Final Project Report

Rylan Keniston, Jaeho Jang

5/12/2020

Overview

The data being analyzed describes musical characteristics of songs streamed in 2017, containing 13 attributes ranging from acoutisness, tempo, and danceability, as well as a target variable which describes whether a specific Spotify user liked the song or not. With this data set, we wanted to see if we could use these song attributes to predict whether a song is liked by the Spotify user or not.

Method

- Linear Model

First, we conducted two different regression models to see which regression would predict the target variable most accurately. Looking at the confusion matrix of the linear model, we were able to conclude the average percent over 200 instances. The data set was split into two sets; a training set that contains a sample of 80% of song observations and a testing set that contains the other 20%.

```
##                Predicted Dislike Predicted Like
## Actual Dislike              135             64
## Actual Like                  73            131

## When predicting whether a song is liked or disliked, our linear regression model has a percent error of 34 %.
```

- Stepwise Linear Model

Moreover, we also looked at a stepwise linear model and also derived its percent error through a confusion matrix, as we believed that a stepwise selection would give us a more accurate prediction of song preference.

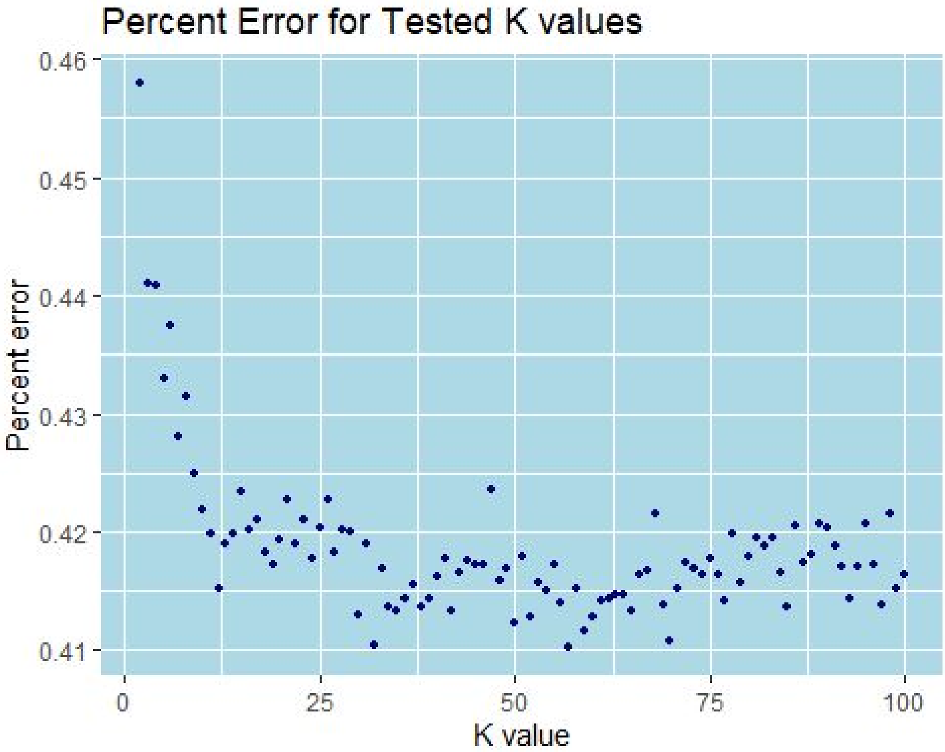
```
##                Predicted Dislike Predicted Like
## Actual Dislike              552            245
## Actual Like                 285            533

## When predicting whether a song is liked or disliked, our stepwise regression model has a percent error of 32.82 %.
```

After running the confusion matrix -as expected - when comparing the two different regression models, the stepwise model seemed to be slightly more accurate with a percent error of 33.09%.

- KNN Model

Next, we used KNN to develop a regression model that predicts whether a song was liked or disliked. To find the number of K nearest neighbors to use, we found the lowest average percent error when the values 2 through 100 were tested 100 times

```
##  The K value of 57 has an average percent error of 41.02 %
```



We used this K value in our KNN model, running the test 200 times to find the true rate at which the model predicts whether a song was liked or not.

```
##                Predicted Dislike Predicted Like
## Actual Dislike              125            73
## Actual Like                 93           111

## Our regression model has an average percent error in classifying whether a song was liked or
disliked for 200 tests using 57 nearest neighbors is 41.29 %
```

Our KNN model only correctly predicted about 60% of the songs correctly, so this model is considered pretty insufficient.

- PCA Model

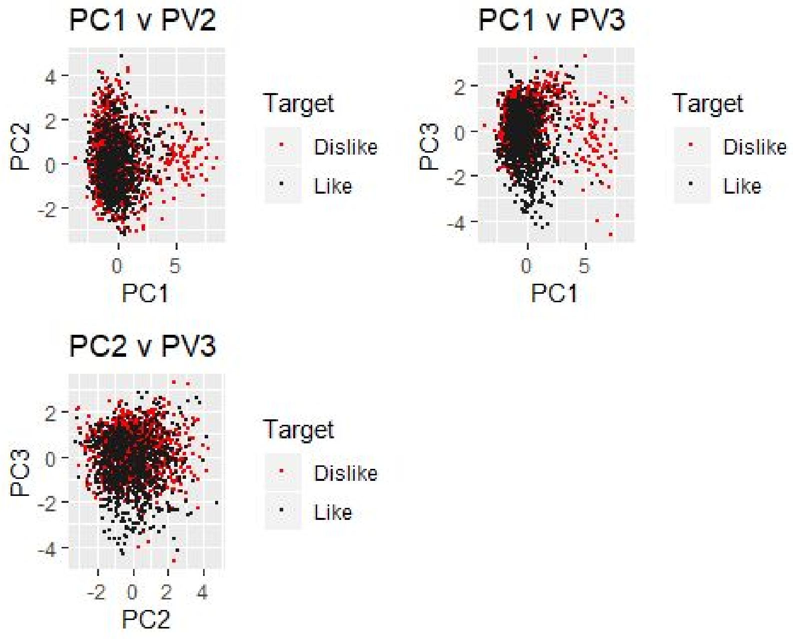
Next, instead of using a regression or clustering to sort the liked and disliked songs in our data, we have created a principal component analysis (PCA) to reduce the number of variables used when describing the data and distinguishing the observations. From the 13 audio variables, the PCA algorithm created 13 new summary variables, named PC1 to PC13.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.6870 1.2351 1.1465 1.06053 1.01840 0.98532 0.94657
## Proportion of Variance 0.2189 0.1173 0.1011 0.08652 0.07978 0.07468 0.06892
## Cumulative Proportion  0.2189 0.3362 0.4374 0.52388 0.60366 0.67834 0.74727
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.89073 0.86636 0.81001 0.73154 0.61950 0.40807
## Proportion of Variance 0.06103 0.05774 0.05047 0.04116 0.02952 0.01281
## Cumulative Proportion  0.80830 0.86603 0.91650 0.95767 0.98719 1.00000
```

Each summary variable is a linear combination that maximizes the amount of variability retained from the original data. Combined, the first five of our summary variables explain about 60% of the variation in our data. The linear combinations of these components are:

```
##                      PC1     PC2     PC3     PC4     PC5
## acousticness      0.4430 -0.1379  0.1425 -0.2108 -0.0636
## danceability     -0.1730 -0.5888 -0.2312  0.0399 -0.0015
## duration_ms       0.2006  0.0780 -0.5256  0.2881 -0.1044
## energy           -0.4962  0.2265 -0.0681  0.2160  0.0905
## instrumentalness  0.2572  0.2229 -0.3833  0.2927  0.0476
## key              -0.0753  0.0857 -0.4186 -0.4103  0.1738
## liveness         -0.1195  0.3641 -0.0717  0.1153 -0.5083
## loudness         -0.5039  0.1490  0.1124  0.0911  0.1217
## mode              0.0652 -0.0643  0.5077  0.4086 -0.1282
## speechiness      -0.1520 -0.0858  0.0048 -0.4494 -0.6453
## tempo            -0.1184  0.3326  0.0798 -0.3959  0.2117
## time_signature   -0.1889 -0.1628 -0.1740  0.1459 -0.4045
## valence          -0.2653 -0.4653 -0.1049  0.0419  0.1730
```

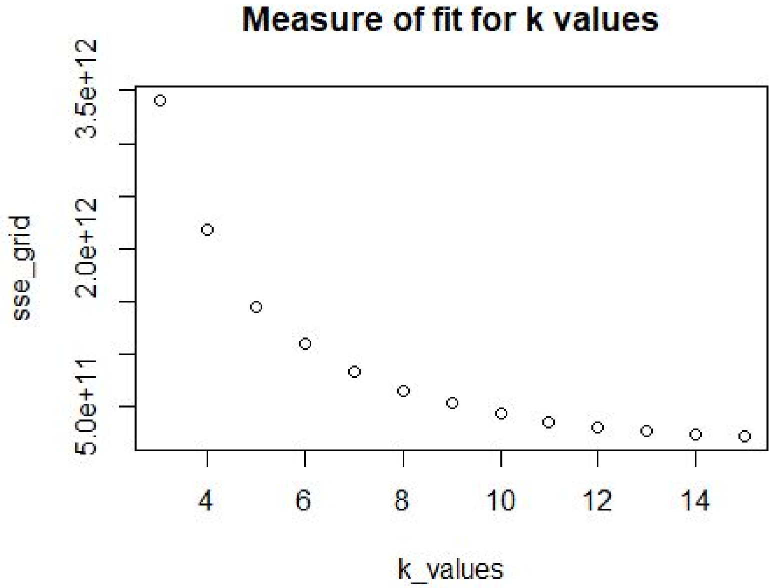
To understand how well PCA performs at identifying similar songs, we looked at plots of our top three summary variables.



In all three of the plots, it is hard to determine exactly how well our principal components analysis was able to distinguish the liked from the disliked songs. From looking at the first two plots, we can see that the disliked songs tend to have more variation in the PC1 component. We can also see from the plots that liked songs are centrally given values in between 0 and -2 for PC2. For the PC3 component, liked songs tend to have more variation in the negative direction than disliked songs. Although none of the plots show substantial clustering, the PC1 v PC3 plot seems to distinguish the songs the best.

- Clustering Model

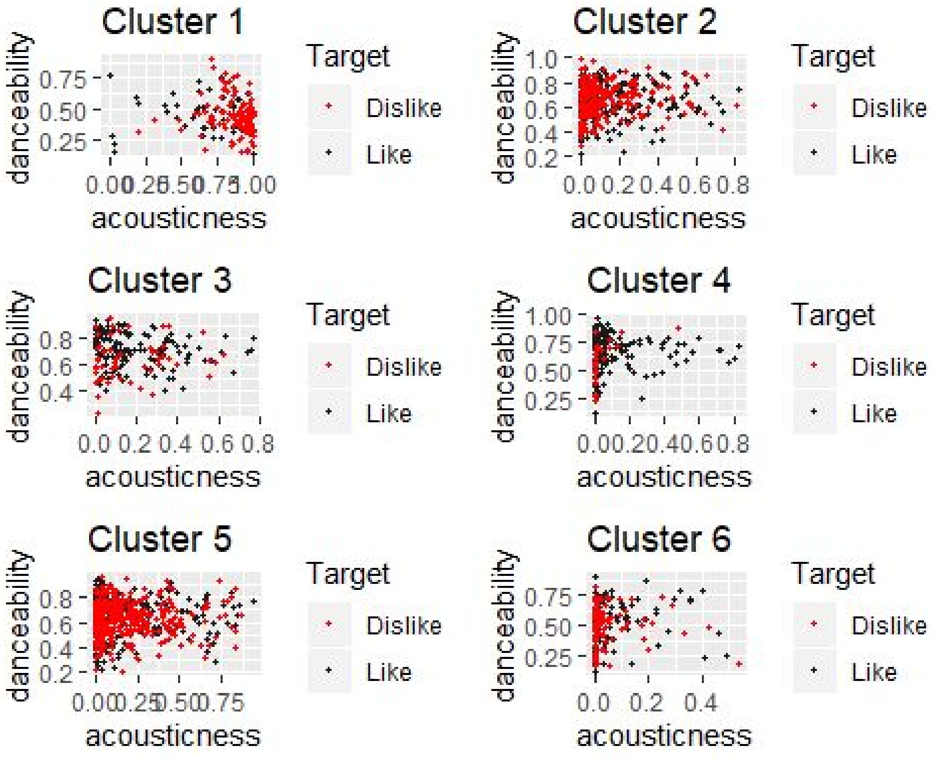
Lastly, we used clustering to determine if it was able to actually differentiate between ‘liked’ songs and ‘not liked’ songs with given attributes. We chose to use the K-means++ method, as the method uses bias of distance when choosing the starting centroid points, therefore, reduces final cluster errors. To find out how many k numbers of clusters will best fit the data, we looked at the plot of trial k's and their associated SSE's.



The plot showed that our 'elbow' point was located at around k =6, and therefore, we chose to create 6 clusters. After running the song data through the K-means++ model using 6 clusters, each cluster's average attribute value was found.

```
##   acousticness danceability duration_ms energy instrumentalness   key  liveness
## 1         0.80         0.16        0.23   6.28             3.99  0.02    125.83
## 2         0.45       -14.86        0.13   0.15             0.55  5.20      4.02
## 3    290530.95         0.71        0.65  -6.04             0.17  0.19      0.53
## 4         0.22         0.04   237611.54   0.00             0.69 -7.18      0.10
## 5         0.35       112.56        0.73   0.07        228415.97  0.62      0.66
## 6         4.64         3.71        0.03 121.95             0.66  0.32 325974.40
##   loudness   mode speechiness  tempo time_signature valence
## 1     0.73   0.06   228359.22   1.00          0.51   -4.73
## 2     0.72 124.14        0.69   0.06      235110.26    0.66
## 3     5.52   3.98        0.02 118.49          0.85    0.10
## 4     0.16   0.49        4.65   3.99          0.13  131.55
## 5    -8.10   0.16        0.14   0.54          6.04    3.97
## 6     0.53   0.64       -6.43   0.05          0.48    0.41
```

To see if clustering the data could actually distinguish ‘liked’ songs and ‘disliked’ songs, single clustered were generated with colors that determined whether they were liked or not.



As seen from the plots, all the plots seemed to have a pretty dominant saturation out of the two ‘target’ choices. This tells us that our K-means++ performs well when predicting if a song was liked or not.

Conclusion

In conclusion, after running various models regarding the Spotify streaming data, we can say that our stepwise and linear regression models performed the best to predict the ‘likeness’ of a song with given attributes. On other hand, we can also conclude that our PCA and KNN models performed the worst. This may be due to the fact that the target “like” variable comes from a single Spotify user’s opinion, personal preference outside of the given variables, such as genre or lyric types, may contribute largely in predicting whether the user liked or disliked a song, which cannot be taken into consideration by the models.As PCA and KNN models look into effectiveness of each attributes, there may be attributes in the data set not included to make to the models as accurate.