Music

Team members: Yukie Wu, Sixiang Yuan, Luoxi Tang

1. Description of research questions / issues

What factors have an impact on the popularity of a song?

2. Description of the data

These datasets are from the CORGIS library. The data was collected by NSF to promote further research into the Music Information Retrieval field. It has a total of 10,000 songs. The response variable is the popularity of a song in December 2010. The independent variables that will be used in this research are artists' popularity, familiarity with the artists, genre of the song, how frequent the artists perform this genre, Average start time of each bar, Duration of the track in seconds, Time of the end of the fade in, at the beginning of the song, Estimation of the key the song is in, General loudness of the track, Average start time of each tatum, measured in tatums, Average start time of each beat, Start time of the fade out, in seconds, Tempo in BPM, and Year when this song was released. For this research, We exclude the data points that have missing values and some variables that weren't able to be interpreted. To reduce some categories in genre, we only choose the genre that the number of the songs of this genre is larger than 50, which reduces the sample size to 953 songs.

3. Statistical analysis

1) Methods

Linear regression: Check linearity between dependent and independent variables. There are lots of categorical variables and insignificant variables. Therefore, we used Variance Inflation Factor(VIF) to eliminate the insignificant variables.

PCR/PCA: To calculate the principal components and then use some of these components as predictors in a linear regression model fitted using the typical least squares procedure. However, we weren't able to reduce some variables as predictors. And for PCA, we can only use numeric variables, which can cause this model performance to be worse.

LOOCV and K-fold CV: To estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Regression Tree and Classification Tree: In a regression tree, we can see how a regression model fits to the target variable using each of the independent variables. In addition, we can use random forests to see which variables are more influential to a song's popularity. For a more direct explanation of the model, or to answer our research question, we change our response variable to a binary variable to make an easier decision on how the independent variables affect a song's popularity. Both trees are able to do feature selection.

LDA, QDA, Naives Bayes, SVM: After we change the response variable to a binary response. We use these methods for classification predictive models. Then compare the test errors to choose a better method to explain this data.

KNN: We use KNN method to determine how the songs are separated into several classes and how they will be classified.

K-means clustering: We use this method to classify the data points to see which cluster they belong to.

2) Results

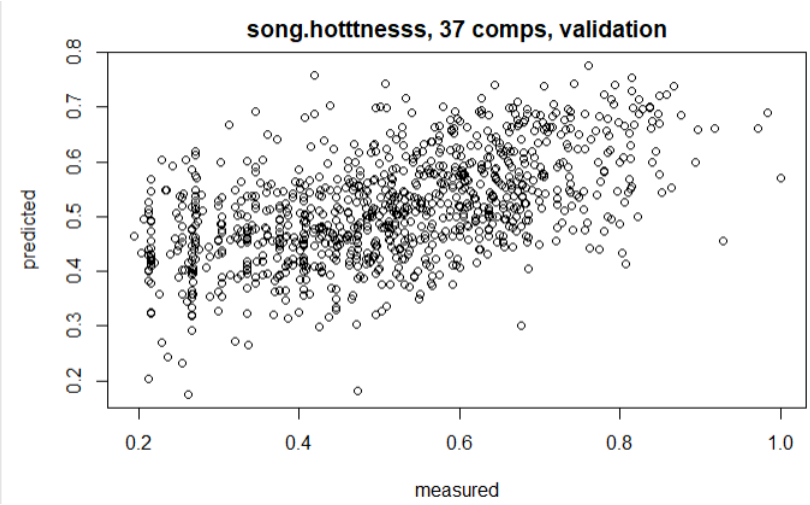

Figure 1.

From Figure 1, the plot performed by PCR showed how the predicted value behaves compared to actual measurements. The test error on the PCR model is 2%.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7
Standard deviation     1.5407 1.2908 1.0627 1.0093 0.99310 0.90810 0.85810
Proportion of Variance 0.2374 0.1666 0.1129 0.1019 0.09863 0.08247 0.07363
Cumulative Proportion  0.2374 0.4040 0.5169 0.6188 0.71742 0.79989 0.87352
                           PC8     PC9    PC10
Standard deviation     0.75664 0.68833 0.46740
Proportion of Variance 0.05725 0.04738 0.02185
Cumulative Proportion  0.93077 0.97815 1.00000
```

Figure 2.

In Figure 2, from the result obtained from PCA, we need at least 8 principal components to be able to explain more than 90 percent of the data. We can also see that the standard deviation for each principal component is large.

We then perform both LOOCV and 10 K-fold CV in the same regression model,

```
Resampling results:

  RMSE       Rsquared   MAE
  0.1375098  0.2988591  0.1091711
```

```
Resampling results:

  RMSE       Rsquared   MAE
  0.1380519  0.2968214  0.1098369
```

Figure 4.

From Figure 4, we can see that the Residual mean squared error, R-squared, and mean absolute error for both methods are pretty close.

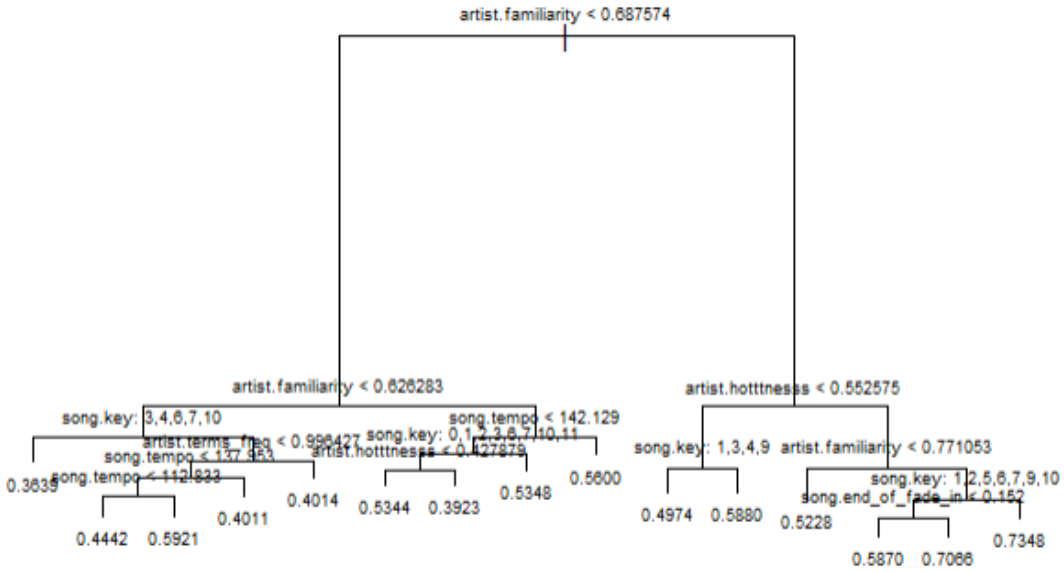

Figure 5.

From this regression tree, the test error is 0.1544546. To improve this regression tree (Figure 5), we use random forests to get the optimal tree size. From Figure 6, it shows that tree size of 3 has the lowest deviance.

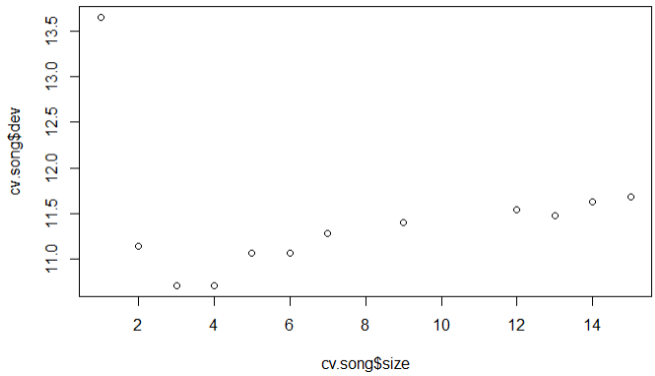

Figure 6.

In addition, we perform bagging on the same tree and then compare the test errors for the three regression trees. When using bagging, the test error is 0.13978 and 0.13802 for random forests. Besides this, we decide to see the importance of the variables, which shows that an artist's popularity and familiarity with the artists are the two most important variables.

Then, we decided to do a classification tree, shown as Figure 7, on whether the song is popular or not by setting the popularity of the song that's greater than 0.5 as popular. And then we prune the tree into size two(Figure 8).

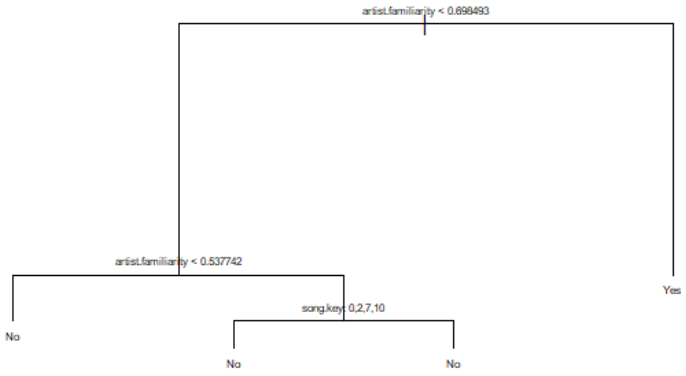

Figure 7.

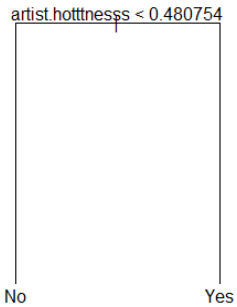

Figure 8.

We then can interpret the tree as for artists' popularities that are less than 0.480754, the songs will not be popular.

To determine the popularity of the song, we performed LDA, QDA, Naives Bayes, and SVM. The test errors are 0.2830189, 0.3228512, 0.2914046, and 0.4884696 (cost=0.01) or 0.3165618 (cost=10), respectively.

Next, for KNN method, we only use three variables: Popular as response variable, and artist's popularity and familiarity with the artists as independent variables. From Figure 9, we can estimate the test error is 0.3165618 at K=53.

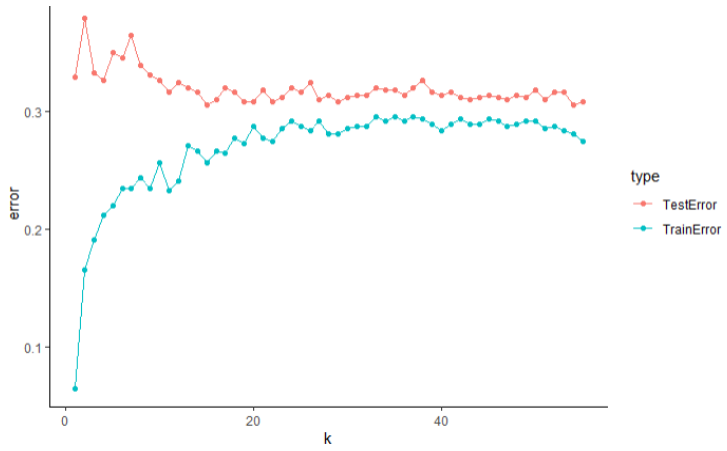

Figure 9.

Lastly, we performed K-means clustering on the same three variables.

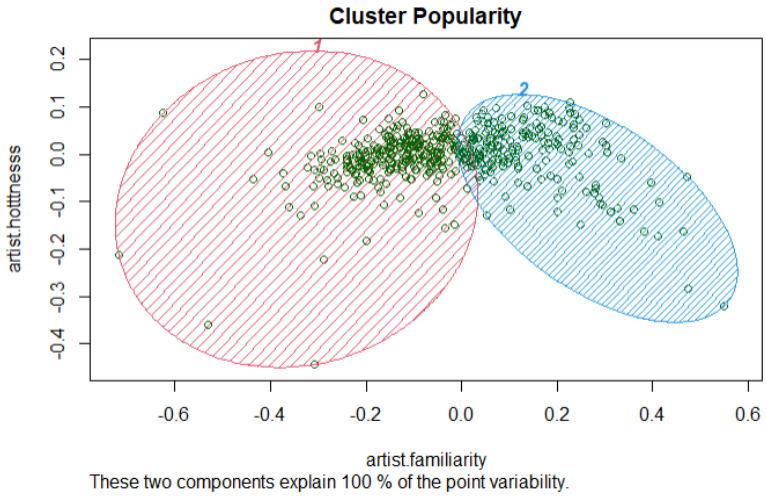

Figure 10.

From Figure 10, the red circle is the cluster that means it is not popular and the blue circle is the cluster means it is popular. But it only accurately explains 70% of the data.

Methods	**Test Error**
PCR	0.002
Regression Tree	0.1544546
Bagging	0.13978
Random Forest	0.13802
LDA	0.2830189
QDA	0.3228512
Naives Bayes	0.2914046
SVM	0.4884696 (cost=0.01) or 0.3165618 (cost=10)
KNN	0.3165618

3) Discussion/Conclusion

From the research, we see that for the regression model, PCR gives the lowest test error, while for classification model, LDA gives the lowest test error. Also, we can conclude that the most influential factors that determine the popularity of the songs are from the artists themselves. The results show that the popularity of the songs are determined by how popular the artists are and how people are influenced by and familiar with the artists. However, this research still has some limitations. For example, we only choose songs with more general genres which can cause bias. Also, the accuracy of each method used for classification is not high.