Report for Spotify (Capstone 1)

8/4/2022
Ashley Mersman

Background and Objective

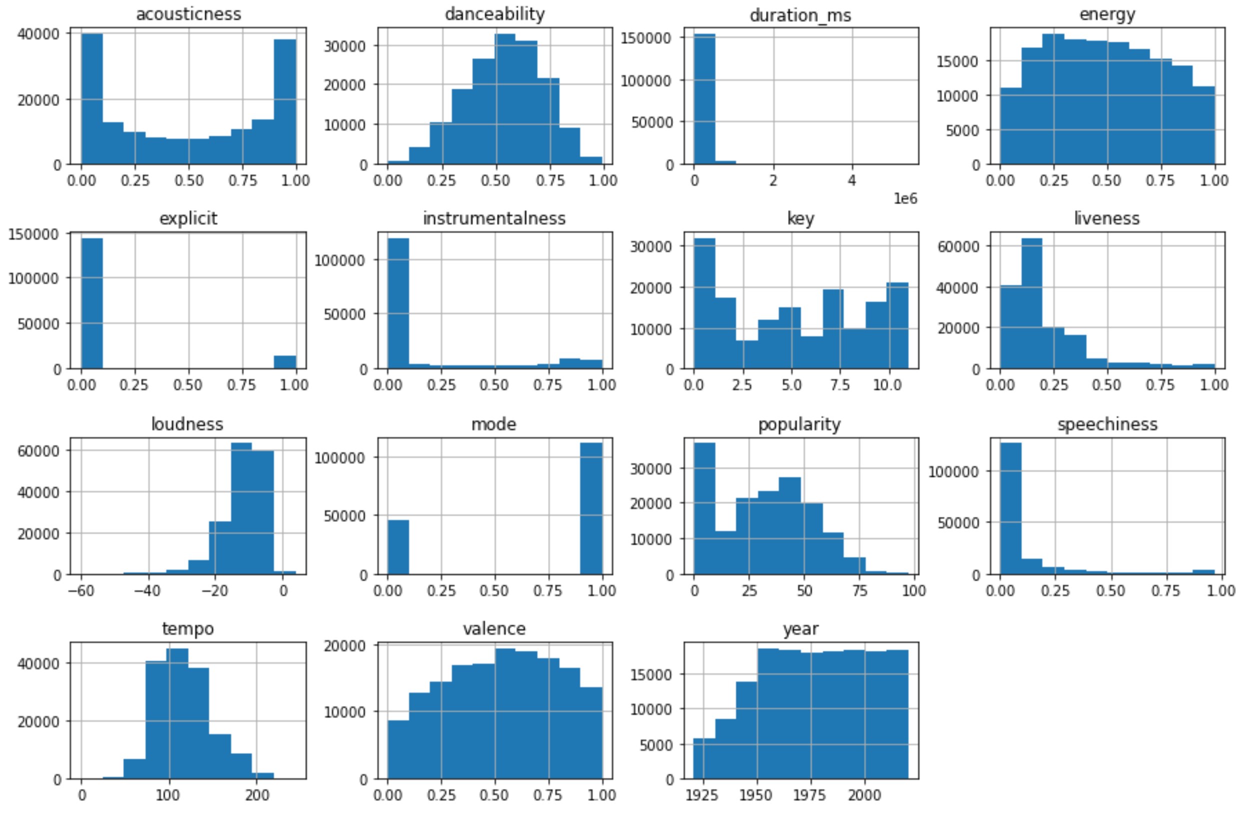
Spotify has changed how the music industry quantifies music. Spotify uses several features to quantify music; acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, and popularity.

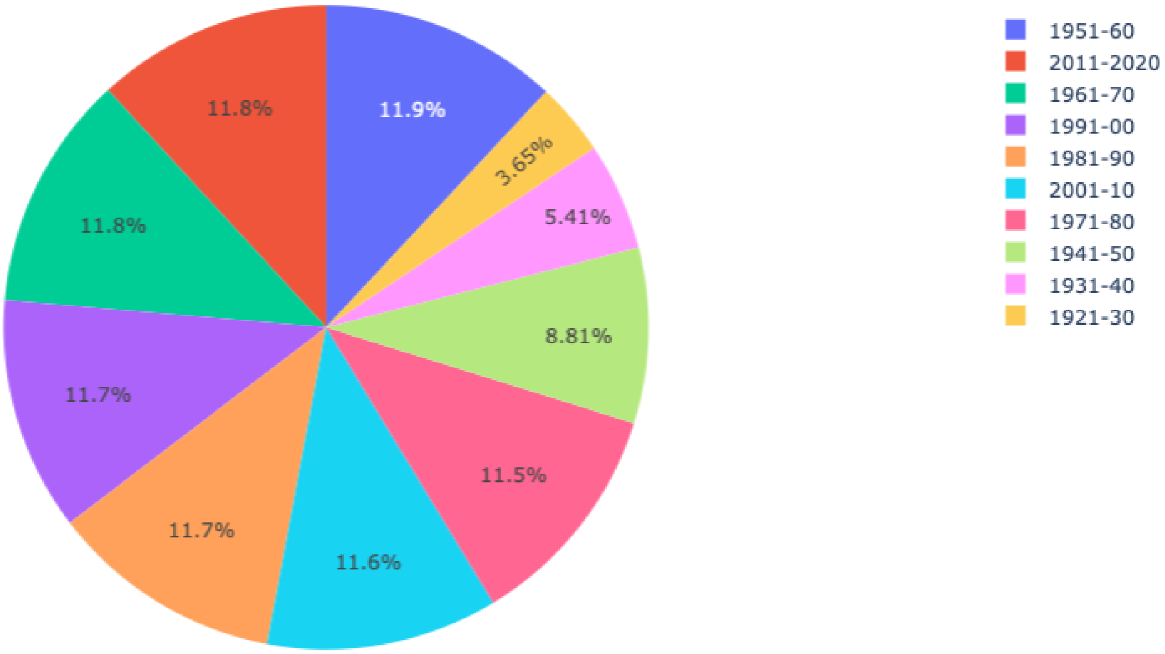
Acousticness	0-1.0 scale; 1 high confidence of acoustic track
Danceability	0-1.0 scale; 1 most danceable combination of elements including tempo, rhythm stability, beat strength, and regularity.
Energy	0-1.0 scale; 1 most energetic Represents intensity and activity using elements including dynamic range, timbre, entropy
Instrumentalness	0-1.0 scale; 0.5 intended to represent instrumental tracks increasing in confidence to 1.0.
Liveness	Over 0.8 indicates strong likelihood that the track is live; detects presence of audience in recording
Loudness	decibels
Speechiness	0-1.0 scale; closer to 1.0 is more likely to be speech (podcast, news, audiobook…), below 0.33 more likely to represent music.
Tempo	beats per minute
Valence	0-1.0 scale; indicative of musical positiveness Higher values are happy, cheerful, euphoric Lower values are sad, angry, depressed
Popularity	0-100, higher values are more popular Uses number of plays and how recent those plays are

Spotify has increased their annual revenue by 18-20% since 2018. They have revolutionized the use of data in the music industry and music streaming. Can we use these attributes to predict popularity of tracks?

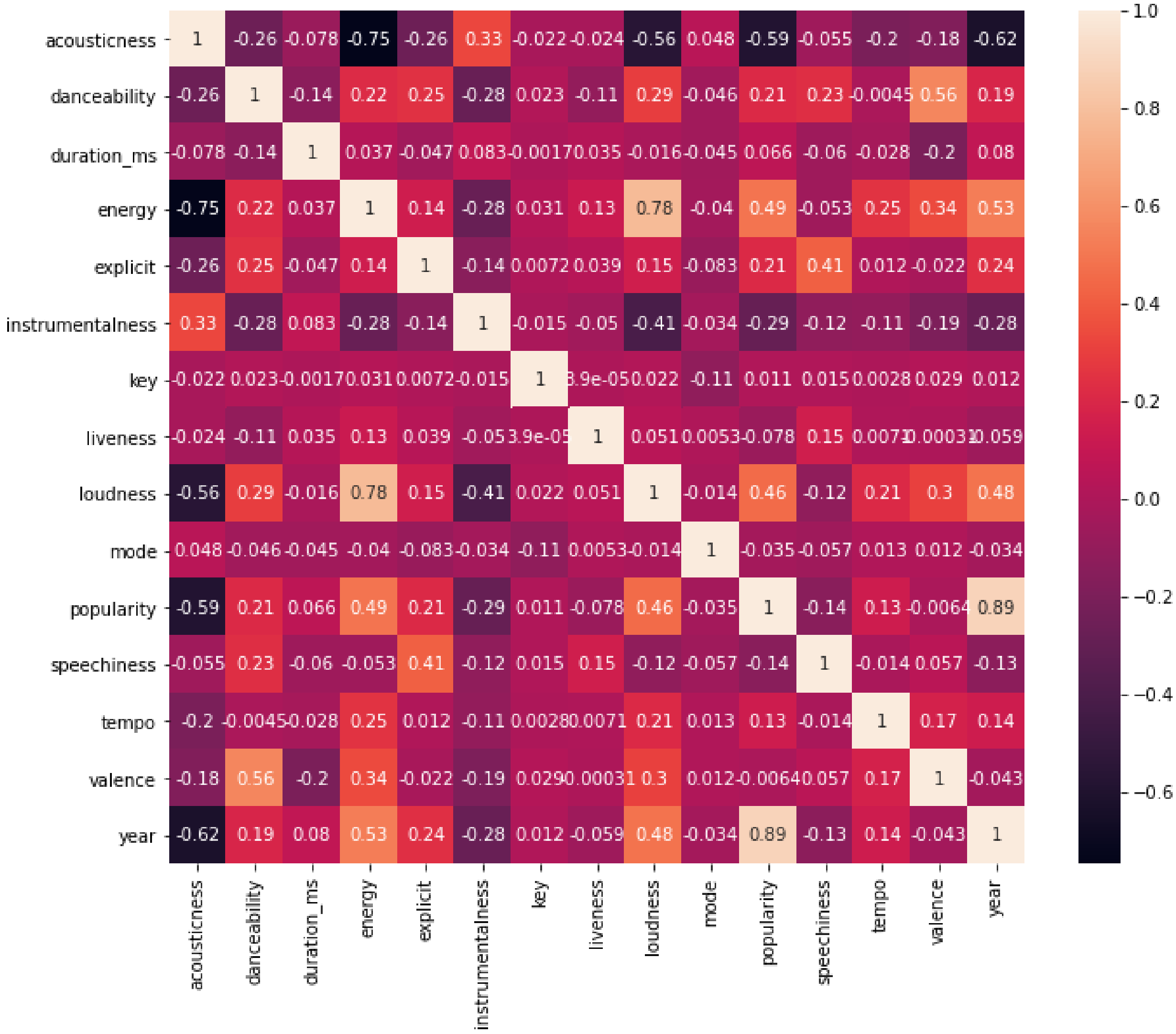
Data Wrangling, Exploratory Data Analysis, and Model Selection

The data used for analysis and to train and test the model was found on Kaggle and included Spotify data from tracks released 1921- 2020. The data set was relatively clean with no missing values. There were some duplications from artists re-releasing the same track on different albums or as single versions. These were filtered out and columns that were not relevant to the analysis, such as Spotify ID, were dropped.





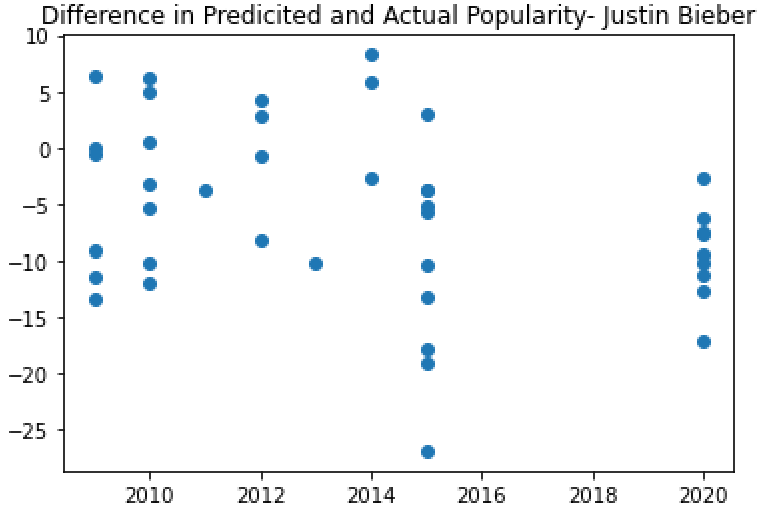
I used a heatmap to find the features with the highest correlation with ‘popularity’. Year was highly correlated, which makes sense due to Spotify's popularity feature prioritizing recent plays. Loudness is relatively high in correlation, as well as energy and acousticness has a negative correlation with a high value.
People seem to prefer tracks that are energetic, happy, and loud.

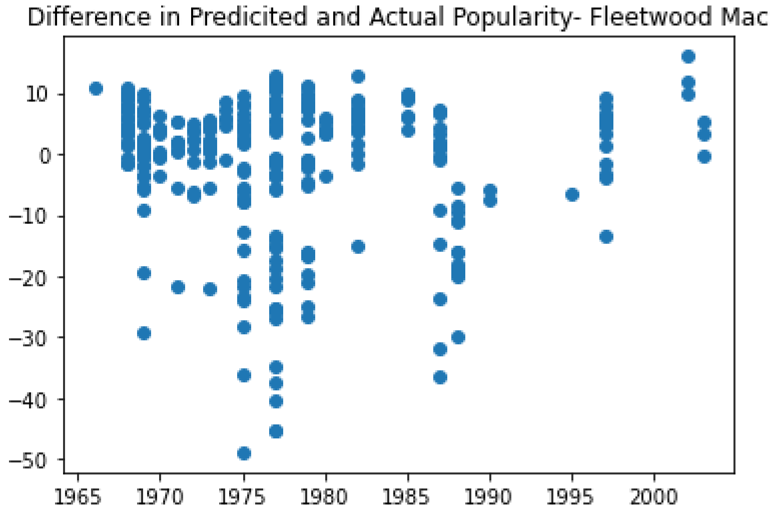


I started with a 70/30 training/testing split. I then scaled the data using StandardScaler. I tested both a linear regression model and a random forest model with various hyperparameter functions. Hyperparameter grid_search found the best linear model was not using scaled data. The model with the best accuracy was a random forest regressor with n_estimators of 124. However the time it took to fit the model to fit and predict was a huge hurdle for only a 3% improvement over the much faster linear regression.

Modeling and Analysis

I used the model to predict popularity and check accuracy on both a recent artist, Justin Bieber, and an artist from the 70’s, Fleetwood Mac. For each occurrence I fit the model using all the data except that of the artist, did a 5-fold cross-validation and used the average. The difference in release date did not make a difference on the accuracy of the popularity prediction. 0.77440 for Justin Bieber vs 0.7745 for Fleetwood Mac. I plotted the difference between the actual popularity and the predicted popularity for both artists.


Difference in Predicited and Actual Popularity- Justin Bieber


Difference in Predicited and Actual Popularity- Fleetwood Mac

Further Analysis and Future Projects

The popularity was based on Spotify’s own metric which skews toward more recent plays. What is the correlation between this metric of popularity and others, for instance the Billboard top 100 for each year? How have the trends on popular music changed over the past 100 years and can we predict the trends in these attributes for the upcoming decade? These answers could help us invest in new artists.

With the new popularity and boom in podcasts we could also apply these same questions to that media and use that to set investment strategies.