Popularity Prediction of Spotify Tracks

Can we predict track popularity with relevant attribute information?

Context

Spotify has changed how the music industry uses data.

There are several attributes Spotify uses to quantify tracks.

- Acousticness
- Danceability
- Energy
- Instrumentalness
- Liveness
- Loudness
- Speechiness
- Tempo
- Valence
- Popularity

Spotify has increased their revenue by 18-20% since 2018. How can we use this data to predict popularity and improve investment strategy?

Results and Conclusions

The attributes can be used to predict popularity with an average 83% accuracy.

The model was fit using all the data except the targeted tracks.

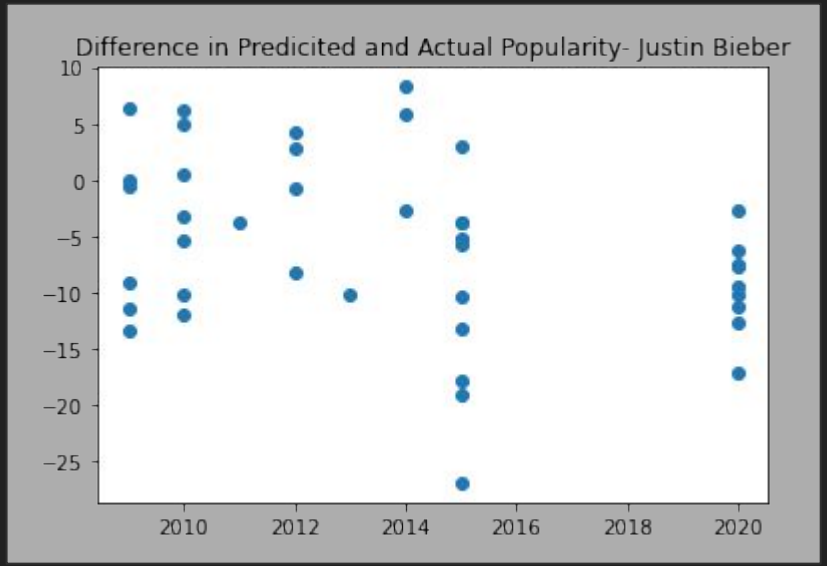
Release date of the track did not have an effect on the average accuracy of the prediction.

Popularity, as defined by the spotify metric, is relatively predictable based on the attributes.

Linear Regression had an average 80% accuracy and was much faster at fitting and predicting.

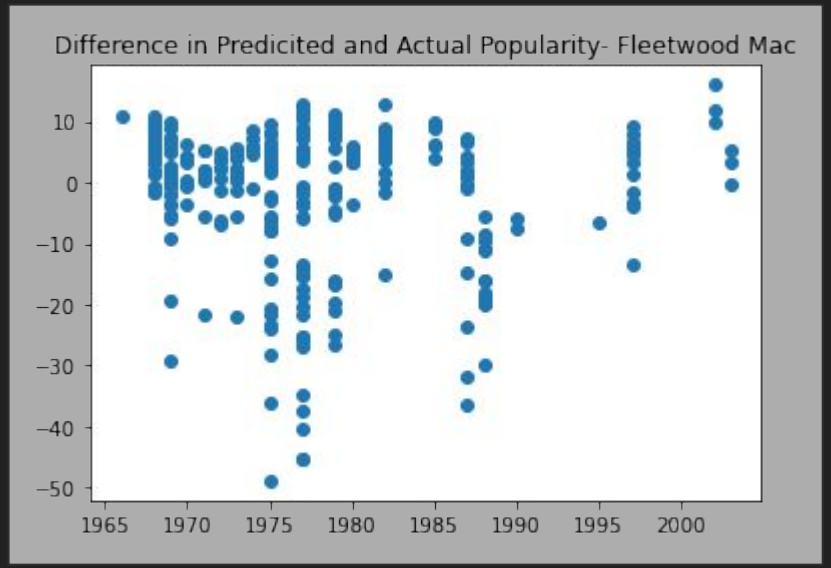
Comparison of Model Prediction Popularity vs Actual Popularity of Tracks by Justin Bieber

The model was used to predict popularity of tracks by Justin Bieber. The difference in predicted popularity vs reported popularity was plotted by release year.



Comparison of Model Prediction Popularity vs Actual Popularity of Tracks by Fleetwood Mac

The model was used to predict popularity of tracks by Fleetwood Mac. The difference in predicted popularity vs reported popularity was plotted by release year.



Data Wrangling

Dataset from Kaggle, contained Spotify tracks and attributes from 1921-2020.

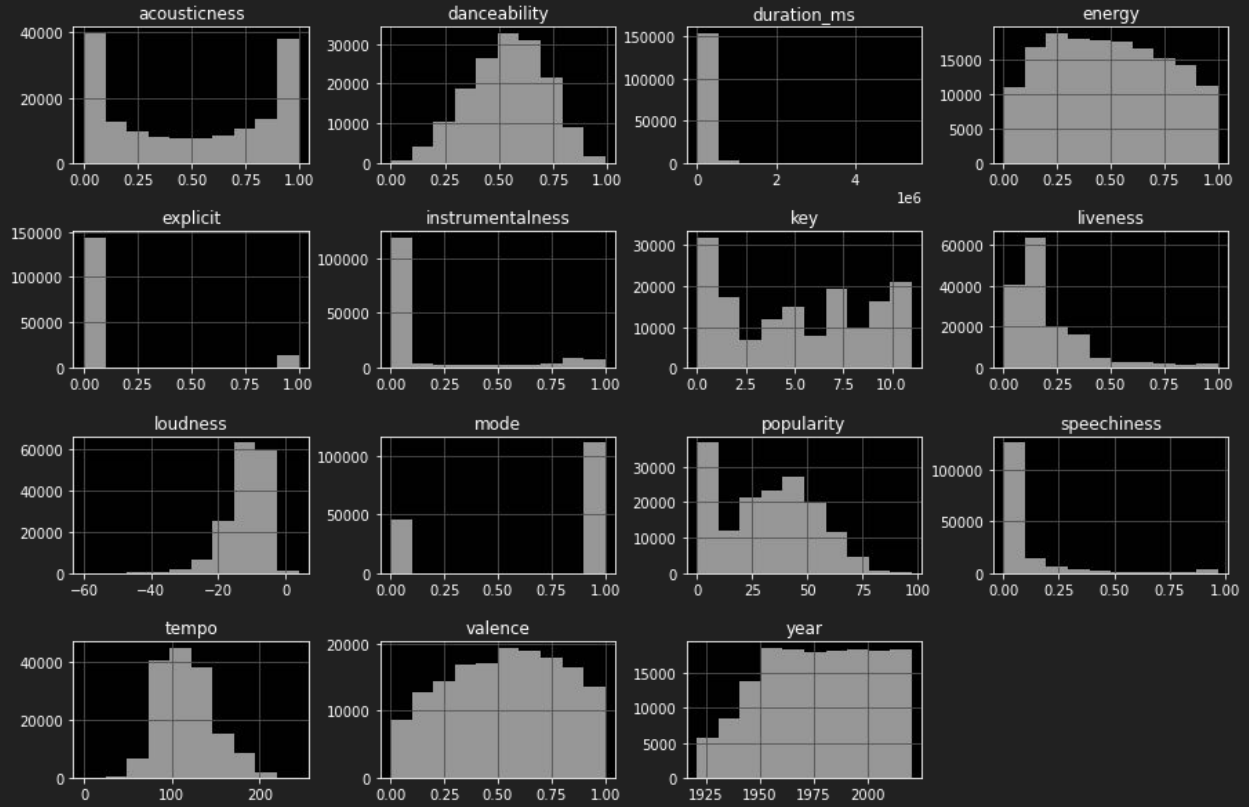
Relatively clean dataset, no missing values

Duplications - Artists sometimes re-release the same track on different albums or as a single

- These were removed using both track name and artist name

Columns such as SpotifyID were dropped

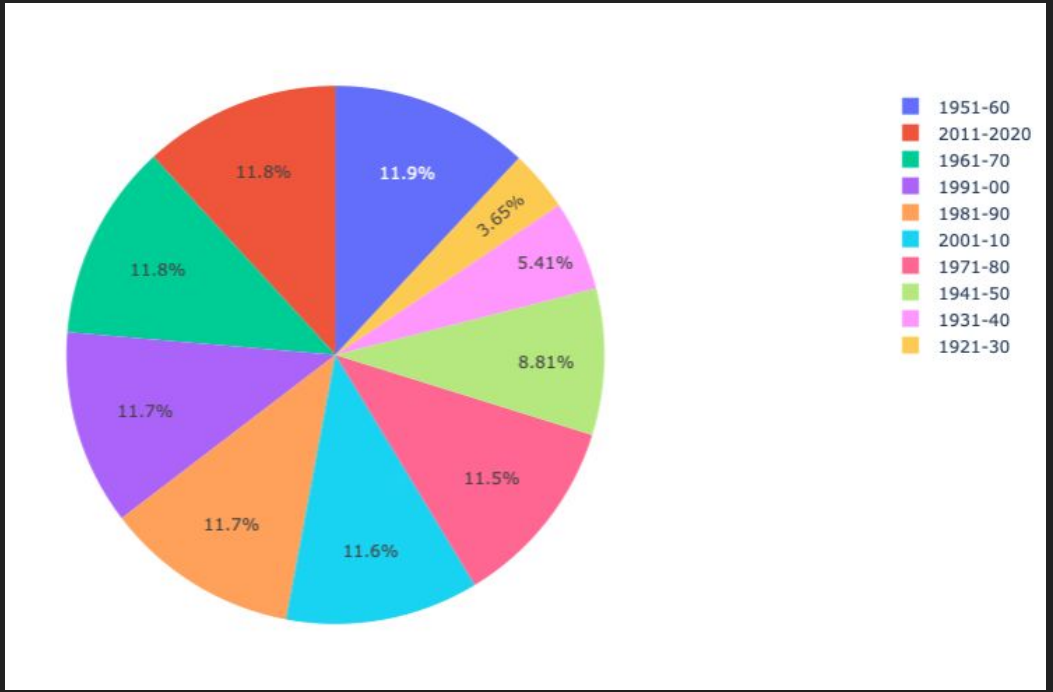
Attribute Distribution



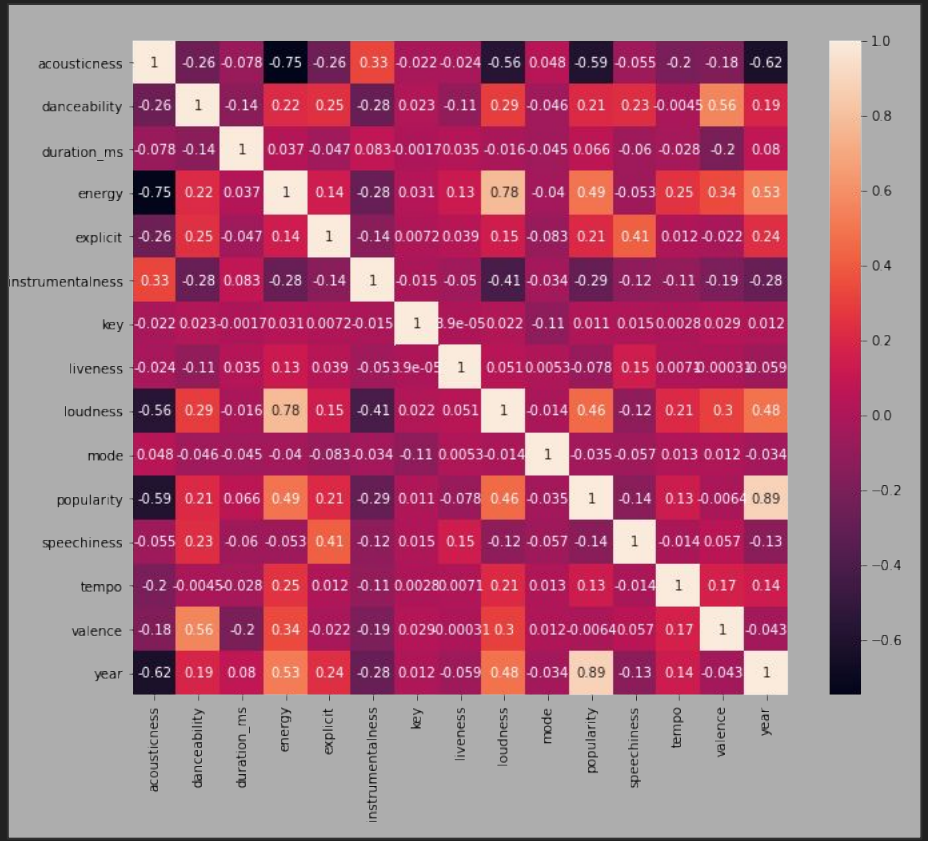
The attributes were plotted using df.hist() to visualize the distribution of the features.

Track Distribution by Decade

The tracks were grouped by decade to view the proportions of each tracc by decade.



Correlation



sns.heatmap() was used to view the correlations between attributes

Year has the highest correlation due to popularity weighting recent plays.

danceability, energy, loudness, and tempo have the highest positive correlations

acousticness, speechiness, and instrumentalness have the highest negative correlations

People seem to like high energy, up beat tracks but not more mellow acoustic/instrumental tracks.

Model Comparison

Compared Linear Regression and Random Forest Regression models and then optimized each using GridSearchCV

train_test_split() was used to split the data for training

Test_size = 0.3, random state = 42

Tested using StandardScaler() vs not and n_estimators for Random Forest Regression.

Random Forest was 3% more accurate but the time for training the model was prohibitive so Linear Regression was used.

Further Analysis Possible

Can we use different popularity metric (one that doesn’t favor recent tracks) to predict the next trends in music?

Podcasts have boomed in popularity:

How does the prediction work if we separate music tracks from podcasts?

Can we predict popularity of podcasts from current attributes? What additional/replacement attributes would be helpful?