DS 3001 Project
Results

Lauren Turner, Dana Pham, Amirah Hossein, Ella Thomasson, Maggie Crowner

See https://github.com/maggiecrowner/DS3001-Project/blob/main/Project.ipynb for code

Introduction

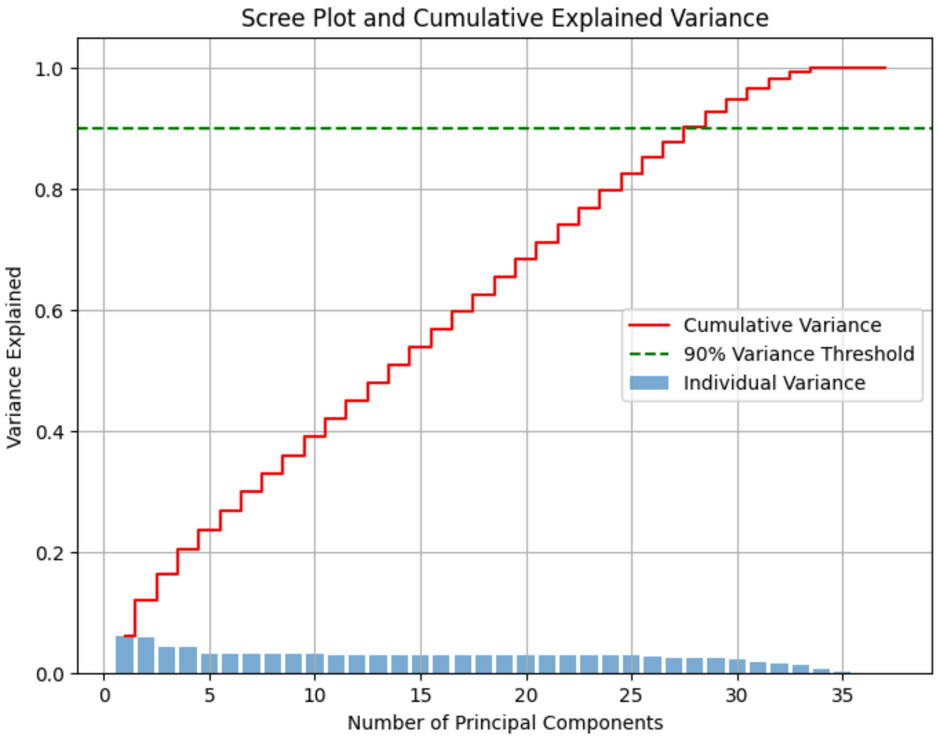
In our study, we sought to answer two main prediction questions:

- How well can we predict song popularity for future songs?
- Which song characteristics are most useful in predicting the popularity of a song?

Our analysis aims to uncover the relationship between various song characteristics and their impact on Spotify popularity, as well as evaluate the predictive performance of our Random Forest Model. By examining the feature importance scores and model accuracy, we can identify which attributes, such as danceability and energy, are most strongly associated with a song's success. This paper presents the key findings and their implications for understanding and predicting song popularity on Spotify.

Principal Component Analysis

In order to select the number of components to use for PCA, we created a scree plot to show the extent to which the variance is explained by each principal component.



The red line represents the cumulative variance achieved by using the amount of components listed on the x-axis in our model building. Since no components explain a majority of the variance, and the 90% variance threshold is not achieved until principal component 28, it is clear that PCA is not necessary for our dataset. The predictors are not correlated with one another enough to build principal components that would be more useful in our analysis than the predictors themselves. If we were to use 28 components, it would be extremely difficult to interpret. Therefore, we will move forward with our model building with the original dataset; we will not use PCA for this model.

Random Forest Model

In order to create our random forest model to predict song popularity, we first performed cross-validation for hyperparameter tuning to determine the optimal values for maximum depth of the trees and the number of trees to calculate in the random forest. We performed 3-fold cross-validation on the training set to test:

- max_depth = 3, 5, 10
- max_features = ‘sqrt’, ‘log2’
- min_samples_split = 3, 5, 10
- min_samples_leaf = 1, 2, 3
- n_estimators = 50, 100, 150

The CV yielded the highest R^2 value for max_depth = 10, max_features = ‘sqrt’, min_samples_split = 3, min_samples_leaf = 1, and n_estimators = 100. We refitted the Random Forest Model with these parameter values on the training set. Then, we predicted on the test set using this model to determine model accuracy and how well it works for prediction.

Training Statistics:

```
SSE (Train): 9447794.79
MSE (Train): 420.72
RMSE (Train): 20.51
R² (Train): 0.336
```

Testing Statistics:

```
SSE (Test): 2693435.91
MSE (Test): 479.69
RMSE (Test): 21.90
R² (Test): 0.228
```

The results from our Random Forest Model with the optimized parameters are above. The root mean squared error of this model for the test set was 21.90 and the R^2 value of this model for the test set was 0.228. An R^2 value of 0.228 means that about 22.8% of the variance in popularity of Spotify songs is explained by the predictors in our model. Although this isn’t as high as would be ideal for prediction, the model is a better prediction than random guessing, so this does provide some information about factors that may make a song more popular.

Additionally, we calculated the R^2 of the training set to be 0.336. Since this is similar to that of the test R^2 but still higher than it, we know that overfitting was not an issue.

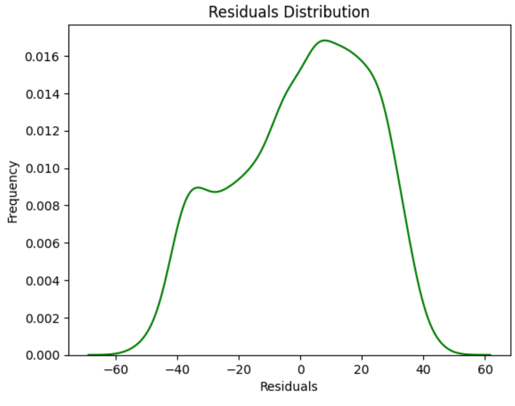
Feature Importance

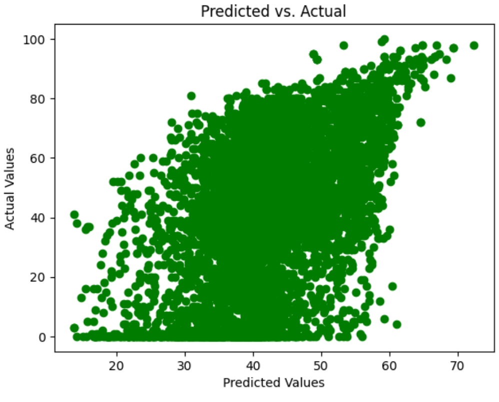
Next, we calculated the feature importance for each variable to see which of our predictors were most useful in predicting the popularity of a song.

```
             Feature  Importance
10              year    0.242562
5   instrumentalness    0.121628
9        duration_ms    0.085787
1             energy    0.079647
2           loudness    0.071333
4       acousticness    0.071176
8              tempo    0.056896
0       danceability    0.053132
3        speechiness    0.046777
7            valence    0.045560
```

The output (for the top 10 most important features) is above. The features that contributed the most to the outcome of the model equation were ‘year’, ‘instrumentalness’, ‘duration_ms’, ‘energy’, and ‘loudness’. So, to answer our previously mentioned research question: the song characteristics that are most useful in predicting the popularity of a song are the year it was made, the instrumentalness, the duration, the energy, and the loudness.

Model Error and Residuals





Looking at the plot to the left above, we can see that the distribution of the residuals is centered around 0. Additionally, note that there is a fairly weak trend between the predicted and actual values for song popularity, as shown in the scatterplot to the right. While we can see that there is a trend between these two values, there does seem to be a substantial dispersion, particularly for mid-range and low popularity scores.

There seems to be bias towards low popularity in the model. Many of the songs are clustered at low popularity levels with few songs accurately achieving a high popularity score. This makes sense because few songs become a “hit,” which would be the ones with extremely high popularity scores. It is interesting, however, that our model appears to be more accurate for predicting the songs that are extremely popular, but not the low-to-mid-range songs. Therefore, this model could be useful in situations in which a music producer or artist is working on a song that they hope to become extremely popular. The Random Forest Model, especially the features mentioned previously as the most important, will help to predict popularity of a Spotify song.