Capstone Project 1 - Final Report

Spotify Audio Features

Objective and Goal:

Question: Is there a relationship between a track's audio features and its popularity?

The goal of this project is to use the audio features of music tracks to predict the songs' popularity among listeners. Being able to predict a song or album's popularity before it is released has many applications in music business. As an example, "The Music Fund", a company based in San Francisco, CA, takes advantage of models developed using data to predict popularity of a newly produced album by an artist. If the model predicts high probability of popularity, the company acts as a label and sponsors the musician. This is a purely data-oriented decision without the need for the sponsor company to know the artist or even like their music work.

Predicting popularity of art work productions, such as music or film, also has applications in recommender systems used in digital entertainment. "Popular on Netflix" and "Recommended for you" features are examples of such applications.

Description of the Data:

The dataset used for this project is called "Spotify Audio Features" and is made available via the official Spotify web API. I accessed the dataset from the Kaggle competitions website(https://www.kaggle.com/tomigelo/spotify-audio-features) in CSV format. This is a dataset of 116,372 music tracks (one row per track) with 17 variables. The variables are Artist name, Track ID, Track name, Popularity (an integer between 0 and 100 that is calculated based on the number of streams a song had) as well as the following 13 audio features that characterize each song:

Acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature and valence. For more details on the meaning of the audio features visit:

https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Data Cleaning and Wrangling:

As expected from a kaggle dataset, this data was relatively clean. The following steps were taken to ensure creation of an analyses ready dataset:

1. **Tidiness:**
    - Each attribute is measured in a separate variable,
    - Each observation (i.e. a track) is in a separate row
    - There are no row duplicates in the data

2. Missingness:
There were no missing values in any column.

3. Examining Outliers:
Histograms of all numeric variables were examined. Of the numeric variables, only duration_ms (length of the track in milliseconds) seemed to have extremely large values (e.g. 90 minutes length). A sample of these observations were checked on Spotify's application. The duration values in all the examined tracks were correct. Some examples are:

- Artist name: John, Tracks: Whatever, 5,040,048 ms (~84 mins) and Ever, 5,610,020 ms (~93.5 mins)

- Artist name: Gentle Whispering, Track: Fluffy Sleepy Whispers, 2,097,245 (~35 mins)

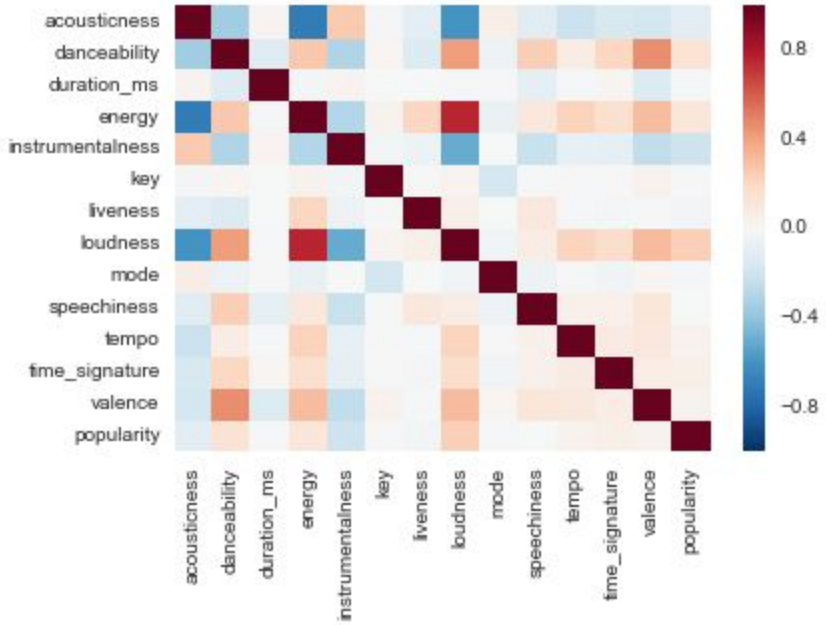
Exploratory Analysis and Initial Findings:

The Main Question: As a reminder, the question we aim to answer in these analyses is whether there a relationship between audio features of songs and their popularity?

Numerical Variables:

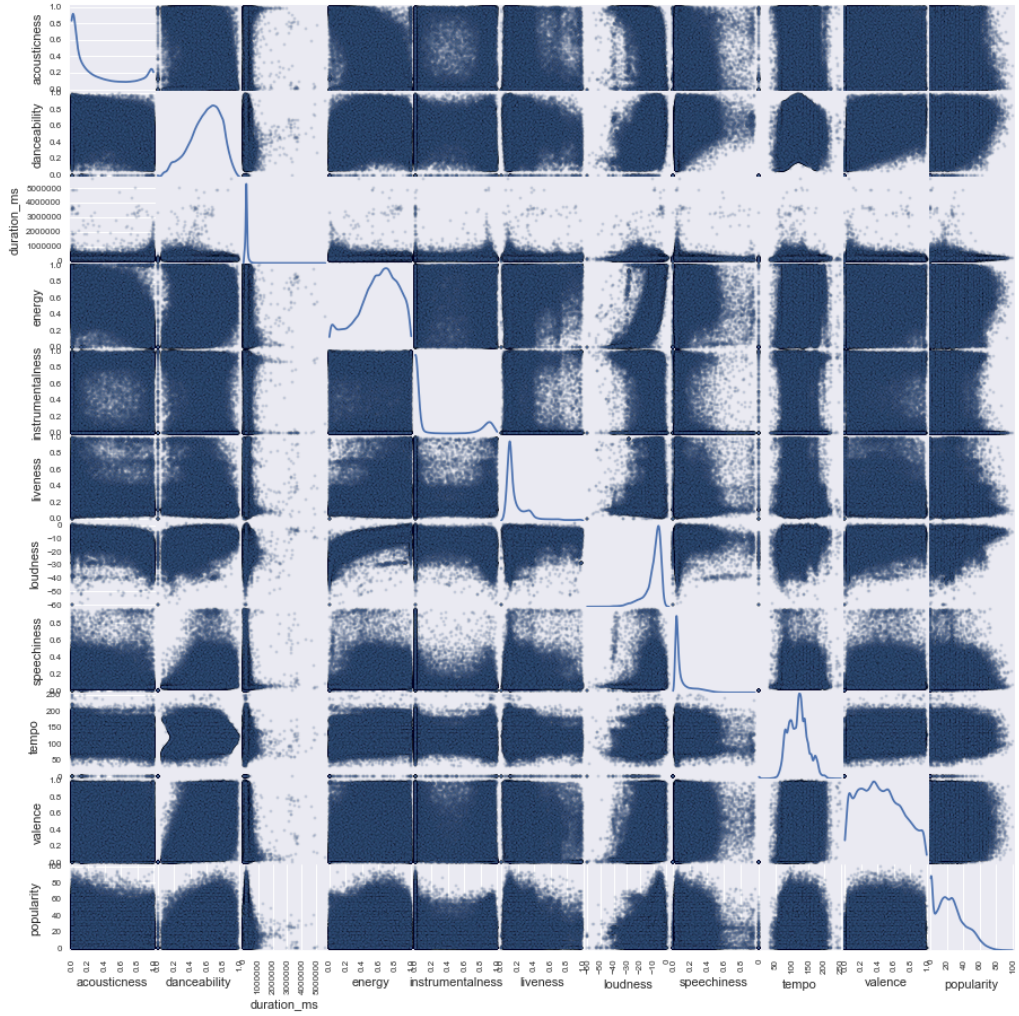
	acousticness	danceability	duration_ms	energy	instrumentalness	liveness	loudness	speechiness	tempo	valence	popularity
count	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0	116372.0
mean	0.3	0.6	212546.2	0.6	0.2	0.2	-9.9	0.1	119.6	0.4	24.2
std	0.3	0.2	124320.8	0.3	0.4	0.2	6.5	0.1	30.2	0.3	17.9
min	0.0	0.0	3203.0	0.0	0.0	0.0	-60.0	0.0	0.0	0.0	0.0
25%	0.0	0.5	164049.0	0.4	0.0	0.1	-11.8	0.0	96.1	0.2	10.0
50%	0.2	0.6	201773.0	0.6	0.0	0.1	-8.0	0.1	120.0	0.4	22.0
75%	0.6	0.7	240268.5	0.8	0.5	0.2	-5.7	0.1	139.8	0.6	35.0
max	1.0	1.0	5610020.0	1.0	1.0	1.0	1.8	1.0	250.0	1.0	100.0

Correlation plot(matrix):



- Loudness, energy and danceability are positively correlated with popularity
    - Louder and more energetic musics are more commonly used for parties and dancing
- Instrumentalness and acousticness are negatively correlated with popularity
    - Instrumentalness (absence of words in the track) limits the audience of the track to listeners who find the music itself appealing. So high instrumentalness limits the audience.
    - Acousticness is a measure of how much acoustic instruments are used in the track, as opposed to electronically produced music. Over the past decade, specially among younger generation, electronic music has become increasingly popular. This explains in negative correlation between acousticness and popularity.

Scatter Matrix:



- Tracks with very long length (e.g. 80 or 90 minutes) tend to be soft and meditative. They are not everyday music. Hence they have limited audience.

- Live tracks are less popular in spotify. Even though listening to live music as it is being played is entertaining, because of the lower quality of sound, the recorded versions of live performances are not popular compared to studio versions.

- Tracks with extreme tempo levels (< 50 BPM or > 250 BMP) are not popular

Formal Tests of Correlation for Numerical Variables:

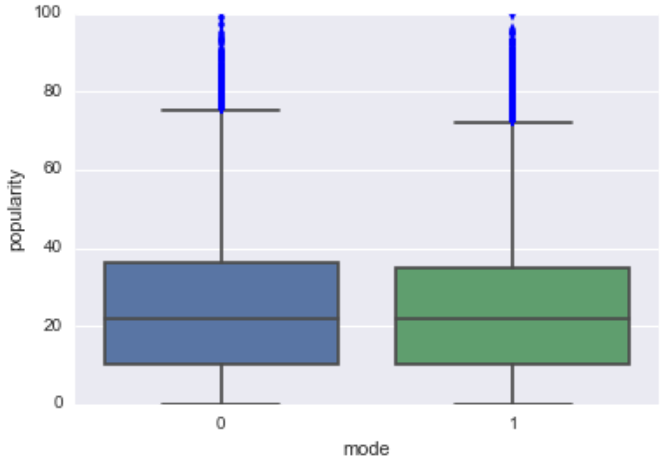
Using the bootstrap method, the following hypothesis tests were performed on the correlation between some of the numerical variables and the outcome (i.e. popularity):

	Hypothesis to Test	**Observed correlation**	**P_value**	**Conclusion**
Loudness vs popularity	H0: No Correlation Ha: Positive correlation	0.235	≈ 0.0	Rejecting the null
Danceability vs popularity	H0: No Correlation Ha: Positive correlation	0.134	≈ 0.0	Rejecting the null
Instrumentalness vs popularity	H0: No Correlation Ha: Negative correlation	-0.21	≈ 0.0	Rejecting the null
Track duration vs popularity	H0: No Correlation Ha: Negative correlation	-0.009	≈ 0.001	Rejecting the null
Liveness vs popularity	H0: No Correlation Ha: Negative correlation	-0.028	≈ 0.001	Rejecting the null

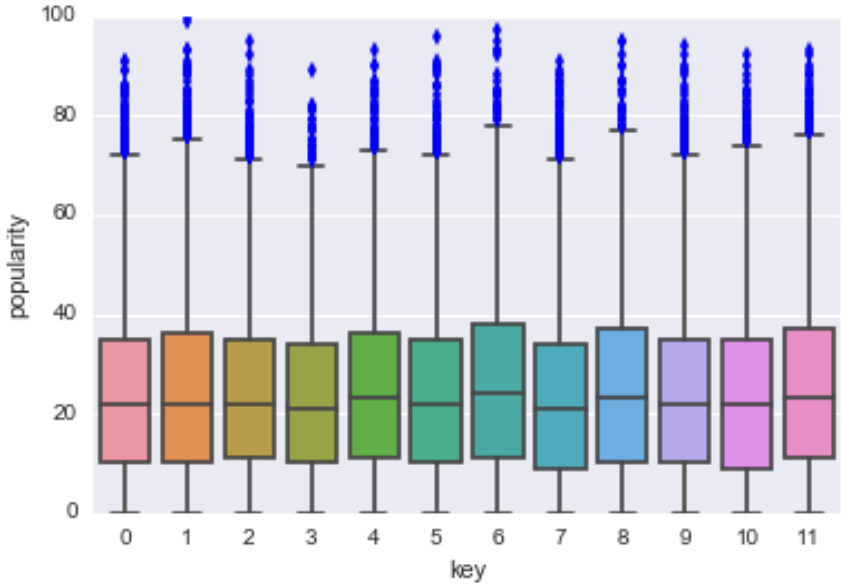
Categorical Variables (Mode, Time Signature, Key):

Variables mode, key and time_signature are represented as integer variables in the dataset. However, since these variables do not have ordinal nature, we will use them as categorical variables:

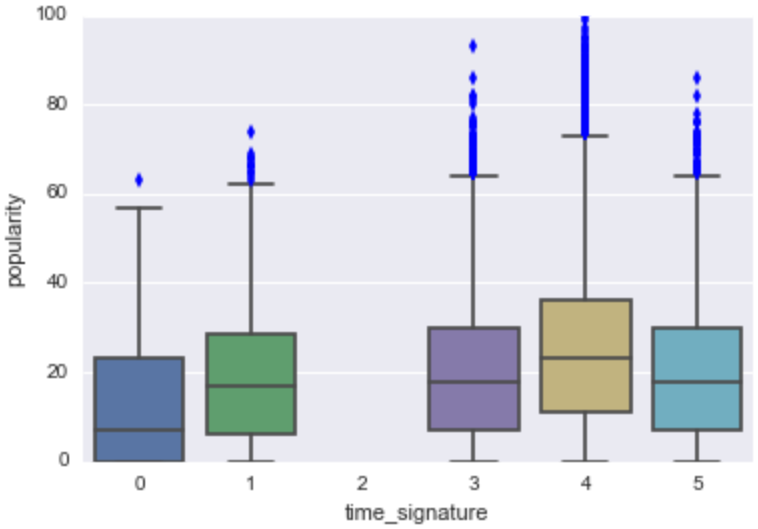
Mode (minor or major) vs Popularity: No visible difference between the two mode groups



Key vs Popularity: No visible difference between the 12 key groups



Time Signature vs Popularity:
Rhythms with 4 beats per measure seem to be more popular. This could be due to the fact that 4-beat rhythm is more common than any other rhythm. This is the most familiar beat to human ear and is the most frequent time_signature in the dataset (~ 86%)



Formal 2-Sample Test of 4-beat Time Signature vs Other Time Signatures:

Bootstrap method was used to test the following hypothesis at 5% significance level:

H0: There is no difference in popularity bewteen tracks with 4-beat time signature and other time signatures
Ha: Tracks with 4-beat time signature are, on average, more popular than other time signatures

Observed difference in means (mu_four - mu_other): **4.56**
P_value ≈ **0.0**
Conclusion: Given the small p_value, we reject the null and conclude that 4-beat time signatures, on average, are more popular than other time signatures

In-Depth Analysis:

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:

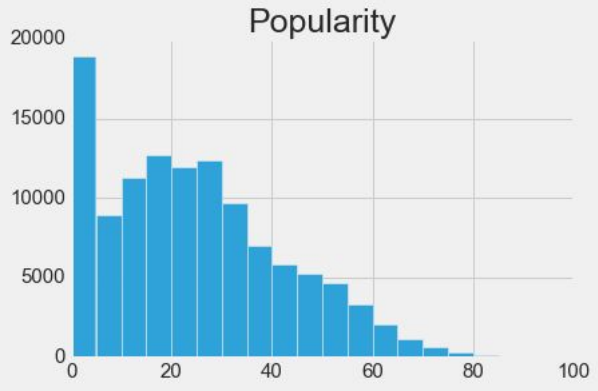
- LASSO
- Gradient Boosting

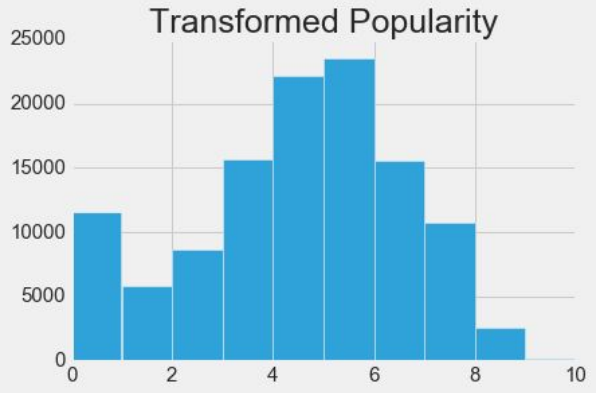
Preprocessing:

- **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:

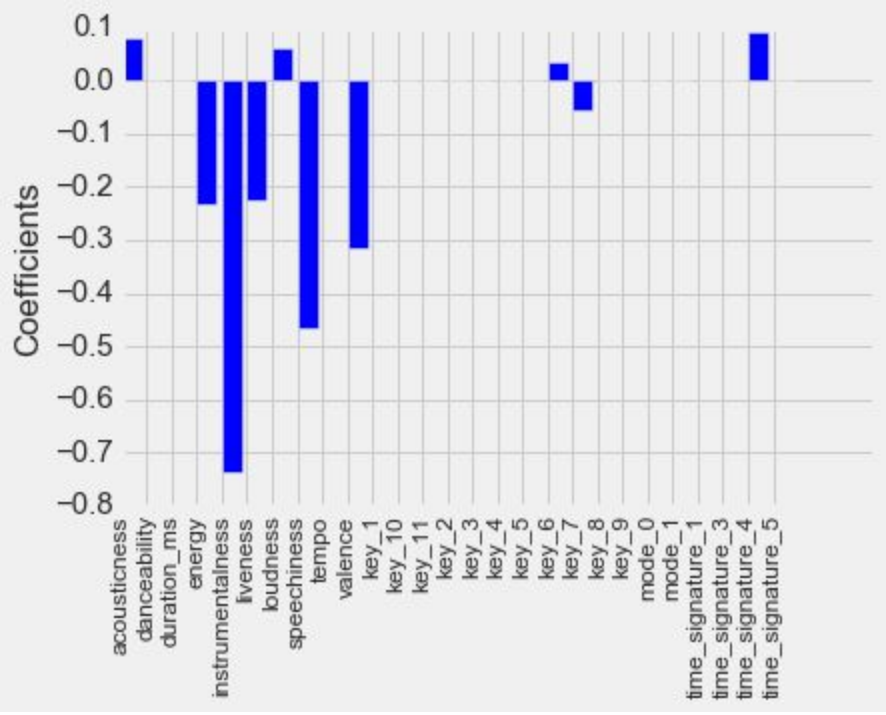
Figure below shows the coefficients. Some observations, a few of which are unexpected:

Expected:

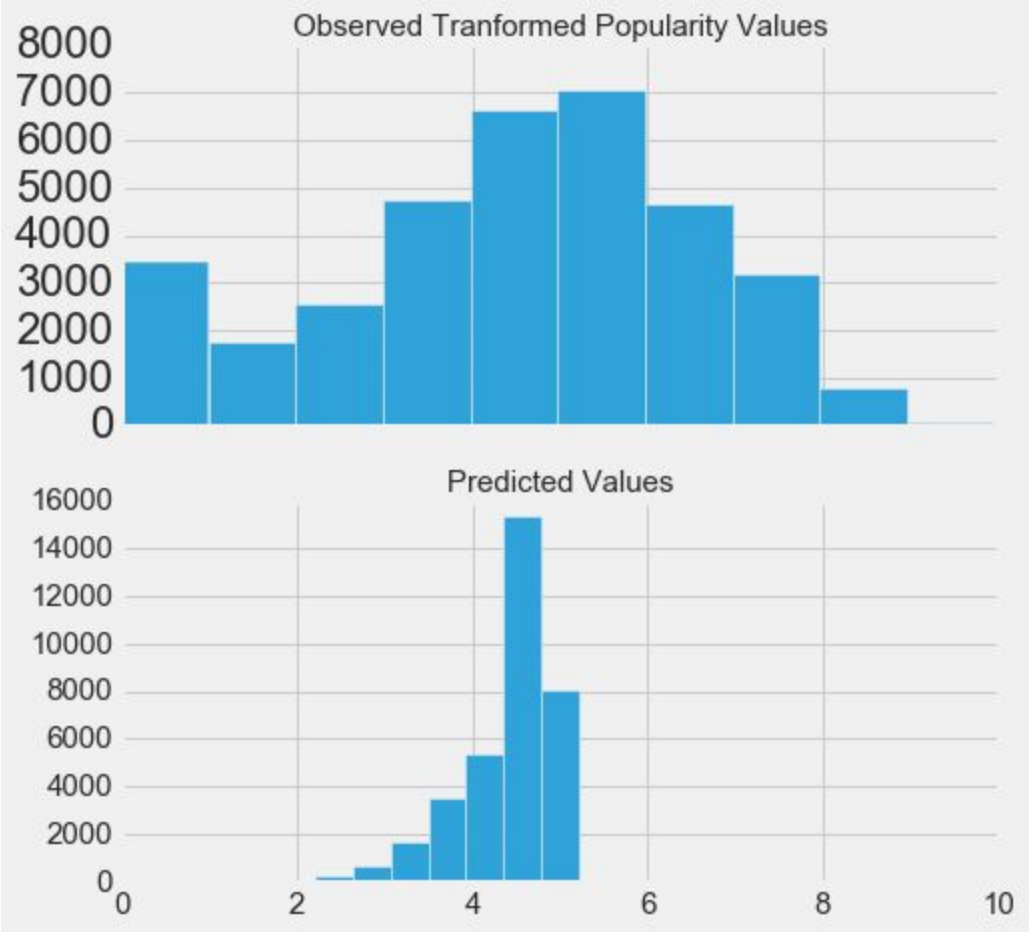
- Instrumentalness is negatively correlated with popularity. That is what we previously saw in correlation plot
- Loudness is positively correlated with popularity. We also saw this in correlation plot.
- Time signature 4 is positively correlated with popularity (expected)

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:

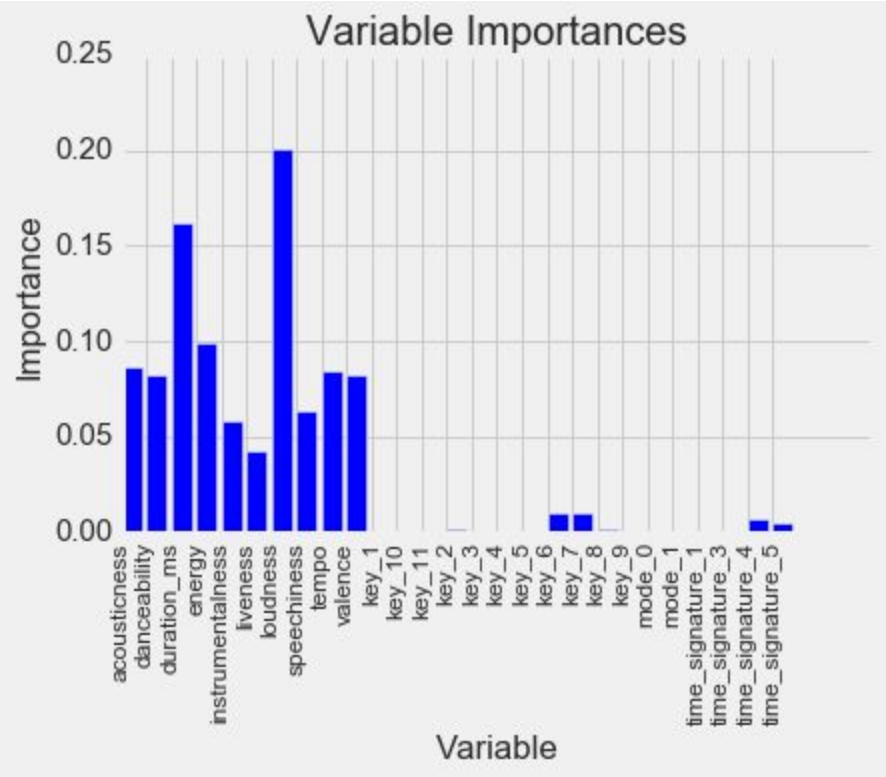
  {'learning_rate': 0.1,
  'max_features': 'auto',
  'min_samples_leaf': 2,
  'min_samples_split': 10,
  'n_estimators': 100}

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

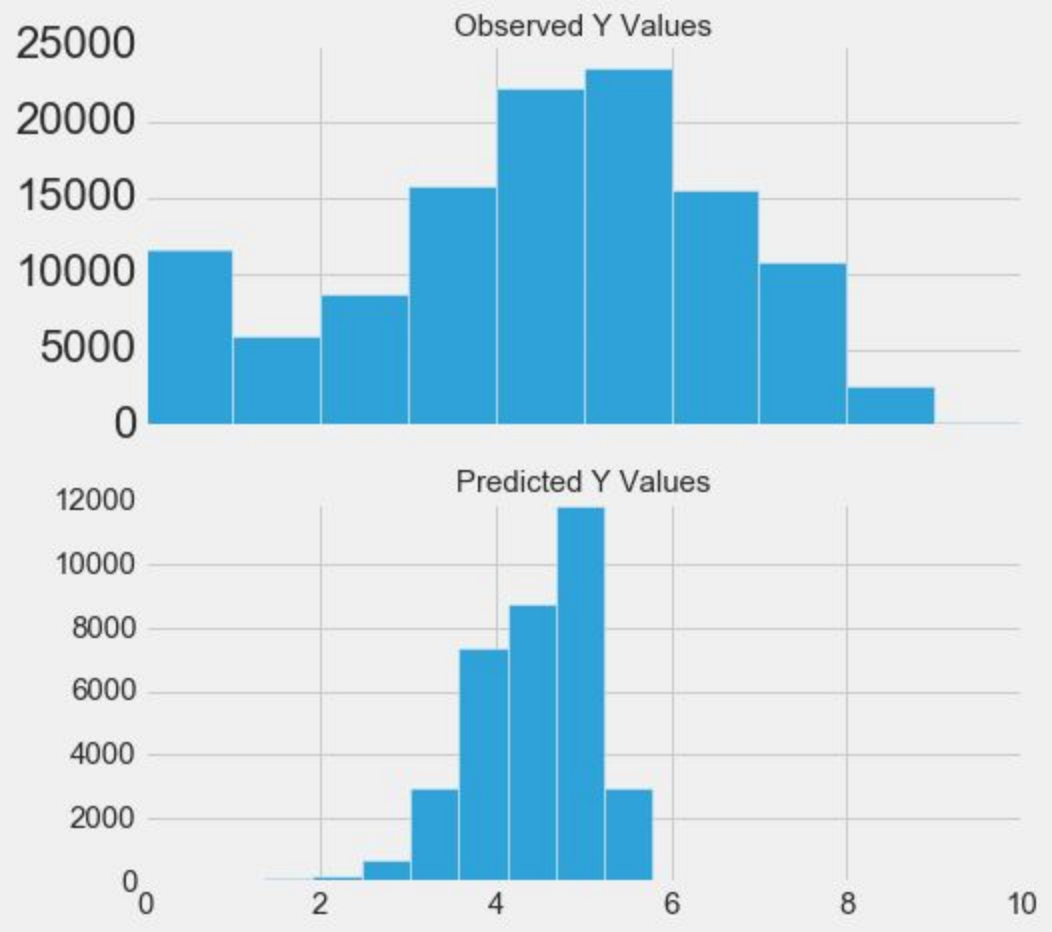
Variable Importance From Gradient Boosting Regression:



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.