**Capstone Project 1: Final Report**

**Problem Statement:**

What makes a song popular throughout the world? What song attributes, words used in titles, & artists contribute the most to a song being more popular & can these attributes be used to predict the popularity of songs not yet released? These are the questions I am attempting to answer using a dataset that has been extracted from Spotify (the leading music streaming service in the world).

This is a useful question to attempt to answer because it most importantly could result in more popular songs being created & produced. It will allow artists & bands to gain insight into how they may need to tailor or alter their songs for the songs to be more popular & have more worldwide appeal. Recording labels & music producers can also use the results of my analysis to develop & release more popular music. Creating & releasing more popular music will also result in more financial success.

**Description of Dataset & Data Wrangling:**

The dataset being used for my analysis was obtained from Kaggle (an online community for data scientists & machine learning) but the data itself was originally extracted from Spotify. The dataset contains attributes of songs including the song’s title, artist, genre, year, beats per minute, energy (how energetic the song is), danceability, loudness, liveness (how likely the song is a live recording), valence (how positive the mood of the song is), duration, acousticness, speechiness (how much spoken word the song contains), & popularity. The numerical attributes & their values were created & assigned by Spotify. The initial dataset contained 603 rows of data each with the 14 attributes just listed.

After obtaining the data, the first step I took in cleaning the data was to rename the attributes to make them easier to understand when importing the data from the excel file. When looking into the descriptive statistics of the numeric attributes, I saw that a lot of them had a minimum value of 0. One song in particular had a value of 0 for almost all of the attributes & this song was removed from the dataset. The minimum values of 0 that remained were valid values that could be given to a song attribute by Spotify’s system. Next, I determined that multiple rows existed for certain songs within the dataset. These rows contained all of the same information except for year & since I was not going to use year as a variable to determine popularity, I only kept one row for each of these songs. One song was listed twice with different popularity values & both of the rows associated with this song were removed. After the necessary data cleaning & wrangling steps were performed, the dataset contained 586 rows of data.

**Exploratory Data Analysis:**

To begin my exploratory analysis, I plotted each numeric attribute vs popularity to see if I could detect a possible relationship or correlation. None of the attributes were seen to have any correlation with popularity individually. Any possible nonlinear relationship between the attributes & popularity could not be detected either through the use of residual plots. A correlation matrix was created to see if any attributes had a strong relationship with each other. Two pairs of attributes could be seen as having correlation (energy/acousticness & energy/loudness). Only these two pairs of attributes having a correlation value above 0.5 was surprising. I would expect other pairs of attributes such as energy/danceability, energy/beats per minute, & acousticness/speechiness to be more strongly correlated. Histograms were also created to see the distribution of each numeric attribute within the dataset & which values occurred the most frequently for each attribute.

When looking into the distribution of number of songs by genre, it was seen that a highly disproportionate number of songs either directly belonged to or were associated with the pop genre. Due to this, it would be difficult to determine the effect of a specific genre on popularity, at least when using this dataset, so genre was not considered going forward. Also, it should be noted that all the songs within the dataset were from the years 2010-2019 but again year will not be used as an attribute in my analysis. The artists with the most number of songs within the dataset were Katy Perry, Rihanna, Justin Bieber, Lady Gaga, & Maroon 5. The words that were used the most within the titles of songs were feat. (indicating the song featured another artist), you, me, love, & I.

Looking specifically into the popularity variable, the values of popularity within the dataset ranged from 0 to 99 with a mean of 66.5 & a standard deviation of 14.4. A 95% confidence interval was constructed for the mean value of popularity using bootstrapping & the interval came out to be between 65.4 & 67.7. 518 of the 586 songs had a popularity score higher than 50, 269 songs had a popularity score higher than 70, & 166 songs had a popularity score higher than 75.

**In-Depth Analysis w/ Machine Learning**:

My analysis to predict the popularity of songs began with creating linear regressions. The first linear regression only included the numeric attributes from the dataset as predictor variables. This regression involving only the numeric attributes had a R^2 value of only 4%. ‘Energy’, ‘loudness’, & ‘length’ were the only significant variables in this model with ‘loudness’ having the most variation & the largest coefficient value of 1.09. This model had a RMSE value slightly over 13 when used to predict the popularity of songs within a test set. Although its R^2 score was very low, this model had actually one of the lower RMSE values observed out of all the RMSE values seen for models used. To see if the interactions of the numeric variables with each other or if their values squared could result in a model with a lower RMSE value, a polynomial regression model (of degree 2) was created & run but this model had a higher RMSE value.

The next step was to incorporate words used in the titles of songs as predictor variables. To do this, I used TfidfVectorizer to get all the individual words used in titles within the dataset & then kept only words that appeared at least 5 times but not in more than 70% of the songs (these are arbitrary restrictions that can be adjusted). I then regressed popularity on the TFIDF (term frequency – inverse document frequency) values of the songs & this resulted in a model with a very high RMSE value due to overfitting to the number of words present. Only 3 words were significant in the model & they were ‘version’ (-20.43), ‘with’ (10.97), & ‘your’ (-19.01) with all 3 having large coefficient values both negative & positive. I converted the TFIDF values into binary values (either 0 or 1 depending on if a word was not present or present within a song’s title) for better explainability. Using binary values instead of TFIDF values did not change the outcome of the regression that was just created. I now combined the numeric attributes with the binary word attributes to run another regression. As expected, this model still had a very high RMSE value due to overfitting of the word attributes & the previous numeric & word attributes that were significant were significant in this model as well.

As I did with words used in titles, I used TfidfVectorizer to get all the individual names of artists or musical groups who created the songs within the dataset. Only names that appeared at least 3 times in the dataset were kept. If the individual names that remained still only referred to one artist or band, only last names were kept to reduce dimensions (this was done by checking if names had the exact same coefficient values in the initial model using artist names). The TFIDF values were converted to binary values (either 0 or 1 depending on if a name was not present or present within the artist’s or group’s name). When regressing popularity on individual names, the linear model had a higher R^2 (0.28, but adjusted R^2 of 0.17) compared to the models that were created up until now but it also had a very high RMSE value due to overfitting to all the names present. The significant names in the model were ‘bieber’ (6.60), ‘birdy’ (-12.85), ‘clarkson’ (34.40), ‘grande’ (10.33), ‘harris’ (12.75), ‘jennifer’ (-47.45), ‘kelly’ (-37.45), ‘lewis’ (30.55), ‘lopez’ (37.90), ‘maroon’ (9.63), ‘martin’ (-15.70), ‘mendes’ (11.85), ‘minaj’ (-20.62), ‘sheeran’ (12.91), ‘smith’ (21.88), ‘spears’ (-10.23), ‘stefani (-18.45)’, & ‘timberlake’ (-7.58). These are all parts of names of well-known artists & some of the names had surprising coefficient values. Even though an artist might be well known, the songs of theirs that are included in this dataset may not have been well received & popular. Artist names were combined with the numeric attributes to run regressions as well but again as expected the models were overfitting & had high RMSE values.

The last step with linear regression models in my analysis was to combine together & use the numeric, words in titles, & artist names attributes to predict popularity. The model including all attributes had an R^2 of 0.41 but a lower adjusted R^2 of 0.21 as well as a very high RMSE value again due to overfitting. So, only the significant attributes from this model were taken to create another regression. The significant attributes were 'energy' (-0.20), 'danceability' (0.16), 'loudness' (1.44), 'be' (-11.07), 'good' (9,94), 'lose' (12.05), 'one' (-13.24), 'your' (-11.97), 'bieber\_artist' (4.51), 'clarkson' (36.17), 'grande' (9,17), 'harris' (14.07), 'jennifer' (-57.27), 'kelly' (-38.32), 'lewis' (6.52), 'lopez' (47.32), 'maroon' (7.62), 'martin' (-16.42), 'mendes' (10.83), 'minaj' (-24.19), 'smith' (23.28), 'spears' (-12.02), 'stefani' (-23.56), & 'timberlake' (-11.00). The model including only these attributes had a R^2 value of 0.23 & a similar adjusted R^2 of 0.20. The coefficients of the attributes in this model are listed next to the attributes above. Most importantly, the model had a relatively low RMSE value of 14.93. This RMSE value is still slightly higher than the RMSE value from only using numeric attributes but the model includes a lot more predictor variables that can explain differences in popularity.

Finally, along with linear regression models, I also ran decision tree, random forest, gradient boosting & kNN regression models only using the numeric attributes to see if these models could be better predictors of popularity. The random forest model resulted in the lowest RMSE value seen (13.16) while the kNN model using 9 nearest neighbors was close with a RMSE value of 13.41.

**Conclusions:**

In conclusion, using this dataset, I was able to explain about 20% of the variation in song popularity scores with the significant attributes listed above. Although this is not a high % of variation being explained, song popularity does depend on human preferences which are difficult to predict. Also, it should be noted that the results obtained through my analysis were skewed by the specific songs present within the dataset used. My analysis consisted of only 586 out of the millions of songs available on Spotify & this was the major factor in my models not having higher predictive power as well as in the obtaining of unexpected coefficient values associated with the attributes used. For example, as mentioned above, certain artists had surprisingly low or high coefficients attached to their names & this could be due to only their unpopular or popular songs being included in the dataset. The effectiveness or ineffectiveness of numeric attributes as well as words used in titles would change as well given new or more song data. Numeric attributes had surprisingly low coefficients in determining the variance in popularity but were still relatively effective in predicting the popularity of songs within test sets by themselves.

Coming to the prediction of popularity of songs, the most effective models created were able to predict the popularity score of songs within a +/- 13-15 score range. The ‘best’ linear regression model, decision tree model, random forest model, gradient boosting model, & kNN model all had RMSE values that fell within this range. With more data & hyperparameter tuning, this error range would be lowered.

**Future Work:**

If this analysis was to be continued, the next step would be to extract more data from Spotify’s API & apply all the steps that were taken so far & models that were created on the additional data to see what insights can be generated & if popularity can be predicted more accurately. With more attributes, PCA can be used to reduce & combine dimensions in a way to predict popularity effectively. More advanced NLP techniques can also be used to extract more meaning from words used in song titles & which artists are associated with a song. If the additional songs being analyzed are more evenly distributed between different genres, genre can also be used as a predictor variable.