Springboard

Capstone 3: Final Project Report

Problem Identification Overview

The focus of this project is to identify the key factors that drive the amount of streams a song will get in a week and create an algorithm that will define the streams based on those factors. Based on these streams, we would like to be able to get a general idea of whether or not a song will chart on the Spotify global charts.

The data will all be pulled via Spotify. First, we will download 53 weeks worth of the top 200 songs on the global Spotify charts. With the tracks and streams identified, we can then pull the track features from the Spotify API.

Unsupervised or Supervised Classification or Regression:

Supervised Regression

Deliverables:

- Jupyter notebooks
- Final report
- Presentation slide deck

Data Preprocessing Steps of Notes

Luckily, the data set captured was very clean and didn't require a lot of pre-processing. Our main work was to get a complete data set from the download site for the Top 200 Charts and get it cleanly formatted. Once we joined this data to the API, we had the full set of data which then needed to be cleaned and formatted to fit into the regression specifications.

Categorical sections were removed. The week ending date was set as the index and to get around dates not being allowed in regression, we calculated the amount of weeks from the release date and week ending date.

The next step was exploratory data analysis to get a better idea of what the data looked like which would help us in defining what the model would need to look at. When looking at the available streams, we had a fairly balanced data set when it came to positions on the charts, but there were large changes in the amount of streams per week in the first position.

Model Description

Input data size and features: 10,600 tracks

Model Algorithm and Parameters:

No models were found to be statistically significant

Along with the streams per week, the data set included track, popularity, sections, chorus hit, time signature, duration, tempo, valence, liveness, instrumentalness, acousticness, speechiness, mode, loudness, key, energy, danceability, artist, and decade features.

Model Performance

The best performance we received was on the linear regression for position 1 tracks in the training sets.

explained_variance: 0.2736

mean_squared_log_error: 0.0212

r2: 0.2736

MAE: 4,112,991.1358

MSE: 54,385,096,712,330.62

RMSE: 7,374,625.1913

Model Findings

We ran a few linear regressions and all possible variations of ARIMA models and found no strong correlations. The best correlation we found was the training set when running a linear regression on the position 1 data with an r squared of .24. However, this model showed to be overfitted to the data set when running against the test data with an r squared of *negative* .31. This meant a linear assignment would be be better than the model prediction.

Next Steps

Based on the poor performance of the model, this suggests that the data required for a solid prediction was not included in the data set we pulled. Further information on the artist and performance of previously released songs may be a good place to start. Perhaps using the global charts was not the best choice in data sets since the patterns in streams may not be consistent. We may want to next apply the approach to a single country and see if that prediction can be used on other countries. As the United States is often the world leader in entertainment, it may be helpful to predict the performance on the US first, then include the US performance on the indicators for the global charts.