ST443 Group Project Part 1

December 10, 2020

1 Executive Summary

Despite the impact of COVID-19 on the Economy, the music industry continues to flourish at a positive growth rate (5.6% in terms of recorded music revenues in the US) especially streaming music according to *the 2020 Mid-Year Music Revenue Report* | *RIAA*. What makes a song a hit is a question that arose a lot of interest and has been studied various times from different perspectives.

The purpose of this report is to predict the popularity of soundtracks on Spotify based on quantitative measures such as danceability, key, etc. Popularity is a figure ranging from 0 to 100 calculated by the algorithm mainly focusing on how many times a particular soundtrack is played and how recent it is played, To investigate what makes a hit song and predict the popularity, various machine learning techniques are used including linear models, non-linear models, tree-based methods and neural networks. Among these models implemented, the generalized additive model (GAM) gives the best performance with MSE around 32.92

2 Data Source

The dataset was uploaded to Kaggle in 06/2020 by *Yamaç Eren Ay*, who used the Spotify Web API for developers to build a data that contains more than 160,000 songs. Each row in the data represents a unique track, identified by a unique ID feature generated by Spotify. The columns are 19 features of the tracks including acousticness, artists, danceability, duration_ms, energy, explicit, id, instrumentalness, key, liveness, loudness, mode, name, popularity, release_date, speechiness, tempo, valence and year (Figure 1 in the Appendix).

After shuffling the index, the original data set was divided into 2 parts—70% for training and 30% for testing.

3 Empirical Results and Analysis

3.1 Linear Regression

Three linear models are used to predict the popularity of songs on Spotify including linear regression, lasso, and ridge regression. Among these linear models, linear regression gives the lowest MSE, which is 47.074 as shown in the table. As for lasso and regression, the MSE for testing datasets are relatively larger than that for linear regression considering the penalty term in the form of $\lambda \sum_{j=1}^{15} |\beta_j|$ and $\lambda \sum_{j=1}^{15} \beta_j^2$ respectively. Additionally, the linear regression model with variables automatically selected by lasso returns a similar MSE (47.262) to the fundamental linear model.

Model	Linear Regression (with all features)	Linear Regression (with only statistically significant variables	Lasso	Ridge	Linear Regression (with variables selected by lasso)
MSE	47.074	47.104	74.049	237.644	47.262

According to the linear model with variables selected with the lasso (Fig 2 in the Appendix), which is simpler in terms of inference and performs well with 86% of the variation explained by the selected features. The type of music and the release year play a significant role in

determining the popularity of the song (Fig 2). For instance, acoustic songs are overall less popular than non-acoustic ones with other features remaining the same, shown by the negative relationship between acousticness and popularity. Accousticness is a confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic, based on Spotify Audio Features for a Track. On the other hand, danceability is positively correlated with popularity, which indicates the popularity of dance music. What's more, as popularity is calculated by algorithm and is based partially on the total number of plays the track has had and how recent those plays are, songs released in recent years are generally more popular than those that were released long ago. This can be proven by the positive relationship between year and popularity.

The performance of the linear model with variables selected has been further assessed with plots in Fig 3 and Fig 4. The linearity holds reasonably well according to the Residuals vs Fitted plot as the red line is horizontal at 0, but the linear model's performance, especially when it comes to prediction, is questionable when the popularity is too either too small or too large as shown in Fig 4. This can be partially explained by the heavy tail of residuals in the Normal Q-Q plot in Fig 3. Apart from that, the linear model's performance is quite good with residuals equally spread along with fitted values and relatively small leverage in most cases.

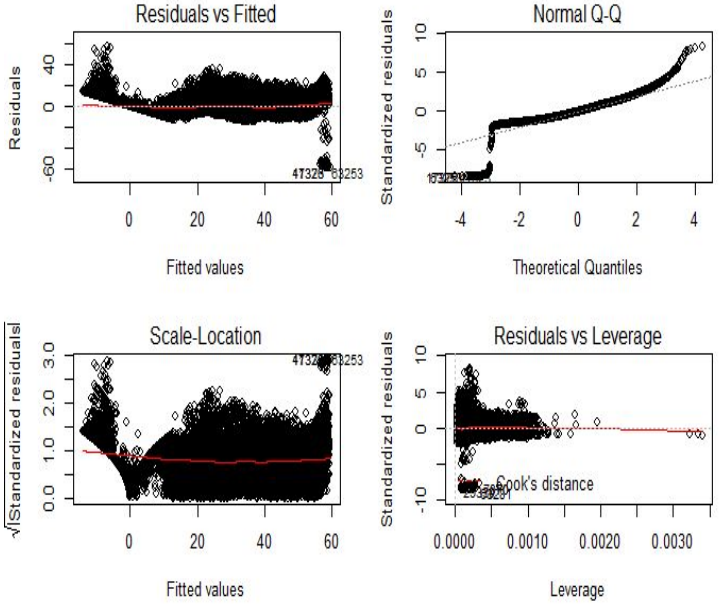

Figure 3 Plots of linear regression

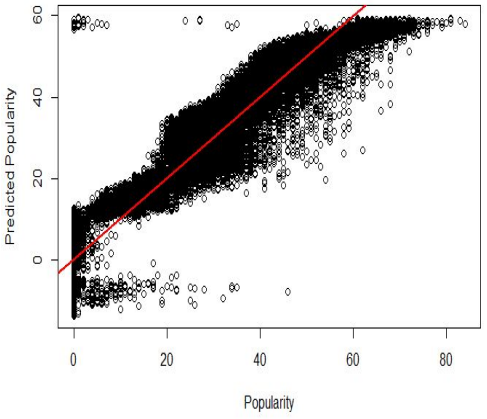

Figure 4 predicted pop vs pop

3.2 Non—Linear Model

3.2.1 Tree Based Model

Firstly, further data processing is applied. Duration_ms, Instrumentalness, and Speechiness, which originally are numerical features, but considering their practical meanings of the value range in tree-based models and also EDA, are transformed to categorical features. Duration_ms was changed to minute-unit which is a more common pattern, and was divided into two categories by "5 min". Due to the consideration of extreme values, the instrumentalness was turned into 3 intervals. And the speechiness was separated by a seeming breakpoint—0.62.

Then, 2 classic tree-based models are used to make a prediction. Model 1 is a regression tree with a default parameter setting. The result is quite surprising, because only 1 feature—year, is used in tree construction, which shows it much more significant and dominant. It shows that for each node only using the feature 'year' can maximum the information gain. It is reasonable because the popularity under the current situation has a great correlation with the songs' 'age'. It

seems to be overwhelming compared to other features. People tend to listen to newersongs than older ones. The prediction ability is not good visually, especially since the predicted values are several fixed values due to the model limitation from branch number, and the MSE is 92.49778. (Fig 5 and Fig 6).

Model 2 is a random forest model. The prediction ability is much better visually, which shows an obvious 45-degree upward trend and more diversified predicted values. (Fig 7) And the MSE is 83.69294, performing better than the regression tree model. Unsurprisingly, the feature “year” is most important. However, some other features like loudness, danceability, energy, valence, acousticness and tempo have an obvious effect on popularity, which makes sense intuitively. For instance, some rock music using a loud voice, dynamic melody and energetic beats to convey power will inspire people or make a hot atmosphere to attract a wide range of listeners. (Fig 8)

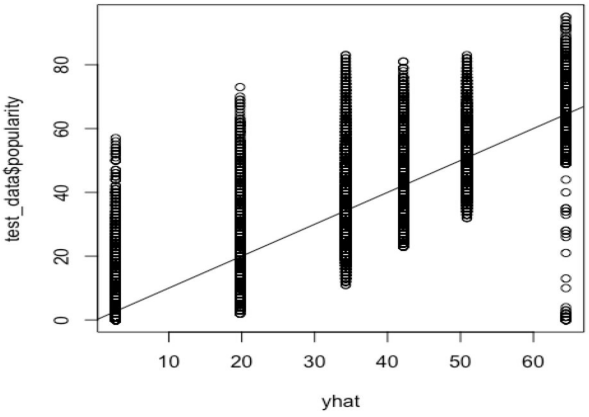

Figure 6 Predicted and Real Values of DT

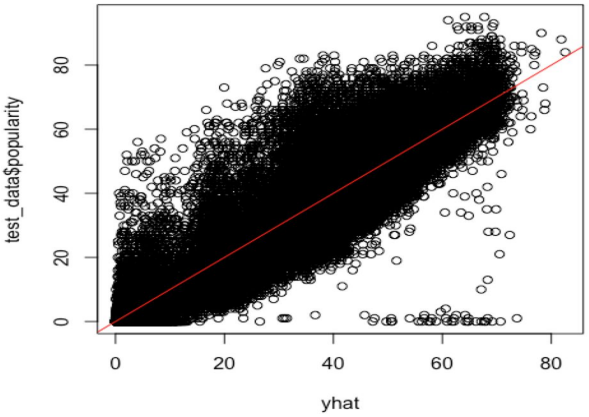

Figure 8 Predicted and Real Values of RF

4.2.2 Neural Network

The first step is also the specific data processing. For neural network, all the input features should be numerical and also share the same dimension. So all the features are scaled at the interval [0,1] through normalization. Under the setting that 1 hidden level, 5 neurons, 1e+07 stepmax and 0.1 threshold (Fig 9), the output network concludes 3 layers, the black line shows the connection between each layer and the weight on each connection, while the blue line shows the deviation term added in each step, and the deviation can be thought of as the intercept of the linear model. But the network is just like a black box without robust explanatory ability. The model performance also shows a 45-degree upward trend in the Predicted and Real Values Figure, but the predicted results are always smaller than the actual ones (Fig 10). Then the predicted values need to scale back to [0, 100] and are used to calculate MSE, which is 247.5759 perhaps because of the accumulation effect of the slight underestimation for many samples. The high error could also be also due to the number of times we back propagated the results, and the size of the training and validation sets. It could also be due to the loss function and the activation functions used on the hidden nodes.

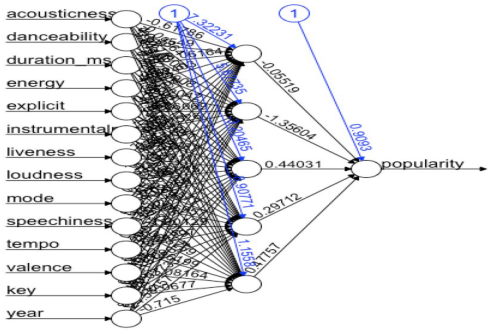

Figure 9 Trained Neural Network

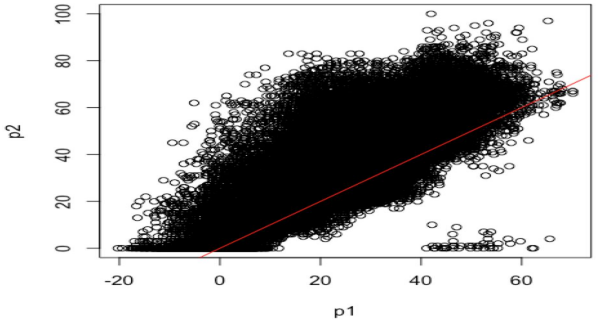

Figure 10 Predicted and Real Values

3.3 Generalized Additive Model (GAM)

With the complexity of the relationship between year and popularity in mind (shown in Fig 11), a generalized additive model in the form of $y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_5(x_{i5}) + \varepsilon_i$ is used where the model used for year is natural spline with $df = 6$ by trials and errors while the linearity of other covariates remains the same. This GAM gives the lowest MSE (32.92) among all the models implemented. Besides, GAM performs better when dealing with popularities above upper quantile and below lower quantiles compared with linear regression with selected variables as shown in the boxplots in Fig 12 and Fig 13.

4 Conclusion

The dataset with little errors contains many samples with diversified professional indicators of music. It can be used in predicting the popularity of music, a practical function, which is useful guidance for musician creation and also provides a reference for music propaganda. And also some research on popularity based on such dataset can help to decide which songs will be played in advertising. Before modeling, proper EDA and real cases research are the foundation for feature selection and categorical variable construction. The next step is applying the dataset in different models (OLS / Lasso / Ridge / Regression Tree / Random Forest / Neural Network / GAM) to get an optimal one for popularity prediction.

For linear models, if only considering the accuracy of the predicted values, the linear model with all the features will be chosen because of its lowest MSE. However, both model practical explanatory ability and model simplicity should be evaluated standard, then the linear model with variables selected with the lasso is the final decision because it only keeps the variables with significance and practical meanings. From the model, acousticness and speechiness show a negative relationship with popularity, and danceability, loudness, and year have a positive relationship with popularity. Intuitively, the fact is that people less prefer acoustic music with high acousticness and speech-like recording with high speechiness (e.g. talk show, audiobook, poetry). Also, this generation enjoys powerful and dynamic music with high danceability and loudness. Keeping with fashion and following the mainstream proves the positive effect of the feature ‘year’.

For non-linear models, GAM has the best performance which absorbs the information from other models. But just as other models, the GAM model tends to underestimate the values especially in the case of extreme fitted values, and also has the problem that output contains the negative value.

But limited by computer hardware handling such scale data (like computing ability and running speed), more complex models like advanced neural networks and XGboost have not been tested. To improve single non-linear model performance, methods like simple looping, grid search, random search and Bayesian Optimization can be equipped for searching optimal parameters of the model. Meanwhile, at the level of model practical effect in business scenario, there is also the possible negative cycle if using these models as reference or guidance to create more and more of the same music (as they're popular due to the model output) but then this huge increase in supply of this type of music makes the market saturated and everyone starts to hate this music as it's become TOO popular.

Appendix

```
── Data Summary ────────────
                           Values
Name                       data
Number of rows             169909
Number of columns          19
_________________________
Column type frequency:
  character                4
  numeric                  15
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
1 artists               0             1     5   661     0    33375          0
2 id                    0             1    22    22     0   169909          0
3 name                  0             1     1   255     0   132940          0
4 release_date          0             1     4    10     0    10882          0

── Variable type: numeric ────────────────────────────────────────────────────────
   skim_variable    n_missing complete_rate        mean          sd    p0        p25          p50        p75       p100 hist
 1 acousticness             0             1       0.493       0.377     0      0.0945       0.492      0.888      0.996 ▅▃▂▃▇
 2 danceability             0             1       0.538       0.175     0      0.417        0.548      0.667      0.988 ▁▃▇▇▂
 3 duration_ms              0             1  231406.     121322.     5108 171040       208600     262960     5403500   ▇▁▁▁▁
 4 energy                   0             1       0.489       0.267     0      0.263        0.481      0.71       1     ▅▅▅▆▇
 5 explicit                 0             1       0.0849      0.279     0      0            0          0          1     ▇▁▁▁▁
 6 instrumentalness         0             1       0.162       0.309     0      0            0.000204   0.0868     1     ▇▁▁▁▂
 7 key                      0             1       5.20        3.52      0      2            5          8         11     ▆▆▇▃▅
 8 liveness                 0             1       0.207       0.177     0      0.0984       0.135      0.263      1     ▇▃▂▁▁
 9 loudness                 0             1     -11.4         5.67    -60    -14.5        -10.5       -7.12      3.86  ▁▁▁▂▇
10 mode                     0             1       0.709       0.454     0      0            1          1          1     ▃▁▁▁▇
11 popularity               0             1      31.6        21.6       0     12           33         48        100    ▇▆▅▂▁
12 speechiness              0             1       0.0941      0.150     0      0.0349       0.045      0.0754     0.969 ▇▁▁▁▁
13 tempo                    0             1     117.         30.7       0     93.5        115.       136.       244.   ▁▆▇▃▁
14 valence                  0             1       0.532       0.262     0      0.322        0.544      0.749      1     ▅▅▅▆▇
15 year                     0             1    1977.         25.6    1921   1957         1978       1999       2020   ▁▂▃▅▇
```

Figure 1 Data Overview

```
Call:
lm(formula = popularity ~ acousticness + danceability + loudness +
    speechiness + year, data = music_train)

Residuals:
    Min      1Q  Median      3Q     Max
-58.940  -5.008  -0.698   4.332  56.660

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept) -1.360e+03  3.481e+00 -390.713  < 2e-16 ***
acousticness -6.787e-01  1.256e-01   -5.405  6.5e-08 ***
danceability  3.453e+00  2.103e-01   16.422  < 2e-16 ***
loudness      2.192e-02  7.740e-03    2.833  0.00462 **
speechiness  -3.877e+00  2.510e-01  -15.443  < 2e-16 ***
year          7.013e-01  1.737e-03  403.766  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.973 on 42204 degrees of freedom
Multiple R-squared:  0.8697,    Adjusted R-squared:  0.8697
F-statistic: 5.634e+04 on 5 and 42204 DF,  p-value: < 2.2e-16
```

Figure 2 Result of linear regression with selected variables

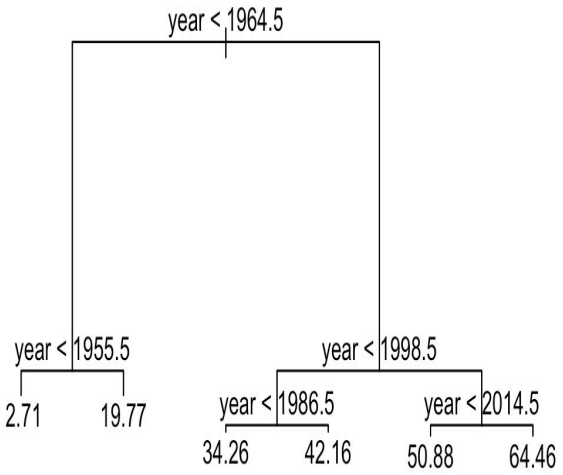

Figure 5 Decision Tree

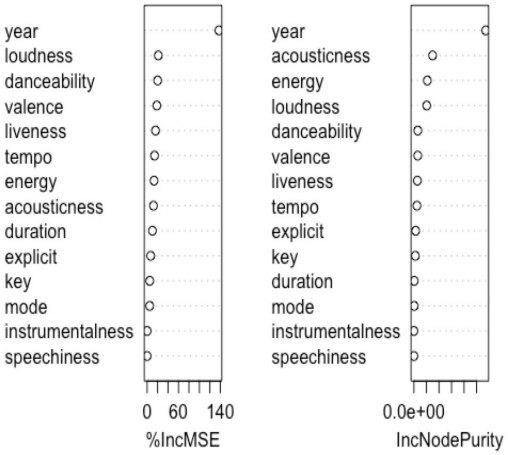

Figure 8 Feature Importance

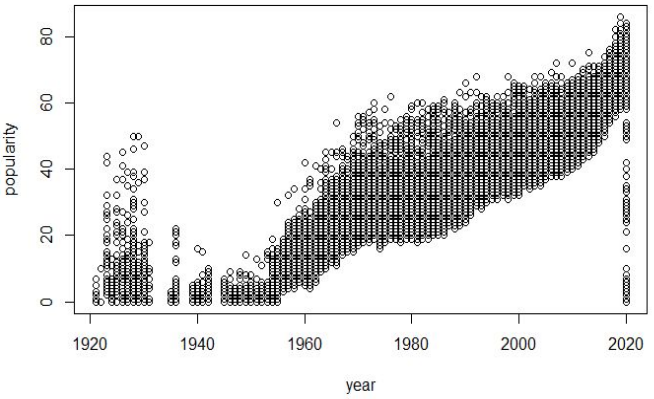

Figure 11 Scatter plot of popularity vs Year

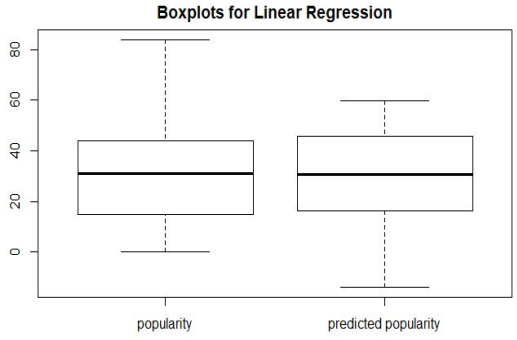

Figure 12 Boxplot of Linear Regression

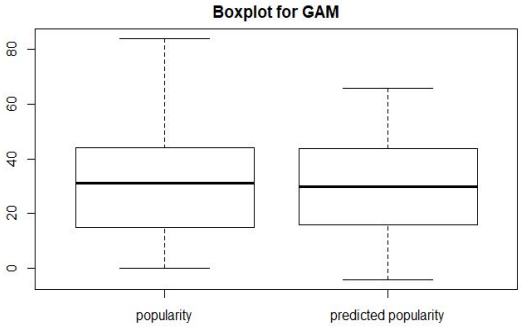

Figure 13 Boxplot of GAM