In order to complete this analysis, we looked at the following datasets available from Kaggle.

top50.csv – top 50 songs from Spotify for the year 2019

spotify\_top50\_2021.csv – top 50 songs from Spotify for the year 2021

top10s.csv – top songs from 2010 – 2019

All the songs contained the same measured characteristics: BPM, energy, danceability, loudness, liveness, valence, acousticness, speechiness, and popularity. Before attempting to run any models we first removed the song and artist title from all datasets as we did not believe this would have a direct impact on the popularity of a song.

To being the analysis we ran a logistic regression to determine if there was any correlation between certain characteristics of a song and its popularity score. As shown below, the results did not correlation that would help us predict a song’s popularity score. In an effort to improve or determine correlation, the model was adjusted multiple times to include different variables (loudness, bpm, etc.) in different combinations. Unfortunately, no correlation could be made. The r-squared value of 0.198 was one of the best values seen. Some of the p-values were below 0.06. We also scaled our logistic regression values but that only hinder the results further. We believed using the largest dataset, top10s.csv, would be the most likely to provide us with worthwhile results. After conducting the first test with the results below, we tried to use the smaller datasets, but results were far worse with some r-squared values of 0.010.

Table

Description automatically generated

Our next plan of action was to run our data through a few other machine learning models.

**Naïve Random Oversampling**

Results from top10s\_2010\_2019.csv shown below. Accuracy of 61%.

**Graphical user interface, text, application

Description automatically generated**

Results from top50.csv shown below. Accuracy of 79%.

Graphical user interface, text, application

Description automatically generated

**Smote Oversampling**

Results from top10s\_2010\_2019.csv shown below. Accuracy of 68%.

Text

Description automatically generated

Results from top50.csv shown below. Accuracy of 79%.

Text

Description automatically generated

**Undersampling**

Results from top50.csv shown below. Accuracy of 70%.

**Graphical user interface, text, application

Description automatically generated**

**Combination (Over and Under) Sampling**

Results from top50.csv shown below. Accuracy of 70%.

**Graphical user interface, text, application

Description automatically generated**

**Balanced Random Forest Classifier**

Results from top10s\_2010\_2019.csv shown below. Accuracy of 68%.

Text

Description automatically generated

Results from top50.csv shown below. Accuracy of 70%.

Graphical user interface, text, application

Description automatically generated

**Easy Ensemble AdaBoost Classifier**

Results from top50.csv shown below. Accuracy of 70%.

Graphical user interface, text, application

Description automatically generated

As shown above the results do not lead us to have significant confidence in any models. Even when we scaled the data and/or increased/decreased the amount of data we were unable to determine any correlation between the data and a song’s popularity score. The analysis may show that the top songs are all consistent different characteristics. The variables that predict a popularity score are extremely different and fluctuate song to song.

In the end, it appeared that some models would generically predict if a song would be popular but would fail in determining to what extent. A model would be able to predict if the popularity score would be over the mean but the deviation from the ‘true’ popularity score was immense.