ORIE 4741 Midterm Report

Akshay Yadava (aay29), Ben Polson (bsp73)

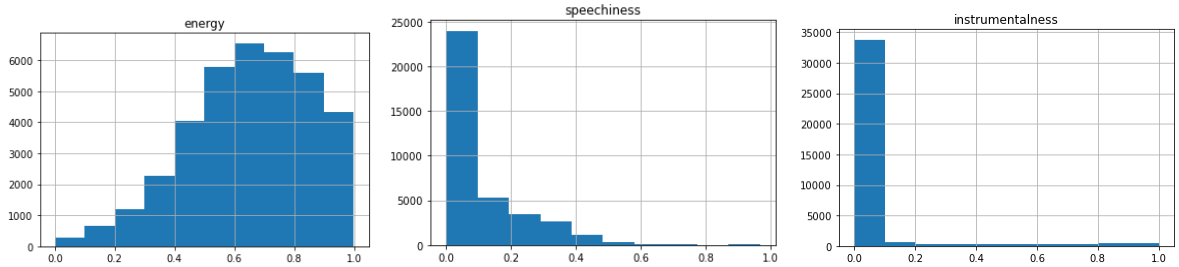
Data Preparation and Visualization

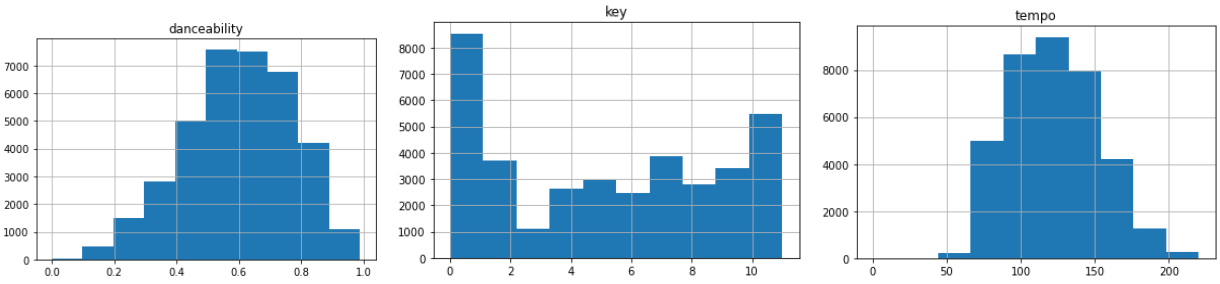
All the data used in this project was pulled from the Spotify API. There is a package labeled “spotipy” that grants users with the correct functions to gather data from the open-source Spotify API. We used the spotipy in python to get data from the API. The only way to look up a song on the API is with its ID, so in an effort to get a huge dataset of popular songs, we first found a list of the 500 most popular artists from the Rolling Stone. We then looked up these artists on the API and retrieved the song IDs of all of each of their songs. We then used those IDs to get general information on the song, and Spotify’s audio metrics on each song. We created a dataframe with this data for over 100,000 songs. We then made the dataset smaller by dropping all of the rows where popularity was less than 30 (this number was decided by looking at songs at different levels of popularity to see if they were well-known). Since all of the information was pulled directly from the Spotify API, less than 100 rows had missing values so we simply dropped them. After duplicates were removed and the song genres were encoded as one-hot vectors for analysis, the final dataset was 37,003 rows by 367 columns. Below is a breakdown of the various data types in the dataset:

Features in the dataset:

- Continuous: “danceability”, “energy”, ”loudness”, “speechiness”, ”acousticness”, “instrumentalness”, ”liveness”, ”valence”, ”tempo”
- Discrete: “popularity”, “key”, “duration_ms”
- Boolean: “Explicit”
- Text: “artist_name”, “alb_name”

Below are a few descriptive histograms describing the data in further detail

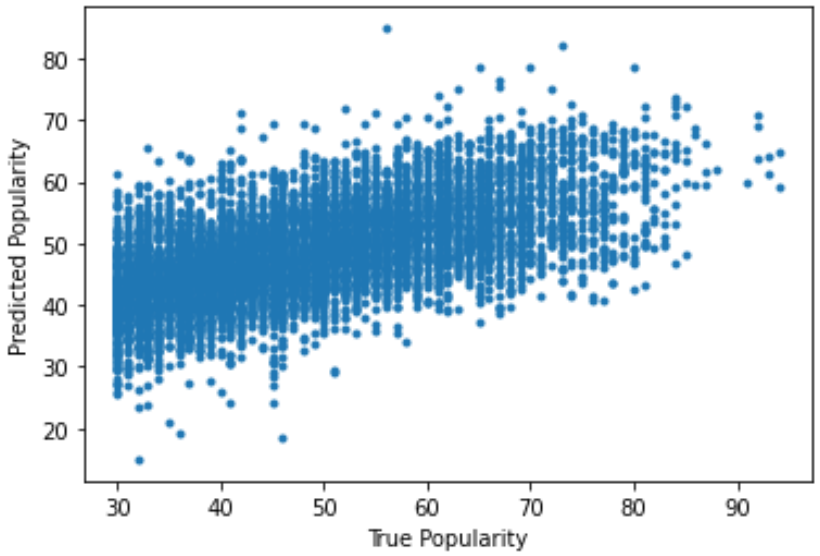




Analyses/Over/Underfitting Tactics

For our preliminary analysis, an initial ridge regression with least squares error was run using the continuous and discrete features, including the genres encoded as one hot vectors. The dataset was split into a train and validation set for both the features and labels. The feature space consisted of the continuous and discrete features of the dataset while the labels consisted of the popularity. This initial split was done to uncover the correlation between the audio features of a song and their popularity. This baseline model will give a sense of which audio metrics tend to be more informative of a song's popularity than others. From our preliminary model, the error of the training set was 104.1381059463284 and the error of the test set was 102.62703682620568. This seems to be very large for the dataset. We can see that our current model on average mispredicts the popularity of a song by about 10.

The initial model is plotted below



Effectiveness

From our preliminary results, we can tell that the model does not overfit the data, yet there is clear underfitting in the general results. This may be the result of running the regression on the entire feature space. Some features may not be as important as others in determining a song's popularity, which will be explored further.

To determine the effectiveness of our current model, we implemented an initial k-fold cross validation set setting k equal to 10. Using ridge regression with 1-2 regularization and varying alpha by various increments, 10 tests were performed and the average mean squared error was noted. From these results, we found the most effective alpha term to be 0.0001 resulting in an average training error using means squared error of 104.01756976749346 and an average test error of 106.32328436868536. This method of k-fold cross validation will remain consistent in assessing the effectiveness of various implemented models.

To Be Done

- Our initial model only utilizes ridge regression to predict the popularity of songs. In the future, various models will be utilized to garner the best regression on the dataset. Particularly, ensemble trees and K-NN may be of exploration for our particular dataset.
- Our initial model uses the entirety of the feature space to predict a given song's popularity. In reality, some features may have minimal influence on the popularity of a song and may be extracted. We plan to perform further feature selection on the dataset to select which features tend to have a high correlation with a song's popularity.
- To further test the effectiveness of the model, we plan to implement boosting in the implementation of our future models to better assess which model method is most effective
- The initial goal of our project was an unsupervised learning project in which a user could input their preferences regarding various audio features in songs and receive a generated playlist that they would most likely enjoy. Currently, we are predicting a song's popularity based on its audio features. Once a clear model is implemented correctly modeling this correlation, we plan to use this knowledge to finalize the initial goal of our project.