Capstone Project 1 - In-Depth Analysis

Spotify Audio Features

This section describes the final analysis and the machine learning techniques used for modeling and prediction. We are to use supervised learning for this problem since the data is labeled. The target variable, popularity, is numerical. That makes this a regression problem. After performing some pre-processing on the data, we will tackle this problem with the following two techniques:

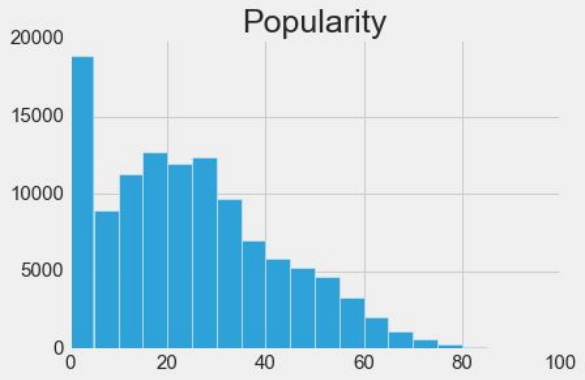
- LASSO
- Gradient Boosting

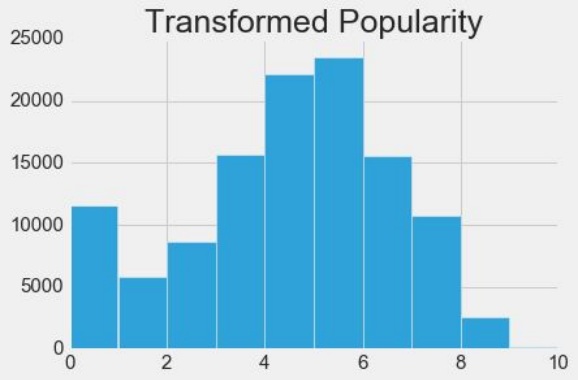
Preprocessing:

- **Turning Categorical Variables Into Dummies:** The 3 categorical variables (mode, key, time_signature) were turned into dummy variables. So the number of features to be fed into the model is now 27.

- **Handling Missing Values:** There were no missing values in the dataset

- **Transforming The Target Variable:** The popularity variable is heavily right skewed. So we did squared root transformation to make it more normal.





- **70 - 30 Train-Test Split:** The data was split by a 70-30 ratio (i.e. train - test) to be used separately for model training and validation.

LASSO Analysis:

- First we established a baseline using OLS regression analysis:
  R^2 = **0.06**
  MSE = **4.5**

- Then performed LASSO utilizing hyperparameter tuning:
  Optimal alpha parameter: 1.01e-4

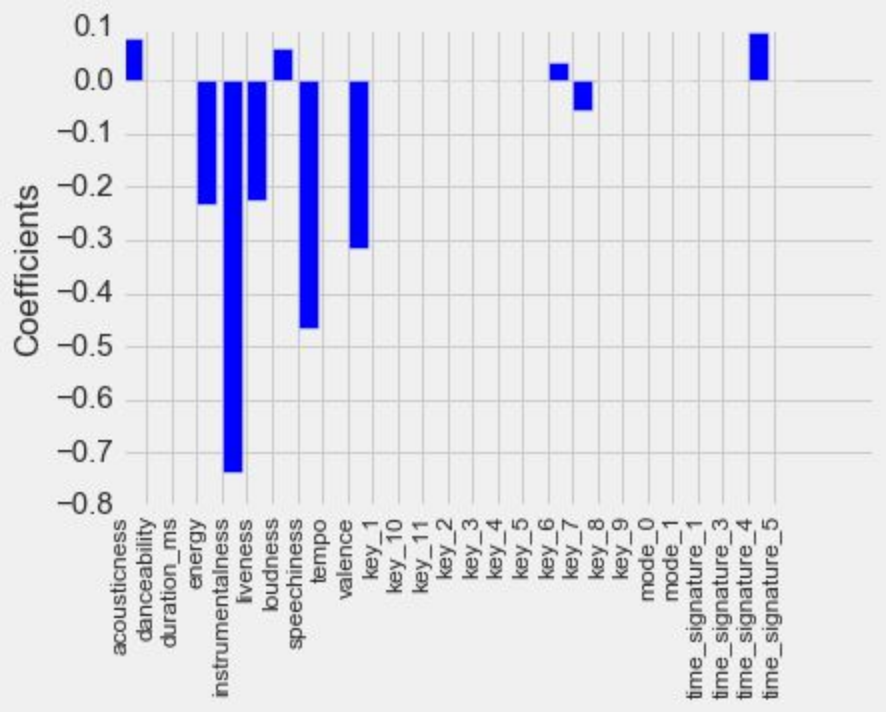
  R^2 on train data = **0.07**
  R^2 on test data = **0.07**
  MSE on test data = **4.5**

Observations:

- The LASSO model performs slightly better than OLS,
- The R^2 and MSE values are not very good, but they are the same on train and test data. This means the model is generalizable to unseen data.

Comparing LASSO Coefficients:

Figure below shows the coefficients. Some observations, a few of which are unexpected:



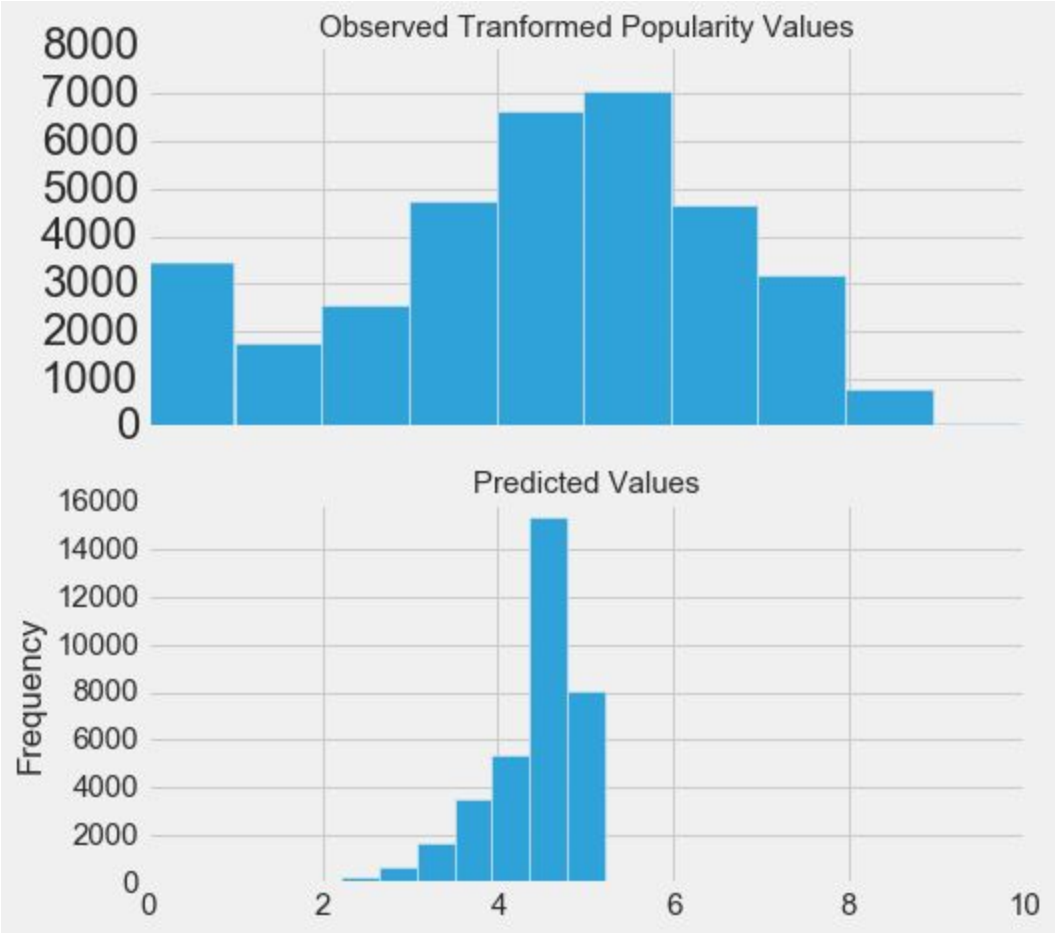
Expected:

- Instrumentalness is negatively correlated with popularity. That is what we previously saw in correlation plot
- Loudness is positively correlated with popularity. We also saw this in correlation plot.
- Time signature 4 is positively correlated with popularity (expected)

Surprises:

- Speechiness is showing as negatively correlated with popularity. In the correlation plot we see a speechiness is almost a neutral variable.
- Valence is showing as negatively correlated with popularity. It was previously almost neutral.
- Energy is showing as having negative correlation with popularity. Previously it was showing as having positive correlation.
- Liveness is showing as negatively correlated with popularity. It was previously almost neutral.

Actual vs Predicted Popularity: Both graphs are centered around 5.



Gradient Boosting:

- First we established a baseline by performing gradient boosting using default parameters: R^2 on train data = **0.14**

R^2 on test data = **0.12**
MSE = **4.2**

- Then performed gradient boosting utilizing hyperparameter tuning(i.e. Grid search):

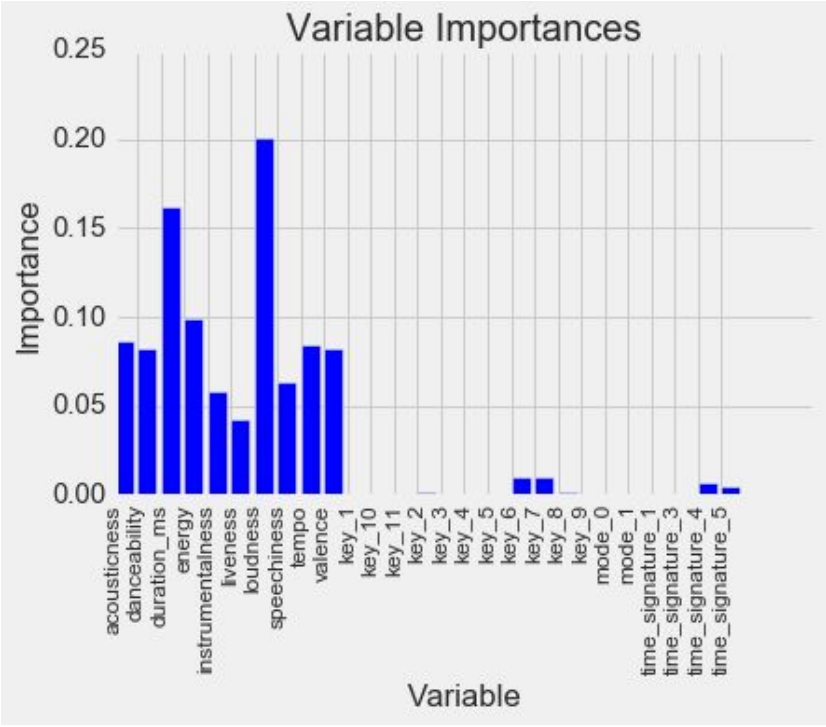
  Optimal parameters:

  {'learning_rate': 0.1,
  'max_features': 'auto',
  'min_samples_leaf': 2,
  'min_samples_split': 10,
  'n_estimators': 100}
  R^2 on train data = **0.14**
  R^2 on test data = **0.12**
  MSE = **4.2**

Observations:

- The baseline model and the model with optimal parameters perform almost the same. This indicates that our performance is near maximum and we will not gain much on performance by tuning the parameters further,
- The gradient boosting model has performed better than LASSO. So this would be the final model for this problem

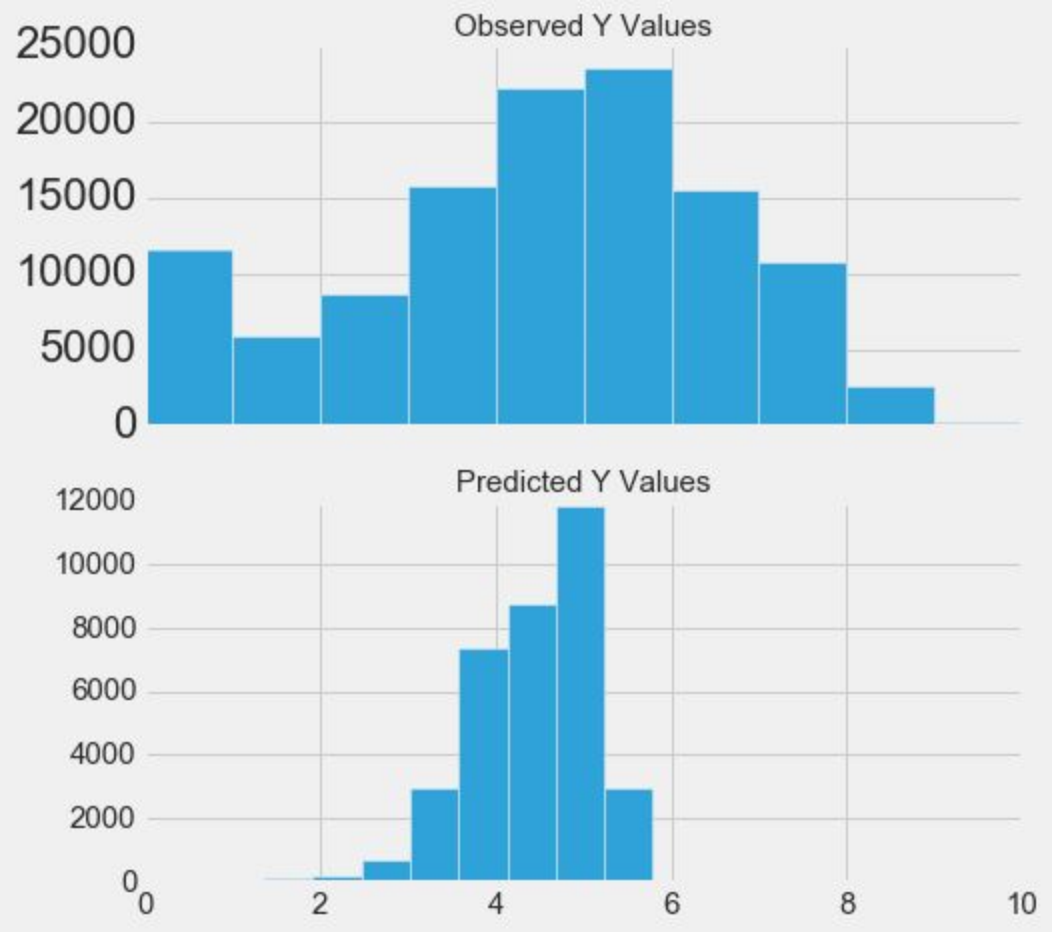
Variable Importance From Gradient Boosting Regression:



What we see in the plot above is in line with what we had seen in correlation plot in the storytelling notebook: https://github.com/hamidniki/Capstone-project1/blob/master/Capstone-Project1-Data-Storytelling.ipynb

We also see that time_signature_4 is more popular than other time signatures, which is also expected from EDA done previously.

Actual vs Predicted Popularity: Both graphs are centered around 5.



Conclusion and Next Steps:

- In this project we concluded that using the audio features available in this dataset, we can only explain about 12-13% of the variability in popularity. The Spotify Audio Features dataset was made publicly available few months before doing this project. This dataset is being expanded and by adding new features we might be able to gain performance,
- Adding 'Artist Name' as a variable to the model will likely add to the performance as well since often listening to a track is motivated by who the artist is. Though this feature

would not be considered an audio feature.