ORIE 4741 Midterm Report

Kevin Van Vorst (kpv23), Manya Walia (mtw62), Saaqeb Siddiqi (ss3759)

November 8, 2020

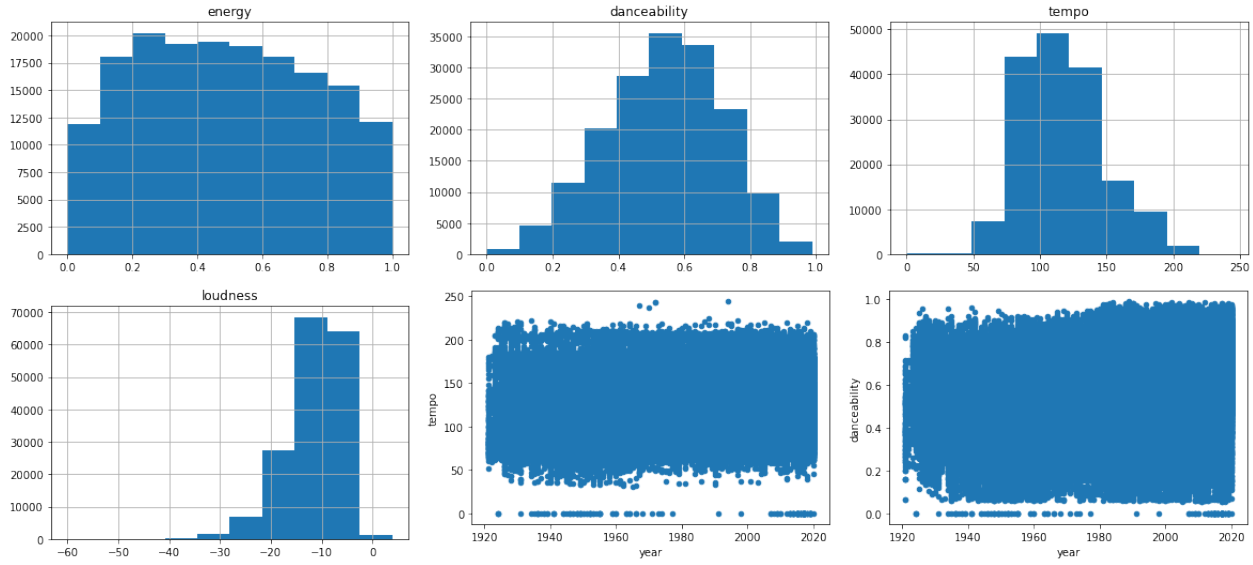
Data Preparation and Visualization

The original dataset contained 169,909 rows and 19 columns. Each row represents a unique song while each column contains the features of the song such as danceability, tempo, key, etc.. First to clean the dataset all rows with missing values were dropped. Surprisingly, the dataset was complete and no rows were dropped. Next, two columns were dropped before the preliminary analysis was performed. The first dropped column was the "id" column which contained for each song, a unique string of integers and letters generated by Spotify to identify the track. That means with 169,909 unique identifiers, this column is irrelevant for our analysis and must be dropped. The other dropped column was the "release_date" column which contained the release date of the song. Compared to the "year" column, this data is too specific for our analysis as we question trends across years and not monthly or daily. Only the "year" column will suffice with our time related analyses and therefore the "release_date" column was dropped. After cleaning, the resulting dataset was 169,909 x 17.

Below is the breakdown of the different data types in the dataset:

- Continuous: "acousticness","danceability","energy","instrumentalness","liveness","loudness", "speechiness","tempo","valence"
- Discrete: "duration_ms", "key", "popularity", "year"
- Boolean: "explicit", "mode"
- Text: "artists", "name"

Before performing our preliminary analyses, our data was visualized through a few scatter plots and histograms.



Preliminary Analyses

As a preliminary analysis, an ordinary least-squares error regression model was fit to 70% of the real-valued data and a summary of the results was generated. The adjusted R^2 of the model is 0.837, which indicates that our model fits the data decently well. The F-statistic probability is 0.00, which proves that all regression parameters are non-zero and that our regression model has good validity in fitting and predicting our data. The Mean Squared Error of the model on our training set is 100.52677995484363, and the Mean Squared Error of the model on our testing set is 101.03827809516288, which once again, are decent MSE values for such a large dataset.

Our model provides us with a headstart on understanding our data and answering our questions. The coefficients of each feature indicates that the danceability, explicitness, and energy levels of a song are largely positively correlated to its popularity, and the speechiness, acousticness, and valence of a song are largely negatively correlated to its popularity. The standard errors also indicate that tempo and key of the song do not have a very significant impact on its popularity.

Based on these results, we can perform feature selection now. It would be most lucrative to drop the tempo and key features in order to get a more accurate prediction. After doing so, we noticed that our MSE values fell by 4 and this gave us a more accurate model.

Effectiveness

This dataset is very clearly underfitted based on all the figures displayed below. It seems that there is a heavy correlation between years and popularity by observing Figures 2 and 3, signifying that there is an inherent disconnect between the realities behind the correlation of year to popularity. For this, some regularization upon the Year's could improve our overall understanding of the dataset.

Once we exclude the Year's column, we notice that our elementary regression model fails to mathematically identify what columns affect popularity even with the isolation of songs that premiered in 2019. However, when we observe Figure 4, we notice that there are two unique clusters that indicate where most of the errors occur within the prediction. For this, it may be beneficial to possibly combine features as well utilize a more complex model to weigh features uniquely.

Another element that could definitely improve our finalized model would be to introduce k-fold cross validation. Whether or not we remove or regularize the year's column, we must introduce this validation method to effectively train our model to decrease the possible overfitting that may occur by lazy training.

These techniques may create a greater and in-depth understanding of the dataset and hopefully provide opportunities to lower the errors.

Going Forward

- By our preliminary analysis, we could see an inherent disconnect between a song's popularity and the year. There is an overreliance on the year the song was released, possibly creating a unique definition of popularity that may be skewed by the current year.
- So far we have been utilizing the entire dataset, ignoring important splits within the dataset that has been provided to us in the form of different files. For example, the three files that could be of importance are data_by_genre.csv, data_by_artists, and data_by_year.csv. These splits may further the model's understanding of what makes a song popular.
- By observing the figures provided to us above, it is clear that clusters and outliers exist within this dataset, proving to us that we may need to improve our model by using higher-order or more complex modeling techniques in order to properly predict a song's popularity.
- While it is clear that our models are heavily underfitted, we must not over complicate the usage and usefulness of some of the features provided to us. While a song's genre combined with other features may fit within a unique cluster, creating additional layers than are necessary may heavily overfit this modeling.

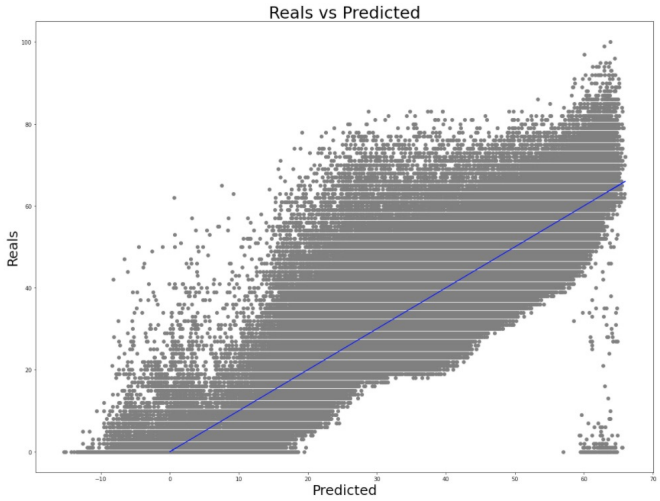


Figure 1: All column values incorporated into regression analysis.

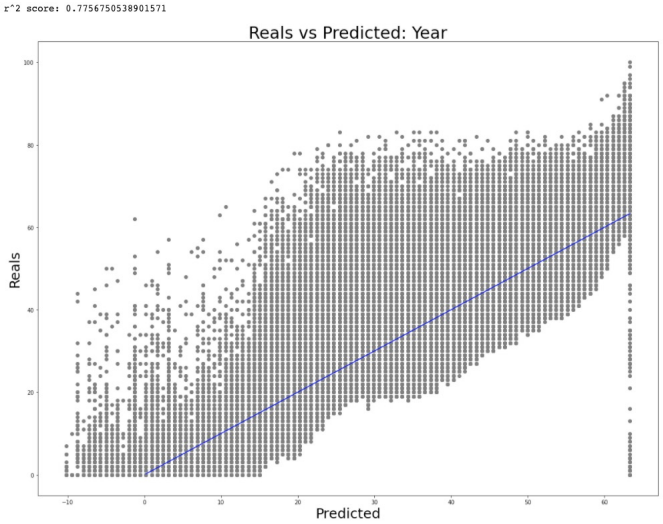


Figure 2: Only Year column values incorporated into regression analysis.