Report for Spotify (Capstone 1)

8/4/2022

Ashley Mersman

Background and Objective

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. With streaming services as the top way people listen to music, it's important that customers are hearing the tracks they want for customer retention and attraction. What trends are we seeing in track popularity and what features are most important in deciding what tracks to push to listeners? Can we predict song popularity to drive revenue and retention?

Data

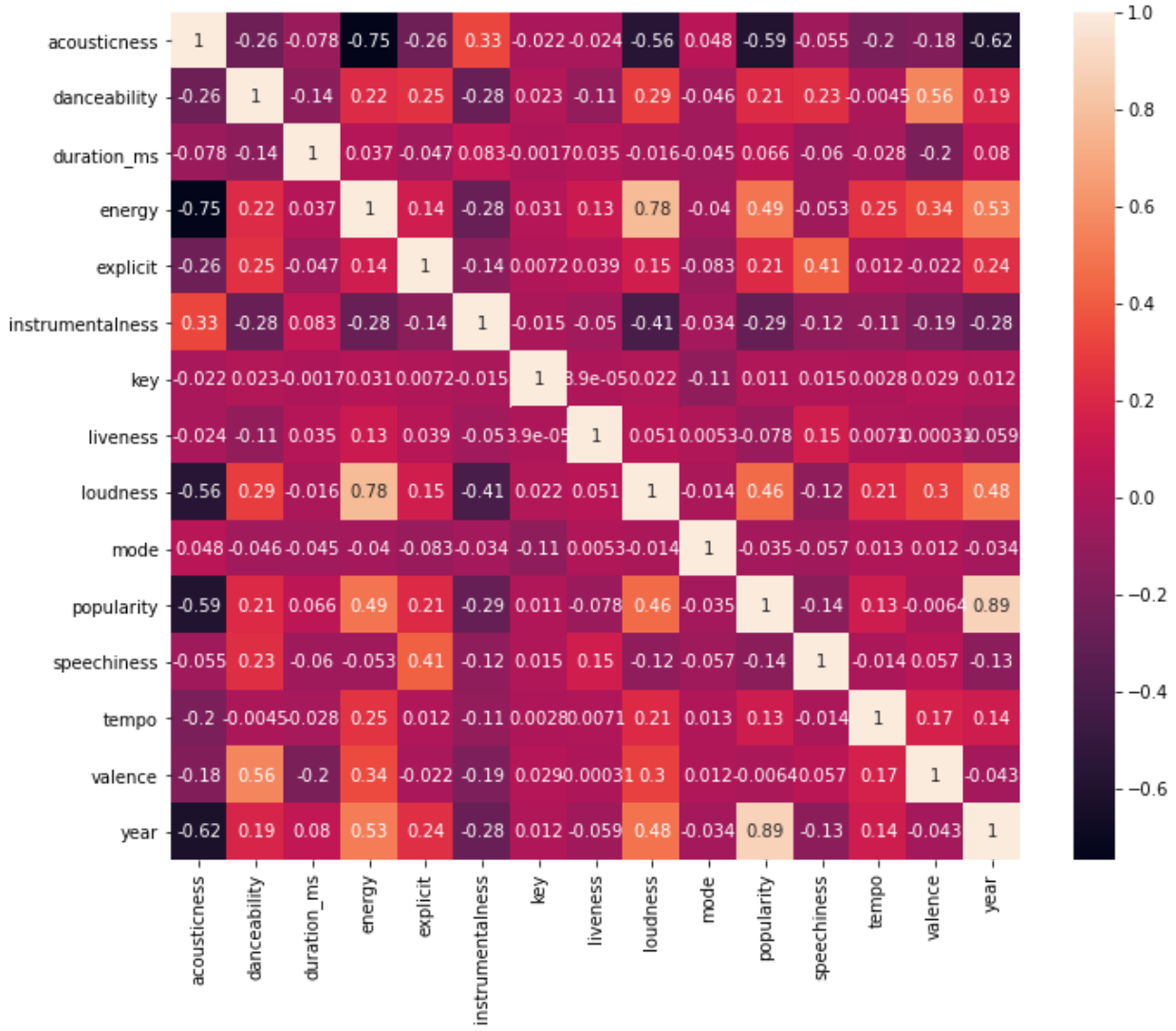
The data used for analysis and to train and test the model were found on Kaggle and included Spotify data from tracks released 1921- 2020. The data set was relatively clean with no missing values. There were some duplications from artists re-releasing the same track on different albums or as single versions.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

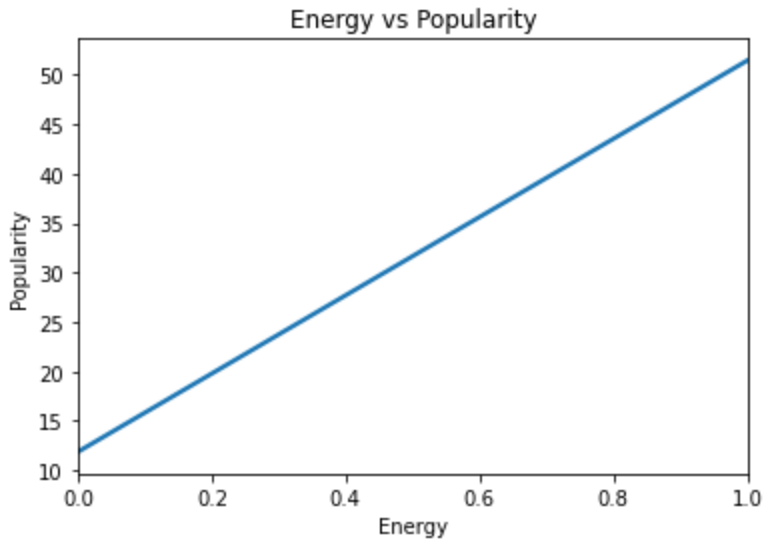
Spotify uses several features to quantify music; acousticness, danceability, duration, energy, instrumentalness, liveness, loudness, speechiness, tempo, valence, and popularity.

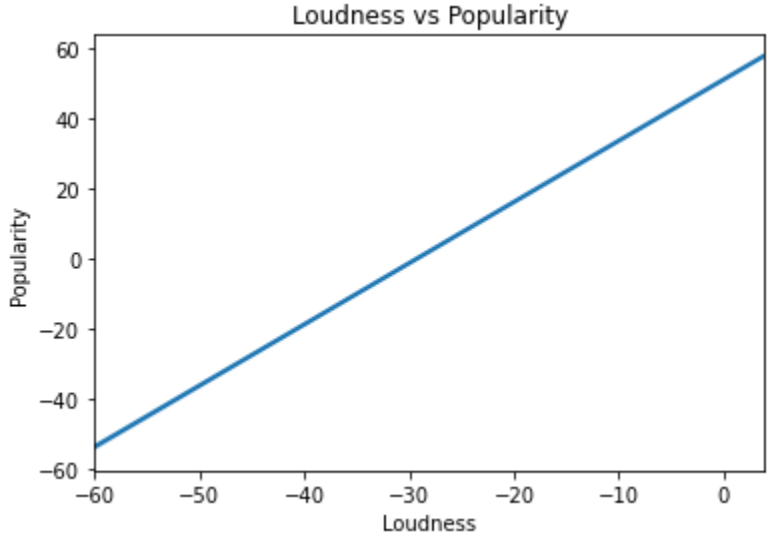
Exploratory Data Analysis

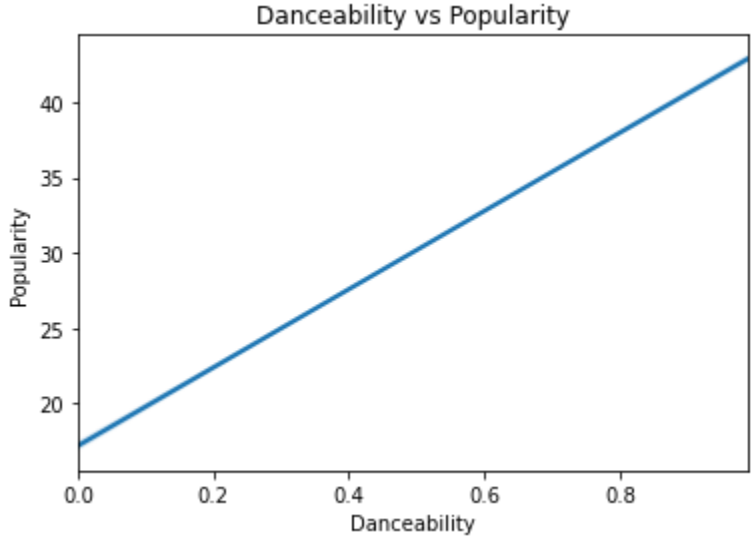
{Look at Popularity, Popularity vs Year, Popularity vs Energy, Popularity vs Acousticness, Popularity vs Loudness and include charts and analysis}

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.

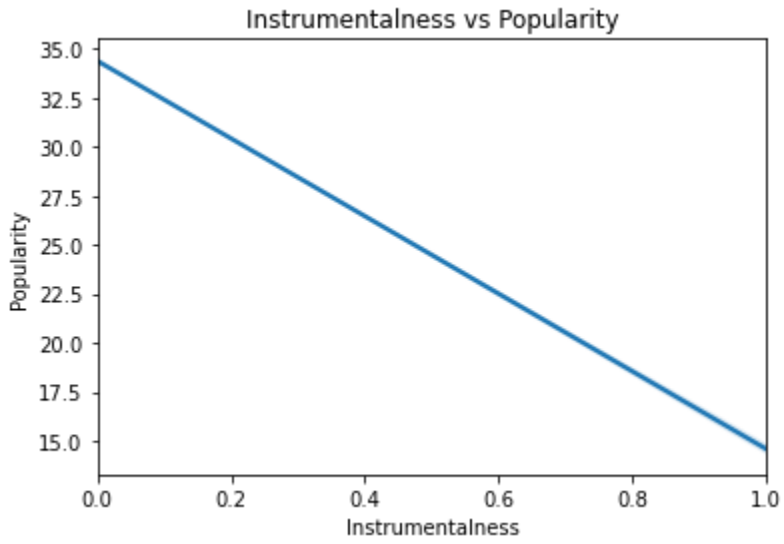

acousticness 1 -0.26 -0.078 -0.75 -0.26 0.33 -0.022 -0.024 -0.56 0.048 -0.59 -0.055 -0.2 -0.18 -0.62
danceability -0.26 1 -0.14 0.22 0.25 -0.28 0.023 -0.11 0.29 -0.046 0.21 0.23 -0.0045 0.56 0.19
duration_ms -0.078 -0.14 1 0.037 -0.047 0.083 -0.0017 0.035 -0.016 -0.045 0.066 -0.06 -0.028 -0.2 0.08
energy -0.75 0.22 0.037 1 0.14 -0.28 0.031 0.13 0.78 -0.04 0.49 -0.053 0.25 0.34 0.53
explicit -0.26 0.25 -0.047 0.14 1 -0.14 0.0072 0.039 0.15 -0.083 0.21 0.41 0.012 -0.022 0.24
instrumentalness 0.33 -0.28 0.083 -0.28 -0.14 1 -0.015 -0.05 -0.41 -0.034 -0.29 -0.12 -0.11 -0.19 -0.28
key -0.022 0.023 -0.0017 0.031 0.0072 -0.015 1 3.9e-05 0.022 -0.11 0.011 0.015 0.0028 0.029 0.012
liveness -0.024 -0.11 0.035 0.13 0.039 -0.053 3.9e-05 1 0.051 0.0053 -0.078 0.15 0.0071 -0.0003 -0.059
loudness -0.56 0.29 -0.016 0.78 0.15 -0.41 0.022 0.051 1 -0.014 0.46 -0.12 0.21 0.3 0.48
mode 0.048 -0.046 -0.045 -0.04 -0.083 -0.034 -0.11 0.0053 -0.014 1 -0.035 -0.057 0.013 0.012 -0.034
popularity -0.59 0.21 0.066 0.49 0.21 -0.29 0.011 -0.078 0.46 -0.035 1 -0.14 0.13 -0.0064 0.89
speechiness -0.055 0.23 -0.06 -0.053 0.41 -0.12 0.015 0.15 -0.12 -0.057 -0.14 1 -0.014 0.057 -0.13
tempo -0.2 -0.0045 -0.028 0.25 0.012 -0.11 0.0028 0.0071 0.21 0.013 0.13 -0.014 1 0.17 0.14
valence -0.18 0.56 -0.2 0.34 -0.022 -0.19 0.029 -0.0003 0.3 0.012 -0.0064 0.057 0.17 1 -0.043
year -0.62 0.19 0.08 0.53 0.24 -0.28 0.012 -0.059 0.48 -0.034 0.89 -0.13 0.14 -0.043 1
acousticness
danceability
duration_ms
energy
explicit
instrumentalness
key
liveness
loudness
mode
popularity
speechiness
tempo
valence
year
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

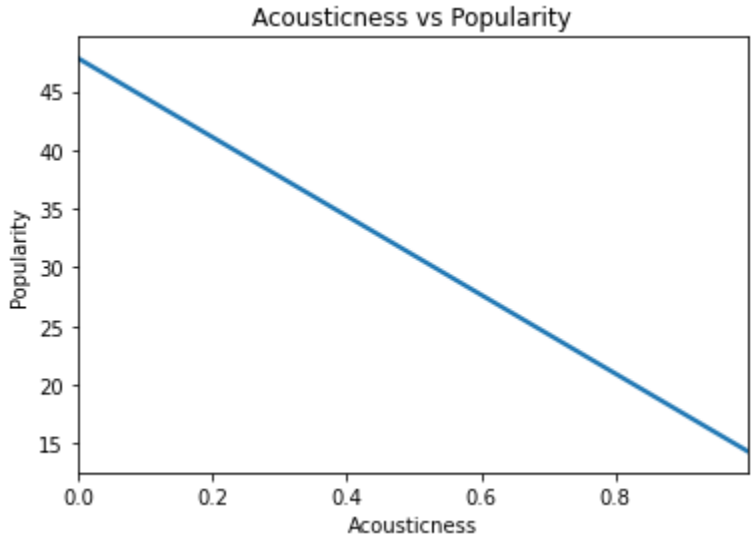

Energy vs Popularity
Popularity
Energy
50
45
40
35
30
25
20
15
10
0.0
0.2
0.4
0.6
0.8
1.0


Loudness vs Popularity
Popularity
Loudness
60
40
20
0
−20
−40
−60
−60
−50
−40
−30
−20
−10
0

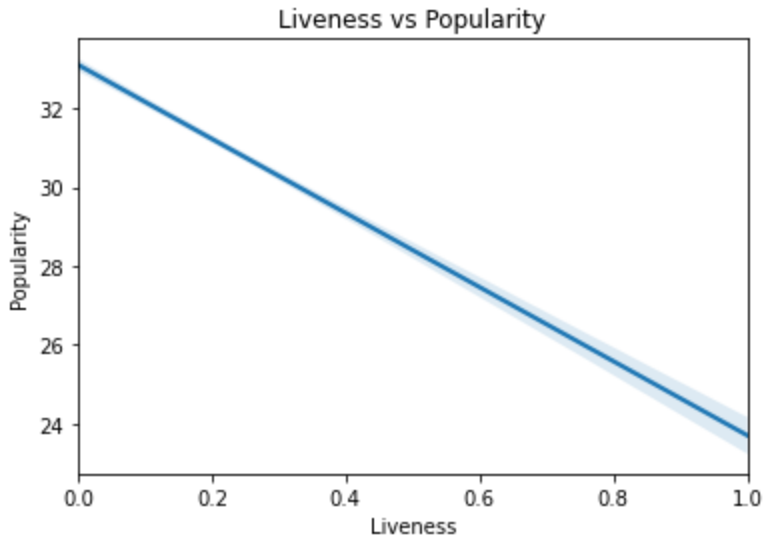

Danceability vs Popularity
Popularity
Danceability
40
35
30
25
20
0.0
0.2
0.4
0.6
0.8

People seem to prefer tracks that are energetic and loud, tracks that have been marked as easy to dance to.



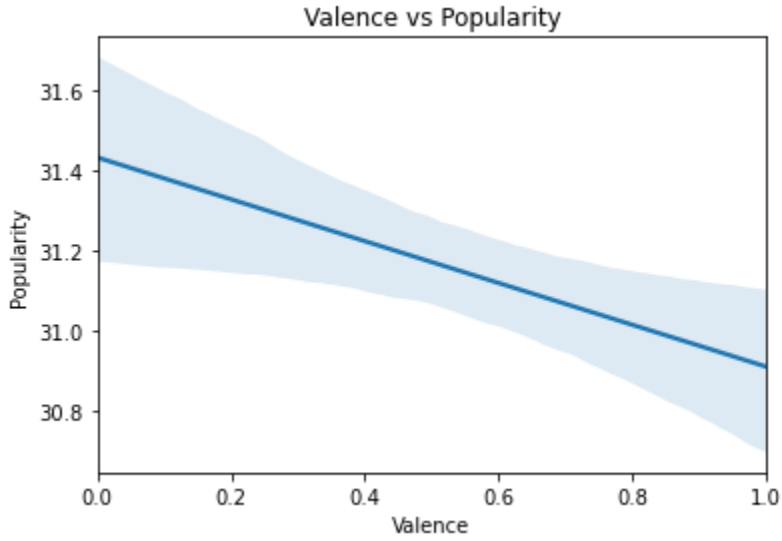


People don't seem to prefer instrumental tracks or acoustic tracks. These have a much more mellow sound so this makes sense when we know that energy and danceability had a positive correlation.



Tracks with a likelihood of being a “live” track fall into these categories and therefore have a negative correlation.

An unexpected correlation was valence. Increased measures of happiness (as defined by Spotify) was inversely related to popularity.



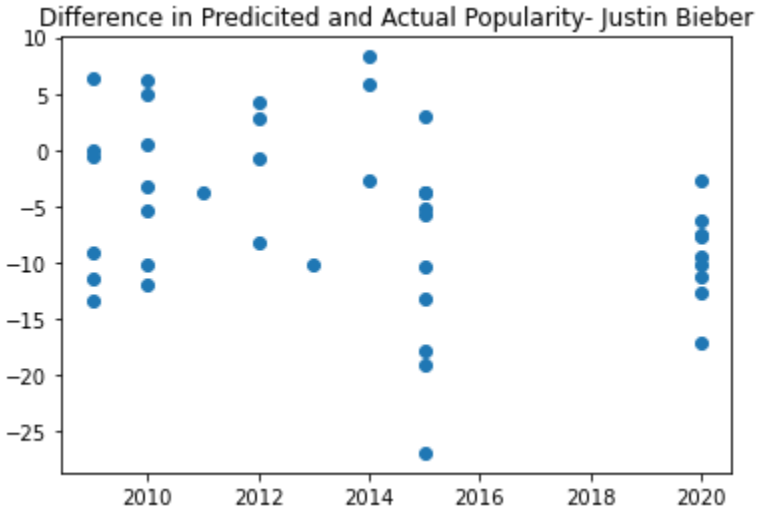
Preprocessing and Model Selection

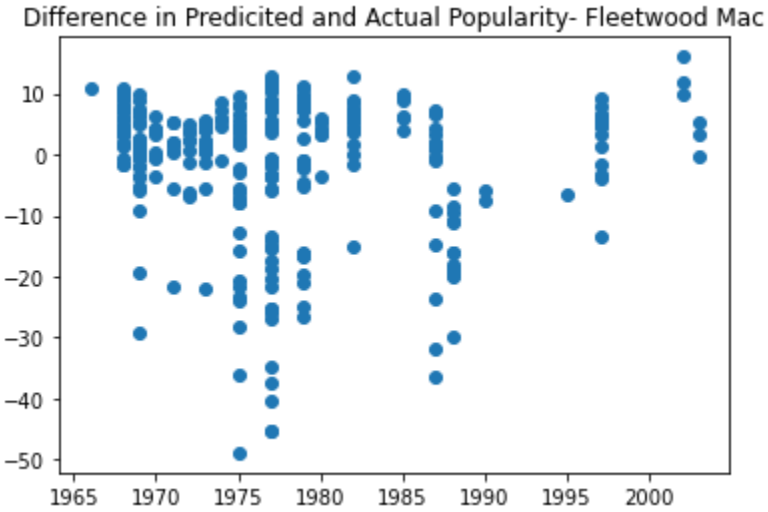
I started with a 70/30 training/testing split. I then scaled the data using StandardScaler. I tested both a linear regression model and a random forest model with various hyperparameter functions. Hyperparameter grid_search found the best linear model was not using scaled data. The model with the best 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 4% improvement over the much faster linear regression.

Model	Best Parameters	CV Score
Linear Regression	StandardScaler: None	0.794
Decision Tree Regressor		0.650
Random Forest Regressor	N_estimators: 195 StandardScaler: Standard Scaler	0.833

Model 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. 0.774 for Justin Bieber vs 0.775 for Fleetwood Mac. I plotted the difference between the actual popularity and the predicted popularity for both artists.





Conclusions

There's some improvements and further research to be done but the model is about 77.5% accurate in predicting popularity on the overall listening audience. This can be used to look at which songs to push to Discover playlists and combined with personalized data can help create playlists, which artists to invest in for Spotify Concerts, which songs Spotify can use in their ad materials to attract listeners, what types of music independent artists can tailor their albums to in order to maximize market share.

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.