Spotify Song Popularity Analysis

Haohang Yan & Jordan Tian

Introduction

According to Counterpoint Research[1], Spotify has accounted for 34% of the market share for music subscriptions. The revenue of such music platforms is linked directly to their ability to make predictions about the user likeability and generate successful recommendations. Besides, we are also undergoing an age of so called music industrialization[2] - a period of deskilling and reliance on studio technology. If we could find the relation between auditory features and popularity, music producers might produce their music reversely to achieve guaranteed popularity. With these potential benefits to the music industry, we aim to investigate the relation between auditory features and popularity and how popularity could be predicted from them.

Data Description

The Song Popularity Dataset is obtained from Kaggle[3], originally extracted from the public Spotify API[4]. Each row of the dataset represents a song and its auditory features. The original dataset contains 18835 songs and their 15 features in total: 3 are categorical(key, audio_mode, time_signature), 11 are continuous(popularity, duration, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo, audio valence), and 1 is string(name). Descriptions of these features:

1. **Song_name**: name of the song
2. **Song_popularity**: song rating from spotify audiences
3. **Song_duration_ms**: duration of the song in milliseconds
4. **Acousticness**: a confidence measure from 0.0 to 1.0 of whether the track is acoustic. (e.g. sound is produced without electrical equipment)
5. **danceability**: a confidence measure from 0.0 to 1.0 of whether the track is suitable for dancing.
6. **energy**: a perceptual measure of intensity and activity from 0.0 to 1.0
7. **instrumentalness:** a confidence measure from 0.0 to 1.0 of whether the track is playing live
8. **key:** the overall key of the song in standard pitch class notation (e.g. 0 = C, 1 = C♯/D♭, 2 = D, …)
9. **liveness:** a confidence measure from 0.0 to 1.0 of whether audiences are present in the song.
10. **loudness:** the average loudness in decibels (dB) of the song
11. **audio_mode:** the modality (major or minor) of a track. 1 represents major and 0 represents minor
12. **speechiness:** a confidence measure from 0.0 to 1.0 of whether spoken words are present in the song
13. **tempo:** the overall estimated tempo(speed or pace) of the song in beats per minute (BPM)
14. **time_signature:** the overall time signature(beats in each bar) of a track
15. **audio_valence:** a confidence measure from 0.0 to 1.0 of whether the song conveys musical positiveness(e.g. happiness, cheerfulness)

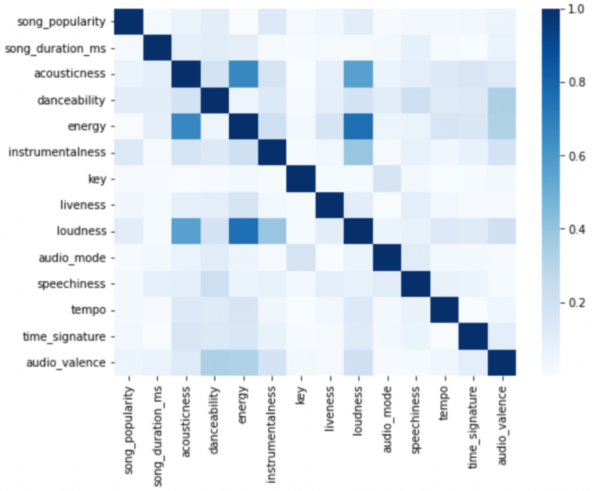
Data Cleaning

There is no missing value but duplicate rows in the dataset. There are 14894 rows left after the removal of duplicates. Due to the large number of outliers, instead of median imputation, we cap these outliers in respect to the 10th and 90th percentile of each variable to maximize retention of the original dataset (datapoints outside the 10th and 90th percentile are imputed with the 10 and 90th percentile). For classification models, we converted song popularity into a binary variable with 0 as non-popular and 1 as popular by doing a median split. As a result, class 0 has 7381 entries and class 1 has 7513 entries. Hence we don't have any issues with data imbalance. Categorical variables are also encoded into dummy variables.

Inference

Which two features are the most correlated?

A correlation matrix can be used to visualize the linear relationships between the variables and identify which pairs of variables are highly correlated. We make a correlation matrix that shows the pairwise correlation between features and find the pair of features with the highest correlation coefficient. After constructing the correlation matrix, we can use a heatmap(Figure 1) to get a clear visualization of the correlation coefficient.



Results

Loudness and energy are the most correlated pair with a correlation coefficient of 0.776.

Are songs that are more danceable more popular than those are less danceable?

We split the songs by the median of danceability. The ratings below the median of danceability are considered less danceable and the ratings above the median of danceability are considered more danceable. The alpha we choose is 0.005. We calculate the effect size and perform a power analysis between these two groups of ratings. The result is 0.996 so we are unlikely to commit a type II error. Before choosing which test we use, we performed Kolmogorov-Smirnov (KS) test on both song popularity and danceability to check for normality. The p-values for KS test on popularity are both less than 0.005, so we reject the null hypothesis that data are normally distributed and claim that song popularity and danceability are not normally distributed. Then we check on the independence. The correlation coefficient is 0.104, indicating that song popularity and danceability are independent from each other. To test the significant relationship between two independent and non normally distributed data, we choose the nonparametric Mann-Whitney U test and make a null hypothesis that more danceable songs are not more popular than less danceable songs.

Results

Mann-Whitney test results in a p-value of 5.16e-29, less than 0.005, indicating that we reject the null hypothesis and conclude that more danceable songs are more popular than less danceable songs.

Prediction

Can song popularity be predicted from the individual features, while controlling confounds key, audio mode, and time signature of the song? If so, which one is the strongest?

Before we began model building, we checked the presence of multicollinearity within the predictors and ran VIF analysis with a threshold of 10 to reduce the variables to a smaller set of acousticness, instrumentalness, liveness, speechiness, and audio_valence, along with the confounds. We built linear regression to predict the popularity in respect to the individual feature and confounds encoded in categories, and evaluate and compare their performance in terms of COD and RMSE on the testing set.

Results

While instrumentalness has a relatively higher performance, there is no evidence that any of the individual features could be a significant predictor for predicting the popularity since the COD are extremely small, which means, only a minimal portion of the variance of popularity is explained by these features. Since these individual features are not insightful, we are now curious in investigating the second question.

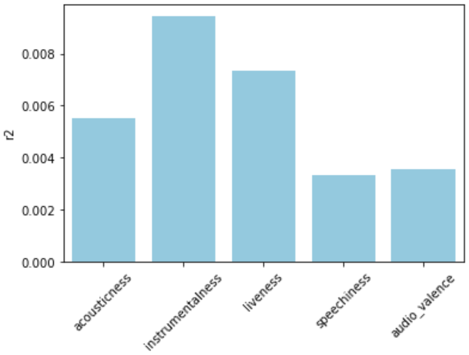


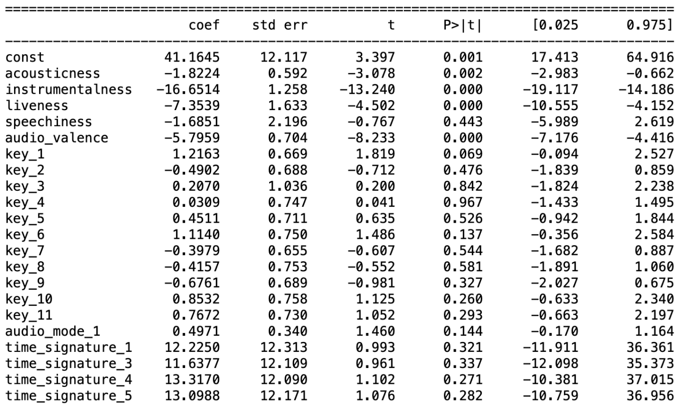
Figure 2: COD of Individual Models

Can song popularity be predicted from features of the song altogether, while controlling confounds key, audio mode, and time signature of the song?

We approached this question by building linear regressions with all features as predictors with confounds encoded in categories and song popularity as target, and thus making conclusions from the significance of the predictors and model performance.

Results

The model summary of the linear regression without regularization is shown below. Acousticness, instrumentalness, liveness, and audio valence have displayed significant non-zero coefficients, which suggests that they could be significant predictors for popularity (Figure 2). However, the overall performance of the linear regression is still bad with a low COD of 0.0170 on testing set - only 1.7% of the variance is explained by the model. By observing the scatterplot (Figure 3) of the predicted popularity and the actual popularity, we found that the predicted values are clustered around the mean of the popularity, which suggests that this model is no better than simply using the mean as the predicted value.



```
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const             41.1645     12.117      3.397      0.001      17.413      64.916
acousticness      -1.8224      0.592     -3.078      0.002      -2.983      -0.662
instrumentalness -16.6514      1.258    -13.240      0.000     -19.117     -14.186
liveness          -7.3539      1.633     -4.502      0.000     -10.555      -4.152
speechiness       -1.6851      2.196     -0.767      0.443      -5.989       2.619
audio_valence     -5.7959      0.704     -8.233      0.000      -7.176      -4.416
key_1              1.2163      0.669      1.819      0.069      -0.094       2.527
key_2             -0.4902      0.688     -0.712      0.476      -1.839       0.859
key_3              0.2070      1.036      0.200      0.842      -1.824       2.238
key_4              0.0309      0.747      0.041      0.967      -1.433       1.495
key_5              0.4511      0.711      0.635      0.526      -0.942       1.844
key_6              1.1140      0.750      1.486      0.137      -0.356       2.584
key_7             -0.3979      0.655     -0.607      0.544      -1.682       0.887
key_8             -0.4157      0.753     -0.552      0.581      -1.891       1.060
key_9             -0.6761      0.689     -0.981      0.327      -2.027       0.675
key_10             0.8532      0.758      1.125      0.260      -0.633       2.340
key_11             0.7672      0.730      1.052      0.293      -0.663       2.197
audio_mode_1       0.4971      0.340      1.460      0.144      -0.170       1.164
time_signature_1  12.2250     12.313      0.993      0.321     -11.911      36.361
time_signature_3  11.6377     12.109      0.961      0.337     -12.098      35.373
time_signature_4  13.3170     12.090      1.102      0.271     -10.381      37.015
time_signature_5  13.0988     12.171      1.076      0.282     -10.759      36.956
```

Figure 3: Linear Regression Model Summary

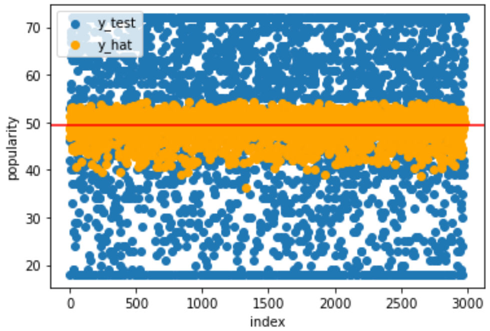


Figure 4: Linear Regression y_test and y_hat

When regularizations are applied to the regression, the performance in terms of COD and RMSE does not improve (Table 1).

	R^2	RMSE	Optimal λ
Linear Regression	0.0170	288.6432	-
Ridge Regression	0.0176	288.4500	$\lambda = 10$
Lasso Regression	0.0170	288.6356	$\lambda = 0.01$

Table 1: Linear Regressions Performances

Rather than predicting the actual popularity of the song, we then approach the question by predicting whether the song is popular or not using logistic regression. The AUC score, or the ability to accurately predict the popularity standing, is 0.5666, which is not significantly better than a random guess. This conclusion can also be reached by observing the ROC curve (Figure 4), which is pretty much in alignment with the diagonal line.

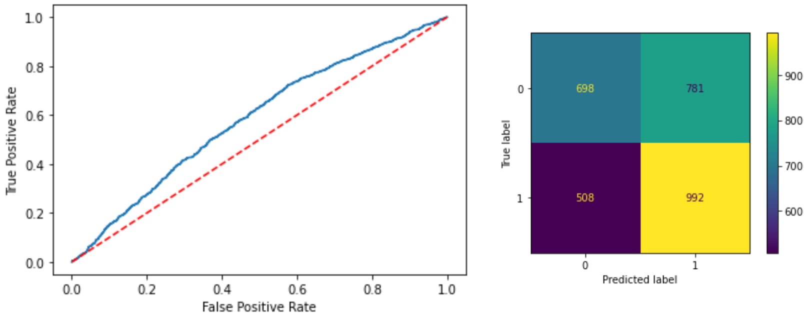


Figure 5: Logistic Regression ROC Curve & Confusion Matrix

In conclusion, song popularity cannot be reasonably predicted from its features given the low performance of all the models.

Classification

Can features of a song successfully classify whether the song is popular or not?

We started with dimension reduction using PCA. As shown in Figure 6, horizontal axis and vertical axis are principal components 1 and 2 respectively. Then, before we do supervised classification, we use the K-Means to do clustering to see whether we can split the data points into 2 clusters. In the end, XGBoost algorithm was implemented for classification. Using the same random seed, we did a 70/30 train-test-split and implemented the grid search cross validation to do hyperparameter tuning.

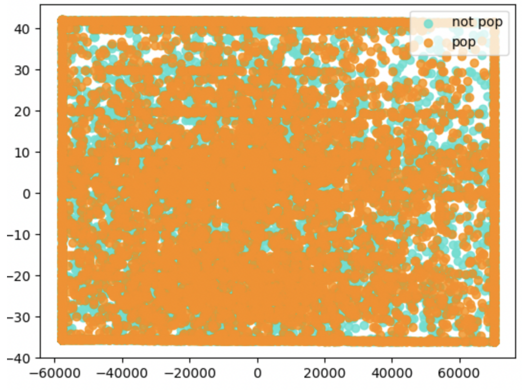


Figure 6: PCA result

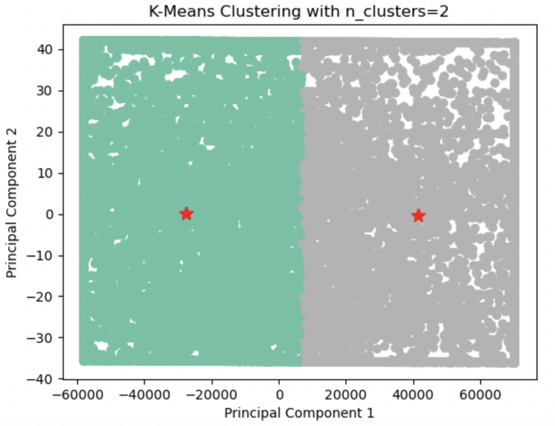
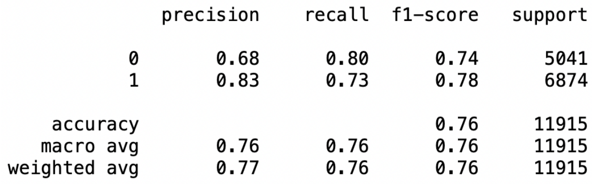


Figure 7: K-Means clustering result

Results

We can see that there is no obvious cluster in Figure 6. All popular and non-popular songs mixed together. Also, the first principal component explains over 97 percent of variance. For the K-Means we should only have 2 clusters, one is popular and the other is non-popular, so we set n_clusters to 2. Figure 7 is what we have for clustering. We can see that it just evenly splits the whole data points into two clusters, which is clearly wrong compared to the real clusters we have earlier. Thus we can conclude that it would be hard to do classification on this dataset. With a three-fold cross validation, we found the best parameters which achieved an accuracy of 0.76 in the training set (Table 2), and 0.59 in the testing set (Table 3). From the table we found that class 1 was better predicted than class 0 in both training and testing data since class 1 has higher precision and f1-score. Figure 8 shows the confusion matrix of the classification results, and Figure 9 shows the ROC curve with an AUC value of 0.62. We can see that the type two error is pretty big, which means we are very likely to fail to reject the null hypothesis. But we still have better accuracy than the result of logistic regression in the previous section. Thus, we can conclude that it is better to use a classification method for this dataset to predict the song popularity than using a regression method, but still, the models do not have great performance. We may need more information in this dataset to predict the song's popularity.

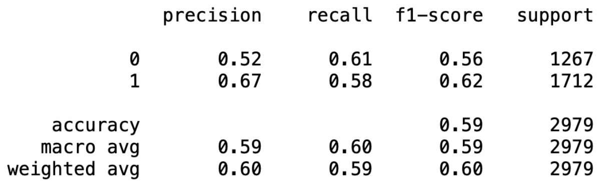


```
              precision    recall  f1-score   support

           0       0.68      0.80      0.74      5041
           1       0.83      0.73      0.78      6874

    accuracy                           0.76     11915
   macro avg       0.76      0.76      0.76     11915
weighted avg       0.77      0.76      0.76     11915
```

Table 2: Classification Report for Train Set



```
              precision    recall  f1-score   support

           0       0.52      0.61      0.56      1267
           1       0.67      0.58      0.62      1712

    accuracy                           0.59      2979
   macro avg       0.59      0.60      0.59      2979
weighted avg       0.60      0.59      0.60      2979
```

Table 3: Classification Report for Test Set

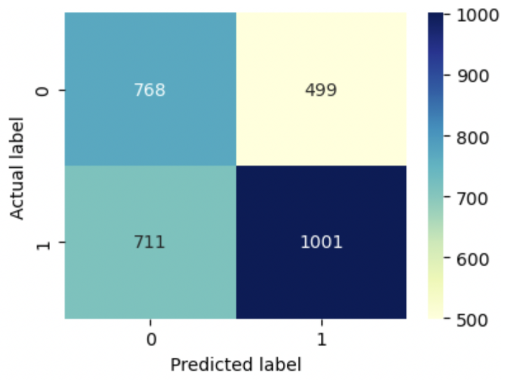


Figure 8: Confusion matrix

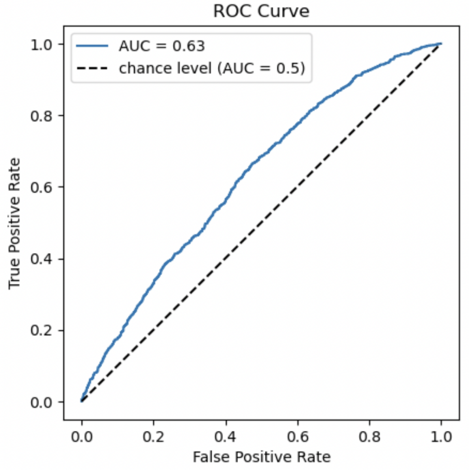


Figure 9: ROC curve

Conclusion

After making statistical inferences, we found some correlations between some of the auditory features. Among those features, we found that some features with high values could lead to a higher popularity (eg. songs that are more danceable are more popular than those that are less danceable). Among all the models in section 2 and 3, the classification model performs the best given its higher accuracy rate. It is thus easier to predict the classes of popularity than the number itself. However, while looking at the accuracy rates and the AUC values, we still could not be optimistic about the classification model. The AUC score for the classification model is 0.62, which is just a little higher than the accuracy rate for the logistic model. Therefore, the classification model could be a better model for predicting the popularity of a song but it is still far away from being a satisfactory model with significantly high accuracy and low RMSE that could fulfill the ultimate goal of prediction.

Integrating these results, we are almost certain that the relationship between auditory features of the song and its popularity is limited, and it is not reasonable to make predictions about the popularity and give recommendations solely based on the "physical" features of the song. Also, "engineering music" reversely seems unrealistic at this point.

As stated in the data cleaning, we used the capping method instead of median imputation to deal with the outliers because there are too many of them. This could inflate the coefficients of the predictors and cause potential problems to the accuracy of our models. Other than limitations that threaten the conclusion we drew, we would also like to address the limitations with the dataset itself. With a high test power and no shortage of rows and features, the relationship we found between features and popularity is still limited. This is very likely due to the fact that the nature of music is highly subjective and not quantitative, and there are so many more variables and confounders that we did not have a chance to consider. Given the status quo of popular music, a large portion of the influence and popularity is in direct relation to the singer, the social message of the song, and the socioeconomic status of the singer's company. If we were given more information on these criterions, a more powerful and careful prediction is very probable. If we want to expand our question of interest into accurate prediction, an ideal dataset needs to be not only well-collected without bias but also needs to include all potential features that could have an effect on the dependent variable.

References

[1] Person, Simon, & Frith. (2017, July 5). The industrialization of popular music: 7 : Taking Popular Music Seriously. Taylor & Francis. Retrieved December 20, 2022, from https://www.taylorfrancis.com/chapters/mono/10.4324/9781315087467-7/industrialization-popular-music-simon-frith

[2] Kumar, A. (2020, October 6). Global online music streaming growth slowed down in Q2 2020. Counterpoint Research. Retrieved December 20, 2022, from https://www.counterpointresearch.com/global-online-music-streaming-growth-slowed-down-in-q2-2020/

[3] Dataset https://www.kaggle.com/datasets/yasserh/song-popularity-dataset

[4] Spotify WEB API https://developer.spotify.com/documentation/web-api/reference/#/