Research skills: Programming with R

Group Project:

Spotify data analysis using R

Introduction

For this research, we used a dataset called "Spotify - All Time Top 2000s Mega Dataset" from Kaggle, which contains audio features of Spotify top 2000 tracks from years between 1956 and 2019. The dataset consists of 15 columns describing each track. It has information about the track's name, artist, genre, year, length in seconds, actual popularity (scale 0-100) and information about audio features of the track (BPM, energy, danceability, loudness, liveness, valence, acousticness and speechiness). Descriptions of these audio features can be found on Appendix A.

It is known that the musical characteristics are significant in contributing to one's musical preference, so the main research question we want to formulate is: **"How do audio features of a track influence its popularity?"**.

Nowadays, due to the high importance of streaming on the music industry revenues, the music producers have higher incentives in making shorter tracks. That's because in order for the music creator to secure payment for a stream on Spotify, the track has to be played for at least 30 seconds, so with shorter songs, the artists can obtain more clicks, avoid track skipping and make more money. So we will also test the hypothesis that **the average length of the songs released after the creation of Spotify (in 2008) is shorter than the songs released before Spotify exists**.

Finally, we also study **what are the main audio features that differentiate music genres**.

Data Preprocessing and Exploratory Data Analysis (EDA)

Since the original dataset had a huge number of different genres, it was decided to group those very specific genres into broader categories in order to improve the readability of our data. To do that, we made use of the website Music Genres and also our own judgements. Our group ended up with 15 main genres, which can be found in Appendix B.

As the first part of the EDA, we calculated the average duration of songs released before and after the year of Spotify creation, to test our hypothesis. We found that the songs released before 2008 have on average 267 seconds and the songs released during and after 2008 have on average 249 seconds (Appendix A) so it reinforces the validity of our hypothesis.

Further exploring the data, we saw that the most frequent genres in our dataset were rock and pop music varieties (Appendix A). We also found that popularity of the songs could not be easily explained with the audio features we have and the only metric that has a fairly bigger impact on popularity is "Danceability" (Appendix A).

The third study, defines the genres with audio features, the average BPM for each genre was also examined in our paper (Appendix A). It turns out that Latin and Jazz music has the highest tempo while Country has the lowest.

Statistical Models and Interpretation of Their Results

In order to test our hypotheses about popularity and genres, we tried to fit two statistical models, KNN and Logistic Regression, into our data. We decided to transform the numerical column "popularity" into a binary column, with values lower than 50 labelled as "unpopular" and values higher than 50 labelled as "popular". We also created two train/test splits of 80/20, one that is balanced around the popularity variable for popularity study and another that is balanced around genre for genre study.

KNN

Popularity

First, we created a KNN model to see which audio features predict and/or influence track popularity. We optimized the model for Specificity to ensure that we do not always predict the class that has most observations in the training set. Our accuracy on the test set is 70.03%, with a sensitivity of 85.27% and specificity of 27.62% (See Appendix B). This means that we are much more accurate in classifying whether a track is popular, compared to classifying that it is unpopular, which makes sense given the unbalanced nature of the data.

Furthermore, we look at the variable importance of all the features. Caret's "varImp" function shows that the loudness and danceability of a track are the two most important in determining whether a song is popular or not. On the other side, we see that the BPM and duration of a track has little to no influence on whether or not the track is deemed popular or not.

Genre study

Additionally, we wanted to see if we can predict the genre of a song based on its audio features and we achieved an accuracy of 39.07% on the test set. It seems that it is even harder to predict genre than it is to predict popularity, which can be explained by the fact that Genre is highly unbalanced[1], which makes it hard for the model to correctly classify genres of tracks that it has not seen much data for.

When looking at the variable importance we can clearly see a different picture. BPM, which was one of the least important features for predicting popularity is now the most important feature across the classes. Also, danceability is seen as an important feature for categorizing genre.

[1] For example, there are 604 rock tracks within our training set, while there are only 2 Latin tracks.

Finally, we see that valence and duration are the least important features for categorizing which genre a track belongs to.

Logistic Regression

Next, we created a logistic regression model, using the same binary popularity variable that we had created as a target variable as logistic regression requires binary values as its dependent variables and included only the audio features for each track. The confusion matrix and summary statistics of the model can be found in Appendix B. Optimizing the accuracy on the basis of all possible combinations of audio features, the accuracy of our model turned out to be 73.8%, with a sensitivity of 98.9% and specificity of 0.3%. Similar to the KNN model we are much more accurate in classifying whether a track is popular, compared to classifying that it is unpopular, which makes sense given the unbalanced nature of the data. Due to the multiclass nature of the Genre variable, we did not perform Genre Study with our Logistic Regression model.

Random Forest

Finally, for our third model we decided to use a random forest model and see whether it could improve our accuracy at predicting either popularity or genre. We used a tunegrid between 1 and 20 to find the best performing mtry values which is the "number of variables randomly sampled as candidates at each split." (Breiman and Cutler, 2018). For predicting popularity, mtry at 7 was found to be the best performing value on the train set and the model gave an accuracy of 73.3% on the test set (See Appendix B). For predicting genre, on the other hand, mtry of 1 was found to be the best performing one, but nevertheless the model was only able to achieve 40.6% accuracy on the test set. Consequently, our random forest models were not able to improve our predictions for either target.

Conclusion

With this research, we found that songs released during and after 2008 (year of Spotify's creation) are on average 18 seconds shorter than songs released before this year, which reinforces our hypothesis that the streaming era is encouraging producers in making shorter songs.

We discovered that danceability is the audio feature that is most correlated with song popularity, and this matches with other studies that have been done in this area. However, some audio features show not to have a significant relation with popularity. This can be due to the unbalanced nature of our data, since our dataset has only the top 2000 songs from Spotify, and those songs are not good representatives of all the songs that exist on this platform. We also need to take into account that some of the songs in our dataset are very old, most of them were created before the streaming revolution (Appendix A) and the audio features that are trending the most nowadays are different from those we were trending in the past.

Our data was also unbalanced in genre distribution (Appendix A). However, we saw that some genres differentiate a lot from others in the amount of beats per minute. For example latin music is characterized for having a much higher tempo than country music (Appendix A).

References

- A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

- Breiman, L & Cutler A. (2018). Breiman and Cutler's Random Forests for Classification and Regression . Retrieved June 20, 2020, from https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

- Datta, H., Knox, G., & Bronnenberg, B. J. (2018). Changing their tune: How consumers' adoption of online streaming affects music consumption and discovery. Marketing Science, 37(1), 5-21.

- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

- Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.5. https://CRAN.R-project.org/package=dplyr

- Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

- Music Genres List. (2019). Spotify - All Time Top 2000s Mega Dataset. Retrieved June 18, 2020, from https://www.musicgenreslist.com/

- Nijkamp, R. (2018). Prediction of product success: Explaining song popularity by audio features from spotify data (master's thesis). University of Twente.

- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

- Singh, S. (2020). Spotify - All Time Top 2000s Mega Dataset. Retrieved June 17, 2020, from https://www.kaggle.com/iamsumat/spotify-top-2000s-mega-dataset

- Spotify. (2020). Get Audio Features for a Track. Retrieved June 18, 2020, from https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

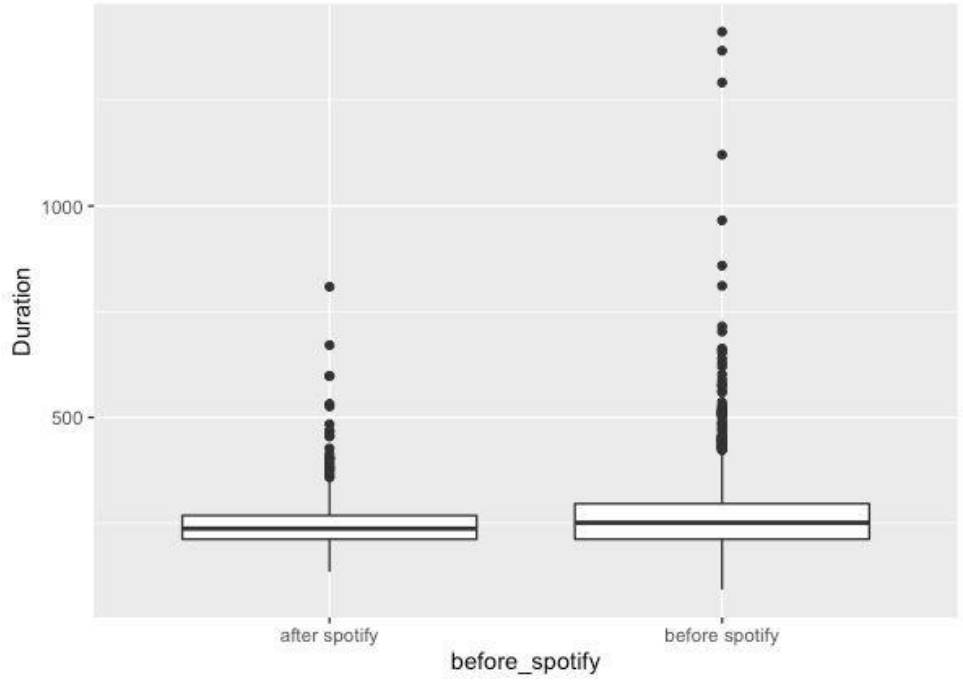
- Teo, T. (2003). Relationship of selected musical characteristics and musical preferences (a review of literature). Visions of Research in Music Education, 3, 1-20.

Appendix A

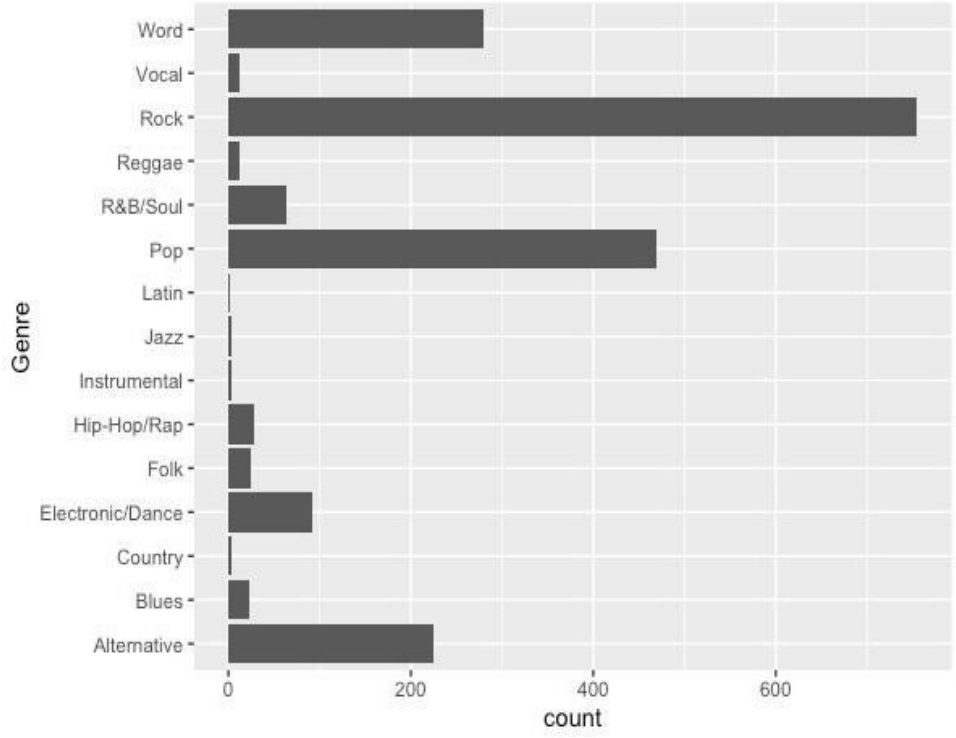
1. Description of the audio features

Audio Feature	Range	Description
BPM	> 0	Track's tempo in beats per minute
Energy	[0; 100]	Measure for intensity and activity of the track (for example death metal is high in energy)
Danceability	[0; 100]	Describes how suitable a track is for dancing
Loudness	typically [-60; 0]	Measures the relative loudness of tracks in decibels and can be also a measure for the quality of the recording (the louder the song the better it is able to communicate emotions)
Liveness	[0; 100]	If its value is above 80 there's a strong likelihood that the track is a live performance
Valence	[0; 100]	Measures how happy a track sounds
Acousticness	[0; 100]	A confidence value that measures if the track is acoustic
Speechiness	[0; 100]	Detects the amount of spoken words in a track (for example rap music has higher speechiness than electronic music)

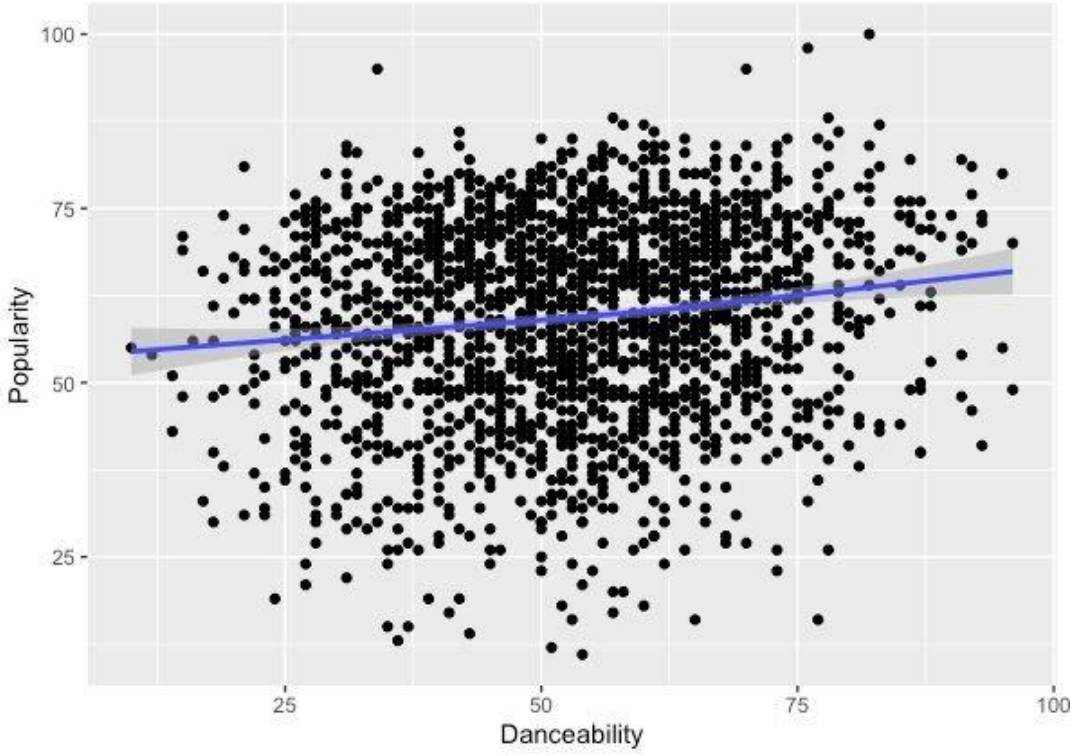
2. Change of Average Song Durations Before and After Spotify



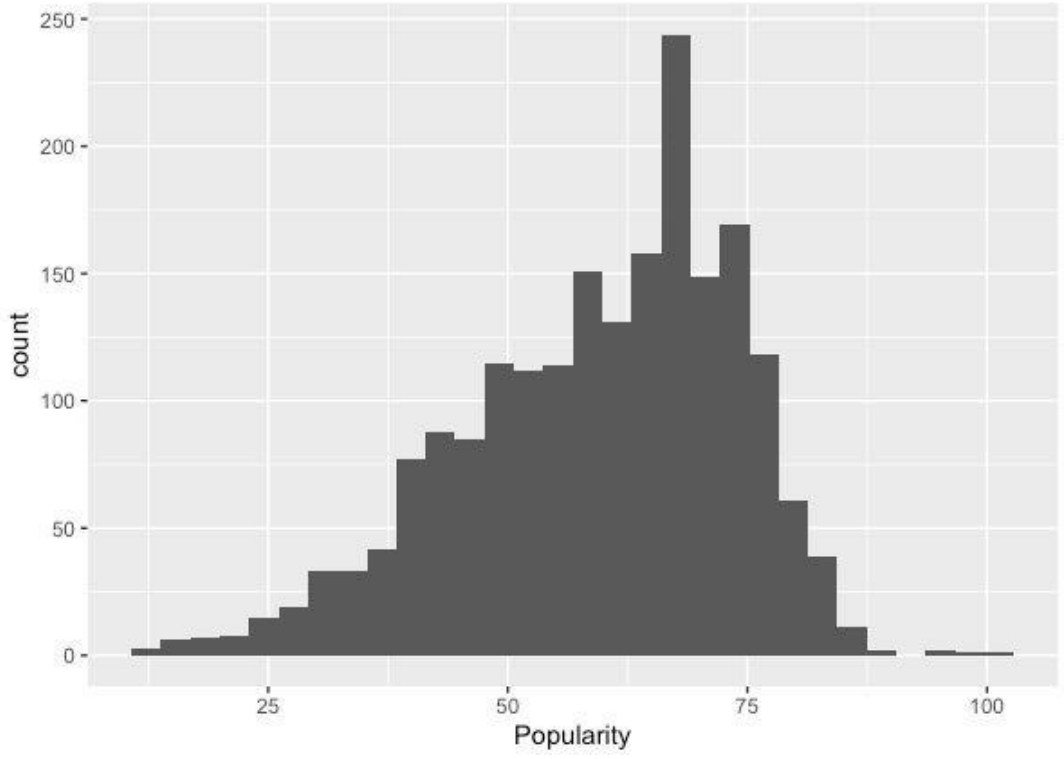
3. Distribution of Genres in the Data Set



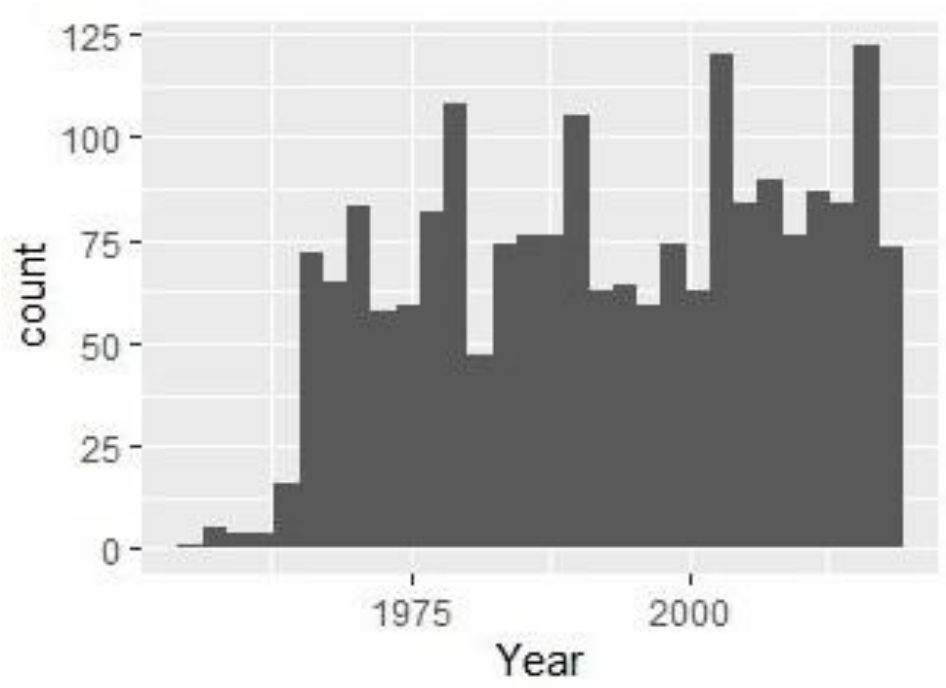
4. Plotting the Relationship Between Danceability and Popularity



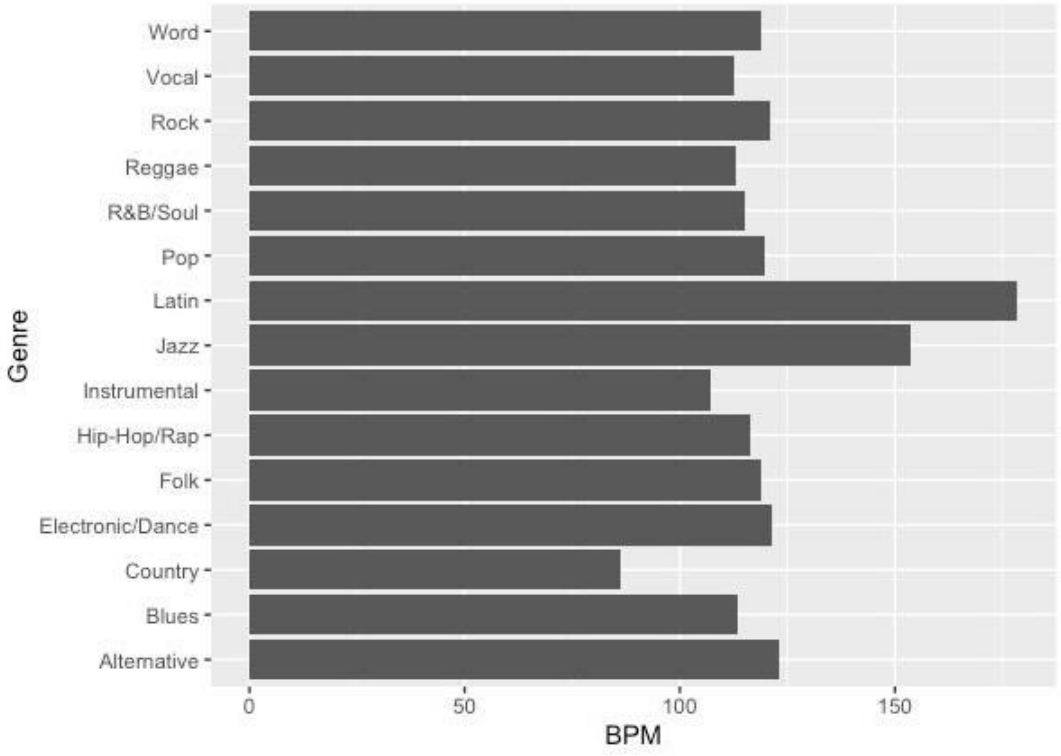
5. Distribution of Popularity



6. Distributed of songs per released year



7. Average BPM for Genres



Appendix B

1. List of Genres

Alternative	Blues	Country	Electronic/Dance	Folk
Hip-Hop/Rap	Instrumental	Jazz	Latin	Pop
R&B/Soul	Reggae	Rock	Vocal	Word

2. Confusion Matrix and the Summary Statistics for KNN Model – Popularity

```
Confusion Matrix and Statistics

          Reference
Prediction  popular unpopular
  popular       249        76
  unpopular      43        29

               Accuracy : 0.7003
                 95% CI : (0.6526, 0.7449)
    No Information Rate : 0.7355
    P-Value [Acc > NIR] : 0.949066

                  Kappa : 0.1434

 Mcnemar's Test P-Value : 0.003352

            Sensitivity : 0.8527
            Specificity : 0.2762
         Pos Pred Value : 0.7662
         Neg Pred Value : 0.4028
             Prevalence : 0.7355
         Detection Rate : 0.6272
   Detection Prevalence : 0.8186
      Balanced Accuracy : 0.5645

       'Positive' Class : popular
```

3. Confusion Matrix and the Summary Statistics of Logistic Regression Model – Popularity

```
Confusion Matrix and Statistics

          Reference
Prediction  popular unpopular
  popular       289       101
  unpopular       3         4

               Accuracy : 0.738
                 95% CI : (0.6918, 0.7806)
    No Information Rate : 0.7355
    P-Value [Acc > NIR] : 0.4809

                  Kappa : 0.0397

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9897
            Specificity : 0.0381
         Pos Pred Value : 0.7410
         Neg Pred Value : 0.5714
             Prevalence : 0.7355
         Detection Rate : 0.7280
   Detection Prevalence : 0.9824
      Balanced Accuracy : 0.5139

       'Positive' Class : popular
```

4. Confusion Matrix and the Summary Statistics of Random Forest Model – Popularity

```
Confusion Matrix and Statistics

          Reference
Prediction  popular unpopular
  popular       278        92
  unpopular      14        13

               Accuracy : 0.733
                 95% CI : (0.6866, 0.7759)
    No Information Rate : 0.7355
    P-Value [Acc > NIR] : 0.5712

                  Kappa : 0.0995

 Mcnemar's Test P-Value : 7.495e-14

            Sensitivity : 0.9521
            Specificity : 0.1238
         Pos Pred Value : 0.7514
         Neg Pred Value : 0.4815
             Prevalence : 0.7355
         Detection Rate : 0.7003
   Detection Prevalence : 0.9320
      Balanced Accuracy : 0.5379

       'Positive' Class : popular
```

5. Random Forest – Mtry and Accuracy Chart for Predicting Popularity