**Capstone Project 1: Final Report**

**Problem Statement:**

What makes a song popular? What song characteristics, artists, & words used in titles contribute the most towards a song being more popular & can all of these attributes be used to predict the popularity of songs not yet released? These are the questions I am attempting to answer within this project using song data extracted from Spotify’s API.

These are useful questions to answer because they could most importantly result in more popular songs being created & produced for public consumption. Artists & bands can gain insight into how they can tailor or alter their music to be more popular. Recording labels & music producers could also use the results of my analysis to develop & release music that will result in more financial success for themselves.

**Description of Dataset:**

<https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks>

The dataset being used for this analysis can be found at the above Kaggle link. It contains song data for approximately 170,000 songs released between 1921-2020. The data comes from Spotify (the leading music streaming service in the world). The numeric attributes within the dataset as well as their values were created & assigned by Spotify itself. The original unaltered dataset contains 169,909 rows with 19 attributes each. After data wrangling (steps described below), the final dataset used has 61,436 rows with 14 song attributes each.

A list & description of the variables within the final dataset used:

Acousticness: How acoustic a song is on a scale from 0-1

Artists: Lists of artists credited with production of a song

Danceability: How suitable a song is for dancing on a scale from 0-1

Duration: Length of the song

Energy: How energetic a song is on a scale from 0-1

Explicit: Binary variable based on whether a song contains explicit content

Instrumentalness: Likelihood the song contains vocals on a scale from 0-1

Liveness: Likelihood the song was a live recording with an audience on a scale from 0-1

Loudness: How loud a song is in decibels

Name: Title of the song

Popularity: How popular a song is on a scale from 0-100

Speechiness: How much spoken word is within a song on scale from 0-1

Tempo: The pace of a song given by beats per minute

Valence: The amount of positiveness conveyed by a song on a scale from 0-1

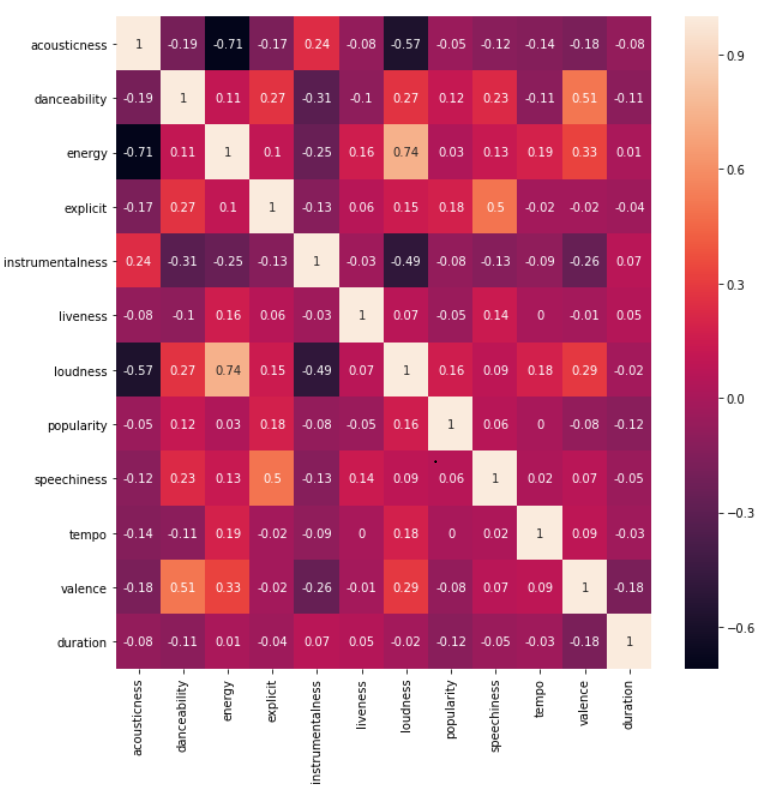
**Data Wrangling:**

After importing the data into a pandas dataframe, I first dropped attributes from which meaningful information was difficult to extract. I then condensed the dataset to include only songs after the year 1990. Since I want to use my analysis to be able to predict the popularity of future songs not yet released, I wanted only relatively recent data to be used within my models. Next, I converted the scale of the duration attribute from milliseconds to minutes as this scale is easier to understand especially in relation to the length of a song. After this, I removed songs with a duration longer than 10 minutes. I did not want songs that were 10 minutes or longer to be considered within my analysis as these could possibly be mixes, multiple songs, symphonies, or even podcasts that could skew the results of my analysis. My last wrangling step was to remove the ‘year’ attribute from the dataset because I want to use only song attributes & not time within the models that will be created to predict the popularity of future songs.

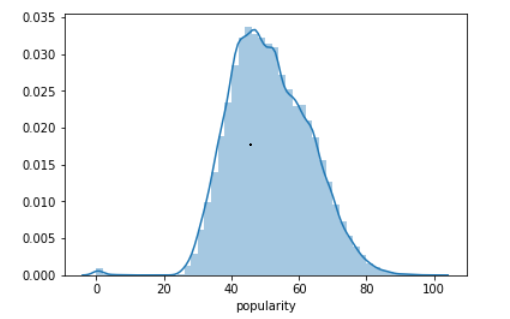
**Initial Findings & EDA:**

Taking a deeper look into the final dataset, you can see the summary statistics for the numeric attributes within the code. Boxplots & histograms were plotted to see the distributions & frequencies of all the numeric variables. Scatterplots & a correlation matrix were created as well to see if the numeric attributes & popularity are correlated. None of the numeric attributes have a significant correlation with popularity individually. The number of songs included per year within the dataset ranges from 1756 songs from 2020 to 2000 songs from both 2017 & 2018. The average popularity of songs through the years ranges from a popularity score of 39.83 in 1990 to a popularity score of 69.66 in 2019 with an increasing trend in popularity overall. The number of songs with a popularity score greater than 50 is 30,251 which is approximately 50% of the dataset. The number of songs with a popularity score greater than or equal to 75 is 1,662 which is only about 3% of the dataset. The 95% confidence interval for the mean value of popularity is between 51.13 - 51.32. Within the dataset, the most frequently occurring song titles are ‘Home’, ‘You’, ‘Runaway’, ‘Stay’, & ‘Forever’. The most frequently occurring words in song titles are ‘the’, ‘you’, ‘feat’, ‘me’, & ‘I’. The artists/bands with the most songs within the dataset are Taylor Swift, Eminem, Red Hot Chili Peppers, Nirvana, & Green Day.

Here is the correlation matrix used showing no significant correlation between predictor variables & ‘popularity’. It does show that there is some correlation between ‘energy’ & ‘loudness’ & between ‘energy’ & ‘acousticness’



Here is the distribution of ‘popularity’ within the dataset showing that ‘popularity’ has close to a normal distribution with a score around 50 being the most common.



**In-Depth Analysis Using Machine Learning:**

I began my machine learning model building to predict the popularity scores of songs on Spotify with linear regression models using just the numeric variables. The standard linear, lasso, & ridge regression models all have an R^2 score of 10.76% & an RMSE value of 11.04 on the testing set when only the numeric attributes were included in the regressions. Valence  & danceability are the attributes with the highest magnitude coefficient values in all three regressions. Next, I transformed all of the numeric variables into polynomials of degree 2 to run a polynomial regression. Even though none of the numeric variables have a quadratic/polynomial form, I wanted to see if including their polynomial forms & mainly the interactions of the numeric variables with each other would result in a better model. The polynomial regression model with the numeric variables has an R^2 score of 17.89% & an RMSE value of 10.59 on the testing set which shows this model is performing better than the linear regression models.

After focusing on linear/polynomial regressions involving only numeric attributes, I shifted my attention to the artists attribute. The artists attribute now contains either a 1 or 0 for each artist value depending on if that artist is credited with production of a song. Running a standard linear regression using only the artist attribute to predict song popularity score resulted in a negative R^2 score on the testing set & very high RMSE value. This shows that the regression is very overfitted to the training set. Switching to a lasso regression resulted in an R^2 score of 29.33% on the testing dataset & an RMSE value of 9.83. The use of the lasso regression reduced the presence of overfitting, I will be using only the lasso regression for linear regression models moving forward. Artist names involving ‘arijit’ & ‘firme’ have the highest magnitude coefficient values in this lasso regression model. Running a lasso regression model including both numeric & artist attributes resulted in an R^2 score of 31.86% on the testing dataset & an RMSE value of 9.65. The names ‘arijit’ & ‘firme’ have the highest magnitude coefficient values in this model as well.

My next models focused on taking into account the words used in titles of songs. Words in titles now also either have a value of 1 or 0 attached to them depending on if that word was used in a particular title. The lasso regression using only the words used in titles to predict song popularity has an R^2 score of only 6.11% on the testing set & an RMSE value of 11.33. The words ‘alguien’ (means someone in Spanish) & ‘bieber’ (referring to Justin Bieber) have the highest magnitude coefficient values. Artist names can appear as words in titles when the artists have a guest appearance aka a feature on a song. The lasso regression using all the attributes: numeric, artists, & words in titles has an R^2 score of 31.87% on the testing dataset & an RMSE value of 9.65. Artist names involving ‘arijit’ & ‘firme’ have the highest magnitude coefficient values again.

Finally, I finished my analysis with random forest & gradient boosting regression models. A random forest regression model involving only numeric variables has an R^2 score of 16.52% on the testing dataset & an RMSE value of 10.68. Loudness is the most important numeric feature within this model. A random forest regression model involving all variables has an R^2 score of 16.38% on the testing set & an RMSE value of 10.69. Loudness is the most important feature in this model as well. The gradient boosting regression model with only numeric attributes resulted in an R^2 score of 22.39% on the testing set & an RMSE value of 10.30. The gradient boosting regression model with all of the attributes resulted in an R^2 score of 25.51%  on the testing set & an RMSE value of 10.09. Loudness is the most important feature in the gradient boosting models as well.

A summary of all the models can be seen below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Training Set R^2** | **Training Set RMSE** | **Testing Set R^2** | **Testing Set RMSE** |
| Standard Linear Regression w/ Only Numeric Variables | 10.81% | 11.05 | 10.76% | 11.04 |
| Lasso Regression w/ Only Numeric Variables (alpha = 1e-08) | 10.81% | 11.05 | 10.76% | 11.04 |
| Ridge Regression w/ Only Numeric Variables (alpha = .001) | 10.81% | 11.05 | 10.76% | 11.04 |
| Polynomial Regression (degree = 2) | 18.38% | 10.57 | 17.89% | 10.59 |
| Standard Linear Regression w/ Only Artist Names | 36.90% | 9.29 | -4.34% | 2.43639E+13 |
| Lasso Regression w/ Only Artist Names | 36.27% | 9.34 | 29.33% | 9.83 |
| Lasso Regression w/ Numeric Variables & Artist Names | 38.40% | 9.18 | 31.86% | 9.65 |
| Lasso Regression w/ Only Words Used in Titles | 12.37% | 10.95 | 6.11% | 11.33 |
| Lasso Regression w/ Numeric Variables, Artist Names, & Words Used in Titles | 42.50% | 8.87 | 31.87% | 9.65 |
| Random Forest Regression w/ Only Numeric Variables | 17.02% | 10.66 | 16.52% | 10.68 |
| Random Forest Regression w/ Numeric Variables, Artist Names, & Words Used in Titles | 17.39% | 10.63 | 16.38% | 10.69 |
| Gradient Boosting Regression w/ Only Numeric Variables | 28.51% | 9.89 | 22.39% | 10.3 |
| Gradient Boosting Regression w/ Numeric Variables, Artist Names, & Words Used in Titles | 30.05% | 9.78 | 25.51% | 10.09 |

Overall, the best performing model is the lasso regression model involving all of the attributes.

We must take into account that even the best performing model in this analysis can only explain about 30-40% of the variability in song popularity scores. I have learned that it can be difficult to predict things that involve human tastes/preference such as song popularity as they can be very unique. More advanced & intricate models along with more data may need to be used to predict song popularity more accurately.