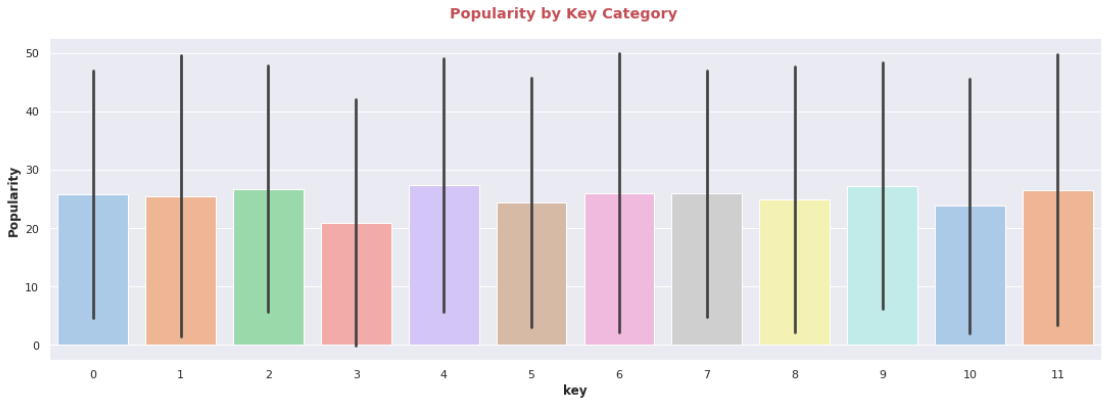
P-value: 9.742906186847722e-51

Null Hypothesis is rejected.

The mode and popularity of songs in the dataset are related.

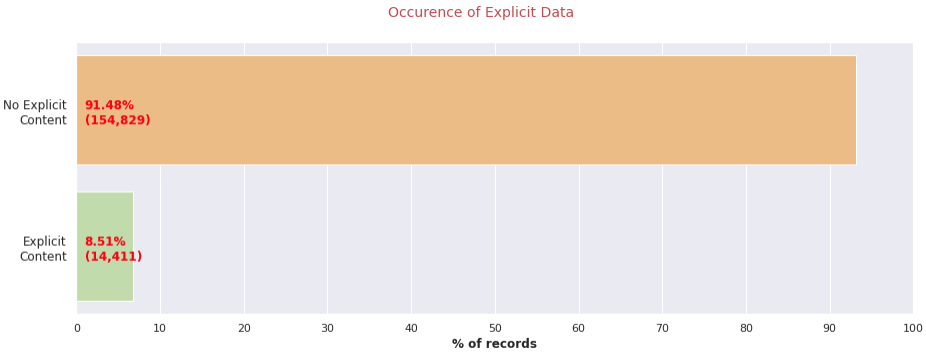
2) Key



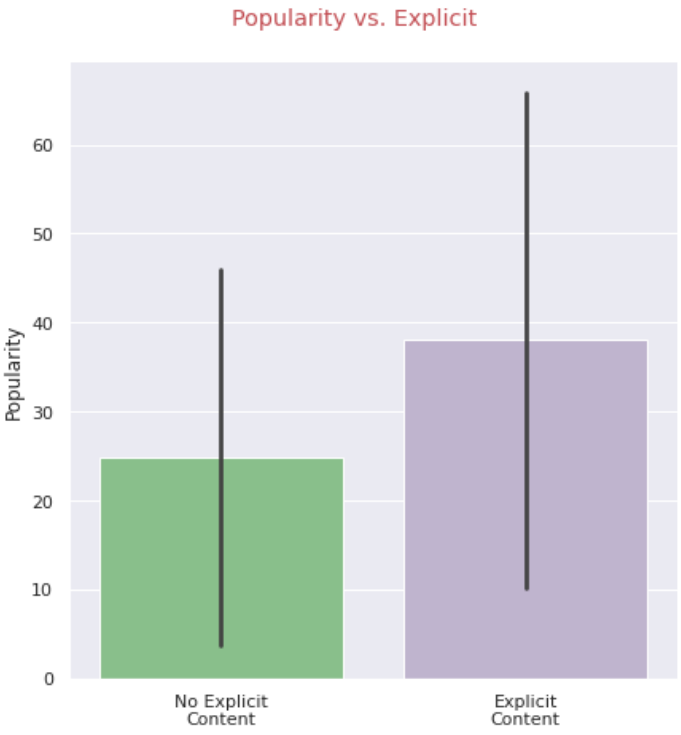


This bar graph shows that even though the popularity of songs in all keys are not the same the difference in mean popularity of songs in different keys is almost insignificant.

3) Explicit Content



The dataset is highly skewed towards songs without explicit content.



Songs with explicit content have a higher mean popularity but this result can also arise because of the inaccuracy of averages in a skewed dataset.

2. Linear Regression using only selected acoustic features

RMSE:

Train = 0.17532
Test = 0.17412

R-squared (uncentered): 0.684
Adj. R-squared (uncentered):0.684
F-statistic: 5.911e+04
AIC: -6.391e+04
Durbin-Watson:2.001

	coef	std err	t	P>|t|	[0.025	0.975]
acousticness	-0.1906	0.002	-91.471	0.000	-0.195	-0.186
energy	-0.0172	0.004	-4.395	0.000	-0.025	-0.010
instrumentalness	-0.0898	0.001	-78.642	0.000	-0.092	-0.088
loudness	0.2272	0.010	23.226	0.000	0.208	0.246
year	0.0002	3.47e-06	45.283	0.000	0.000	0.000

3. Linear Regression – only selected acoustic features and artist information

RMSE:

Train = 0.13327
Test = 0.13902

R-squared (uncentered): 0.838
Adj. R-squared (uncentered): 0.838
F-statistic: 1.180e+05
AIC: -1.555e+05
Durbin-Watson: 2.000

	coef	std err	t	P>|t|	[0.025	0.975]
acousticness	-0.0679	0.002	-44.414	0.000	-0.071	-0.065
artists	0.0092	2.54e-05	361.122	0.000	0.009	0.009
energy	-0.0464	0.003	-16.583	0.000	-0.052	-0.041
instrumentalness	-0.0379	0.001	-45.754	0.000	-0.040	-0.036
loudness	0.0753	0.007	10.748	0.000	0.062	0.089
year	3.565e-05	2.51e-06	14.214	0.000	3.07e-05	4.06e-05

4. Linear Regression – all available information

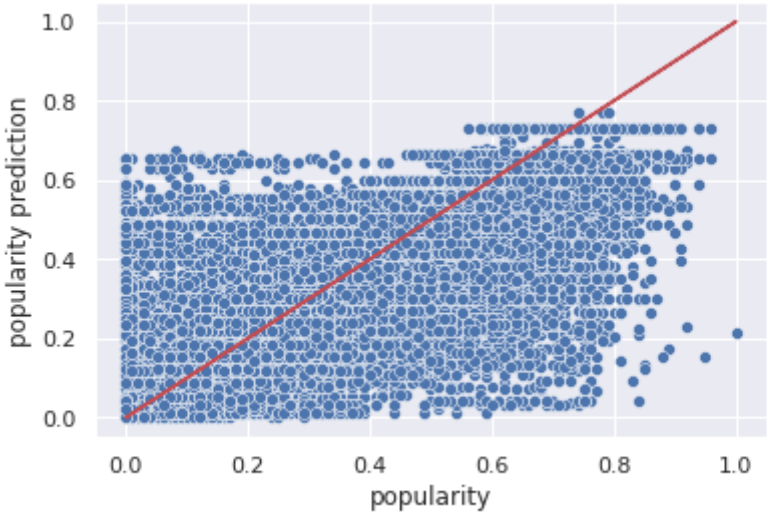
RMSE:

Train = 0.131639
Test = 0.137098

R-squared (uncentered): 0.843
Adj. R-squared (uncentered): 0.843
F-statistic: 2.938e+04
AIC: -1.597e+05
Durbin-Watson: 1.999

	coef	std err	t	P>|t|	[0.025	0.975]
acousticness	-0.0555	0.002	-34.203	0.000	-0.059	-0.052
artists	0.0089	2.64e-05	335.175	0.000	0.009	0.009
danceability	0.0220	0.003	7.997	0.000	0.017	0.027
duration_ms	-0.0637	0.014	-4.718	0.000	-0.090	-0.037
energy	-0.0180	0.003	-6.122	0.000	-0.024	-0.012
explicit	0.0789	0.002	50.218	0.000	0.076	0.082
instrumentalness	-0.0419	0.001	-48.973	0.000	-0.044	-0.040
liveness	-0.0572	0.002	-27.171	0.000	-0.061	-0.053
loudness	-0.0174	0.007	-2.415	0.016	-0.032	-0.003
mode	-0.0014	0.001	-1.633	0.102	-0.003	0.000
speechiness	-0.0777	0.002	-35.407	0.000	-0.082	-0.073
tempo	-0.0122	0.003	-4.507	0.000	-0.018	-0.007
valence	-0.0030	0.002	-1.622	0.105	-0.007	0.001
year	7.288e-05	2.83e-06	25.766	0.000	6.73e-05	7.84e-05
key_1	0.0006	0.002	0.353	0.724	-0.003	0.004
key_2	-0.0021	0.002	-1.403	0.160	-0.005	0.001
key_3	-0.0076	0.002	-3.688	0.000	-0.012	-0.004
key_4	0.0004	0.002	0.239	0.811	-0.003	0.004
key_5	-0.0014	0.002	-0.889	0.374	-0.004	0.002
key_6	-0.0004	0.002	-0.229	0.819	-0.004	0.003
key_7	-0.0053	0.001	-3.633	0.000	-0.008	-0.002
key_8	0.0009	0.002	0.525	0.600	-0.003	0.004
key_9	-0.0039	0.002	-2.555	0.011	-0.007	-0.001
key_10	-0.0055	0.002	-3.232	0.001	-0.009	-0.002
key_11	-0.0054	0.002	-3.002	0.003	-0.009	-0.002

5. Decision Tree – only selected acoustic features and artist information



RMSE:

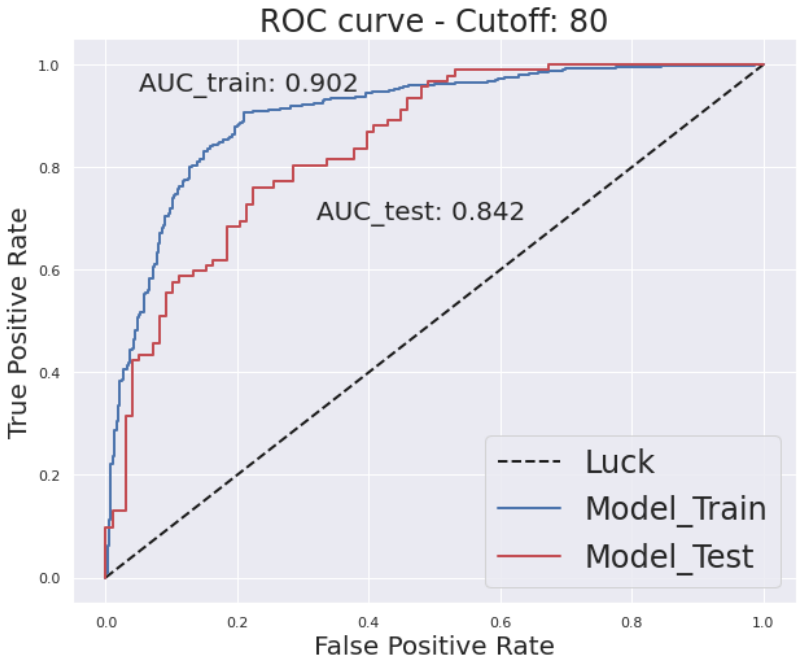
Train: 0.105
Test: 0.114

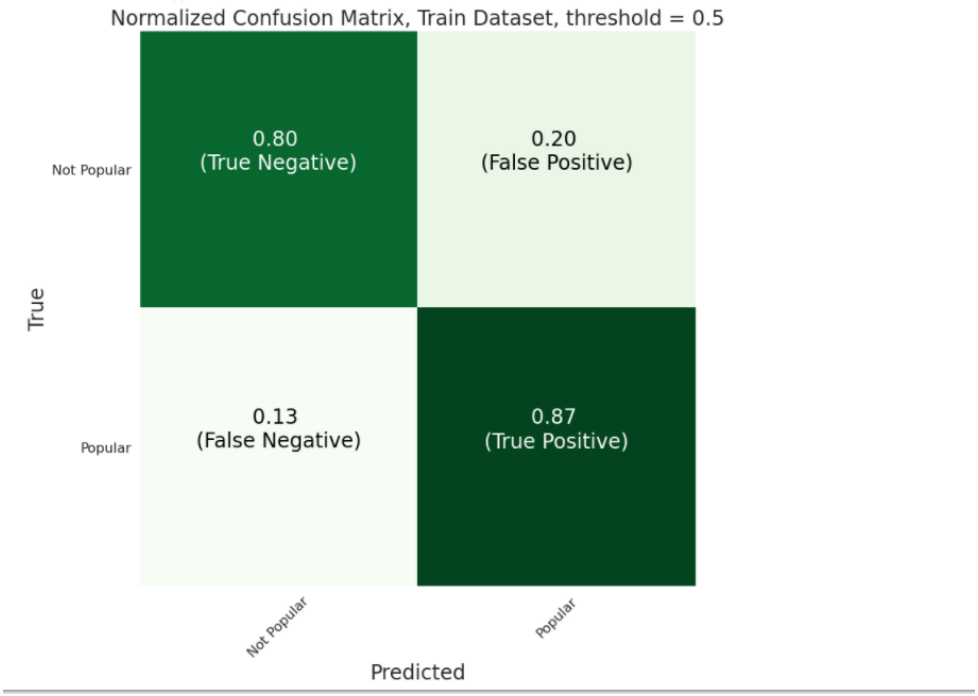
R-squared (uncentered): 0.838
Adj. R-squared (uncentered): 0.838
F-statistic: 1.180e+05
AIC: -1.555e+05
Durbin-Watson: 2.000

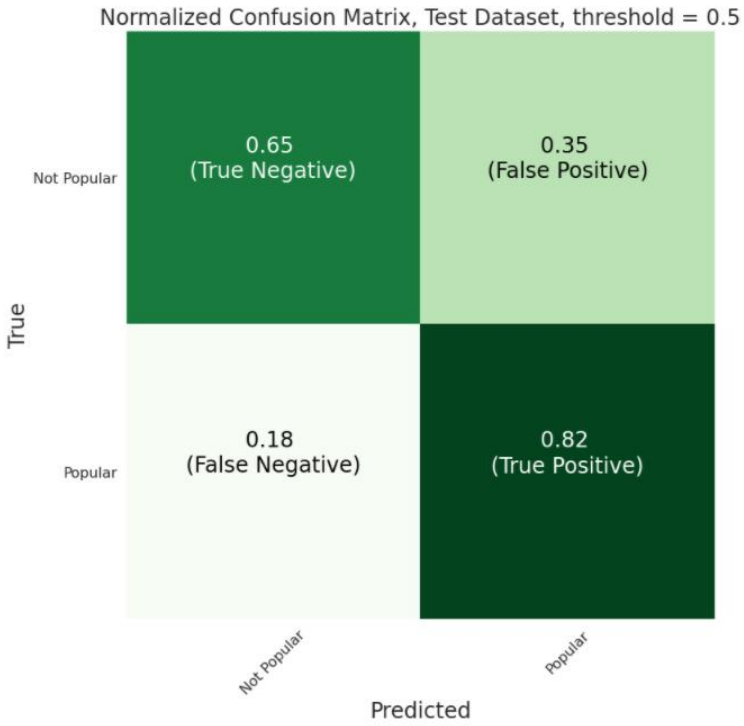
	coef	std err	t	P>|t|	[0.025	0.975]
acousticness	-0.0679	0.002	-44.414	0.000	-0.071	-0.065
artists	0.0092	2.54e-05	361.122	0.000	0.009	0.009
energy	-0.0464	0.003	-16.583	0.000	-0.052	-0.041
instrumentalness	-0.0379	0.001	-45.754	0.000	-0.040	-0.036
loudness	0.0753	0.007	10.748	0.000	0.062	0.089
year	3.565e-05	2.51e-06	14.214	0.000	3.07e-05	4.06e-05

6. Logistic Regression – all available information

Train accuracy: 0.84
Test accuracy: 0.73
Train recall: 0.87
Test recall: 0.82
Train precision: 0.82
Test precision: 0.69
Train AUC: 0.902
Test AUC: 0.842





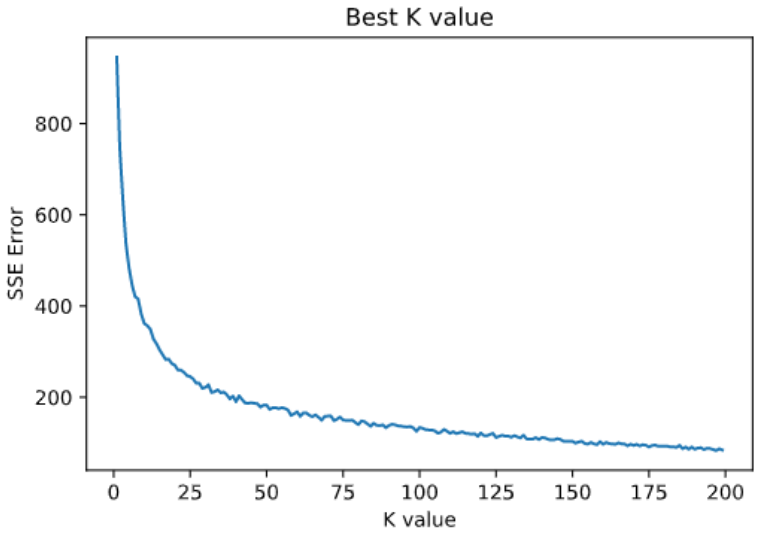


Delta in p-Value: -0.2482833

7. **Clustering-then-Regression: K-Means Clustering and Decision Tree using only selected acoustic features and artist information**

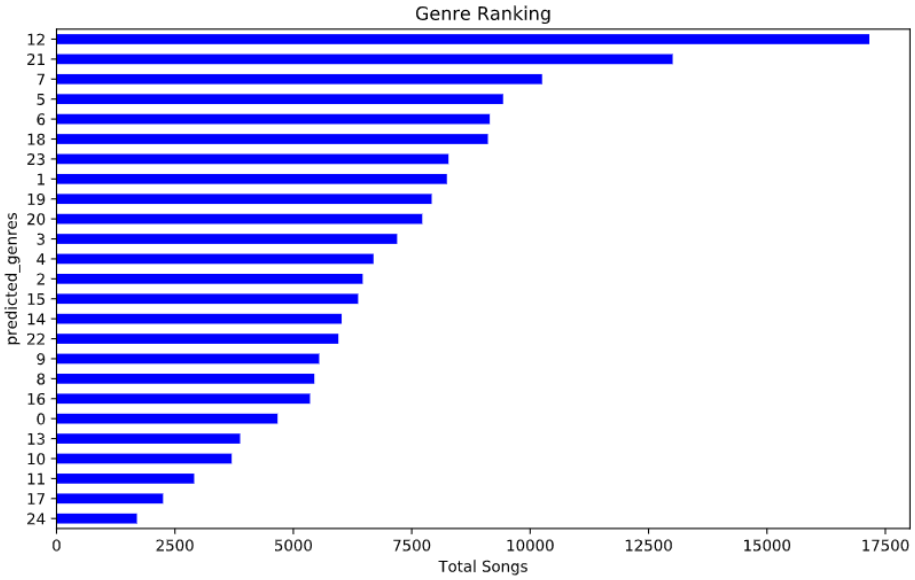
K-Means Clustering

Choosing k :

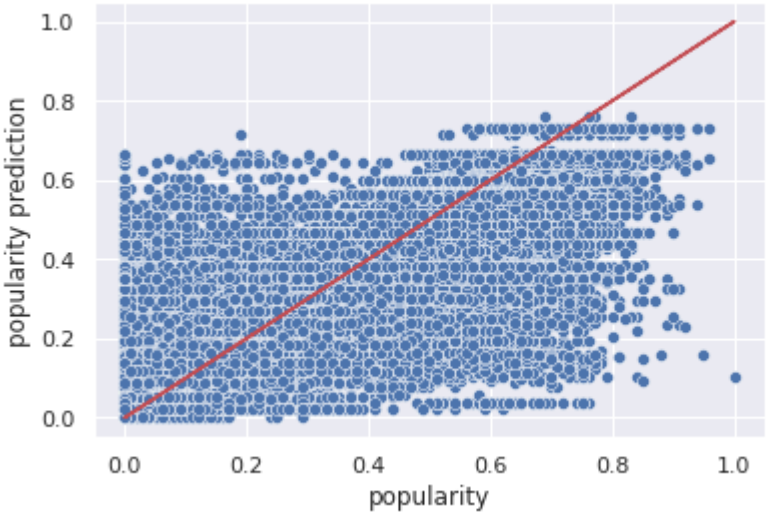


Using the “elbow ” method the k-value was chosen as 25.

Clustering Result :



Decision Tree on modified Dataset :



RMSE:

Train: 0.10421681427960094
Test: 0.11232989759946345

R-squared (uncentered): 0.838
Adj. R-squared (uncentered): 0.838
F-statistic: 1.012e+05
AIC: -1.556e+05
Durbin-Watson: 2.001

	coef	std err	t	P>|t|	[0.025	0.975]
acousticness	-0.0664	0.002	-43.273	0.000	-0.069	-0.063
artists	0.0092	2.53e-05	362.106	0.000	0.009	0.009
energy	-0.0459	0.003	-16.368	0.000	-0.051	-0.040
instrumentalness	-0.0537	0.001	-43.816	0.000	-0.056	-0.051
loudness	0.0771	0.007	10.852	0.000	0.063	0.091
year	1.064e-05	2.45e-06	4.352	0.000	5.85e-06	1.54e-05
genre	0.0007	5.36e-05	13.244	0.000	0.001	0.001

Conclusion

In this project a methodology to predict the popularity of a song using data collected from Spotify was presented.

The best results were achieved using a decision tree using only the highly corelated features, artist data and the genre data obtained using clustering. The logistic regression algorithm performed very well for the case of binary classification with a training AUC of 0.90 and a test AUC of 0.84.

With the available dataset currently a logistic regression model performs much better than any linear regression model.

Even though the project is limited to data on Spotify, the models can be extended to other platforms.

Furthermore, while this project explored models that use only acoustic information, historical data about the previously popular songs and other information about the current social media trends can help greatly in improving the accuracy of the models.

References

[1] D. Herremans, D. Martens, and K. Sorensen, “Dance hit song prediction,” Journal of New Music Research, vol. 43, no. 3, pp. 291–302, 2014..881888

[2] V. Dhar and E. A. Chang, “Does chatter matter? the impact of user-generated content on music sales,” Journal of Interactive Marketing, vol. 23, no. 4, pp. 300 – 307, 2009.

[3] J. Lee and J. Lee, “Music popularity: Metrics, characteristics, and audio based prediction,” IEEE Transactions on Multimedia, vol. 20, no. 11, pp. 3173–3182, Nov 2018.

[4] https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks