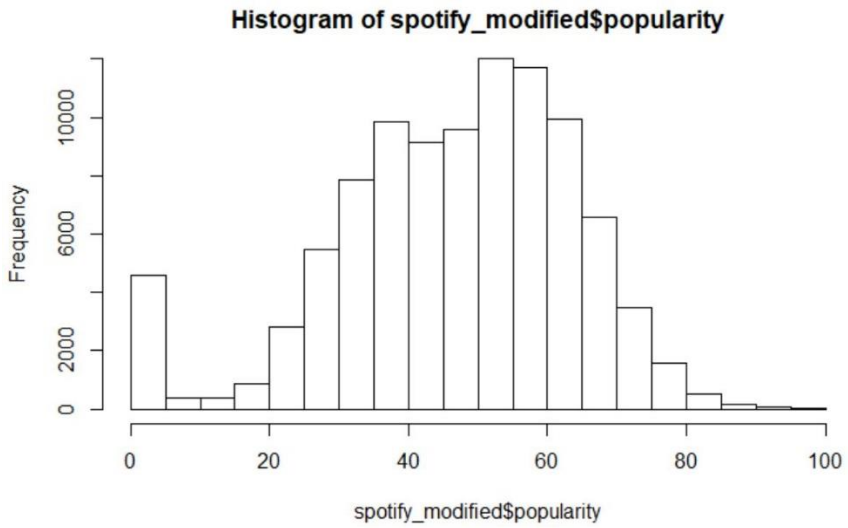
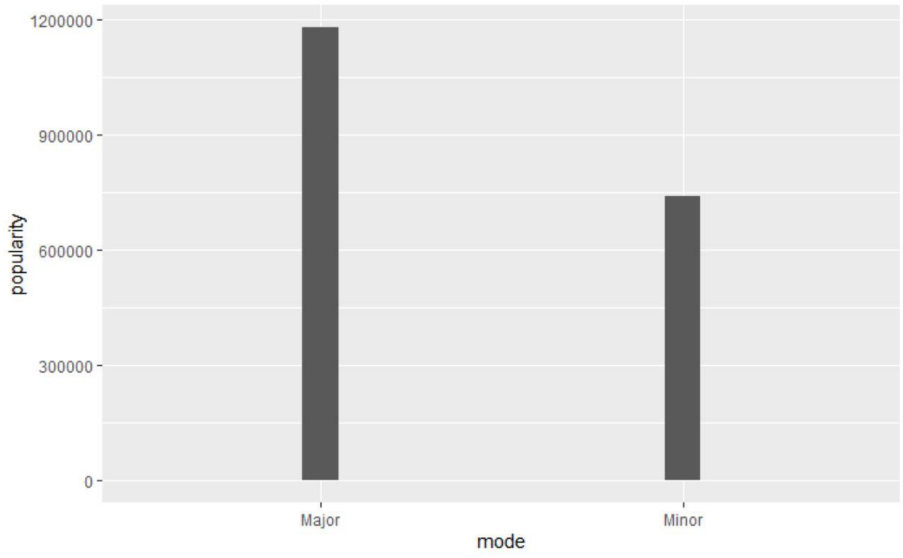
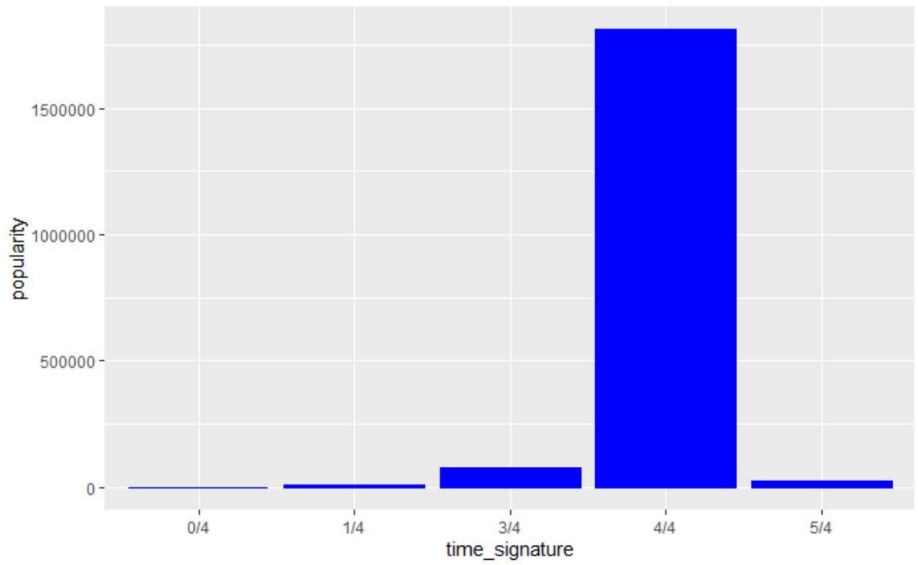
PROJECT REPORT – SONG POPULARITY PREDICTOR



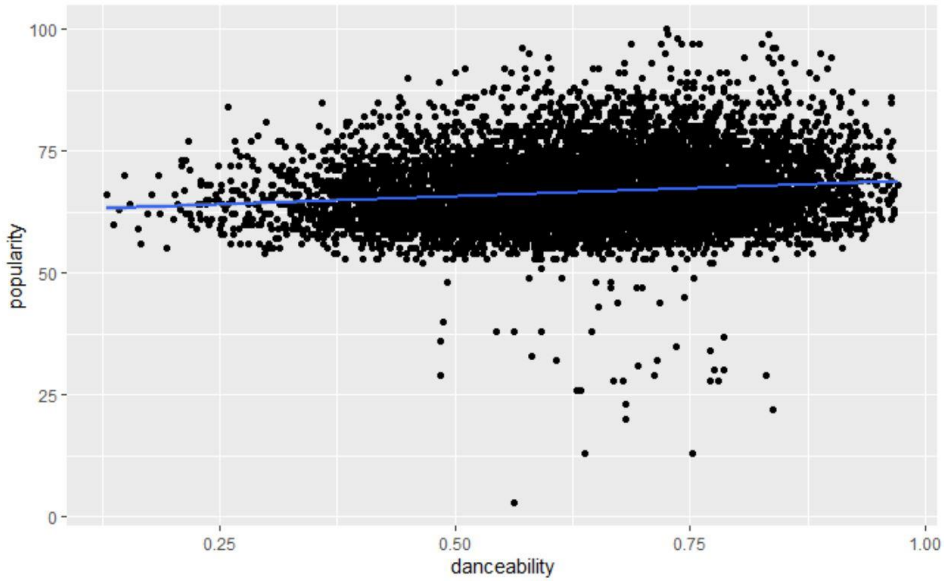
Distribution of songs based on Popularity (Histogram): Suggests that major distribution of dataset is slightly above popularity score 50



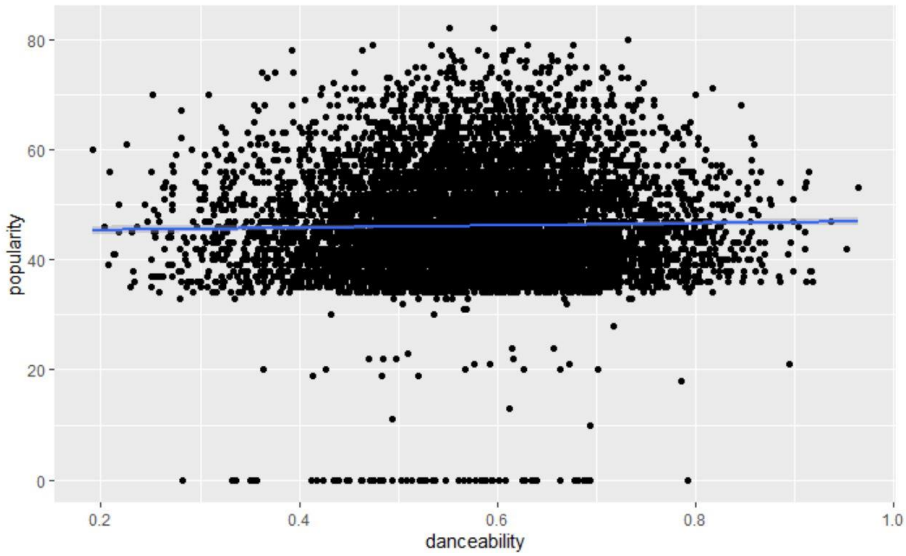
Distribution of songs based on Mode (Bar Plot)



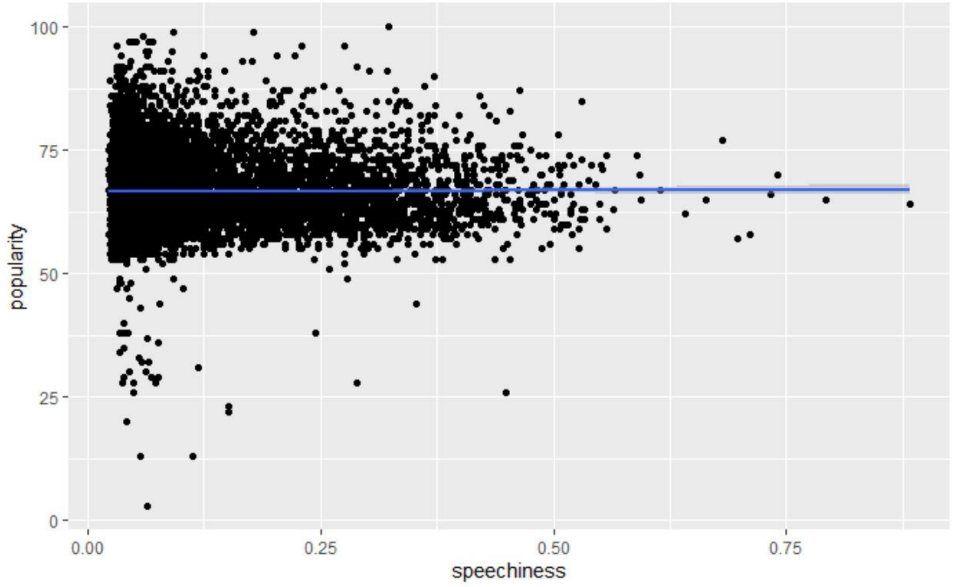
Distribution of songs based on Time Signature (Bar Plot): 4/4-time signature seems to be either the most famous ones or the most used ones



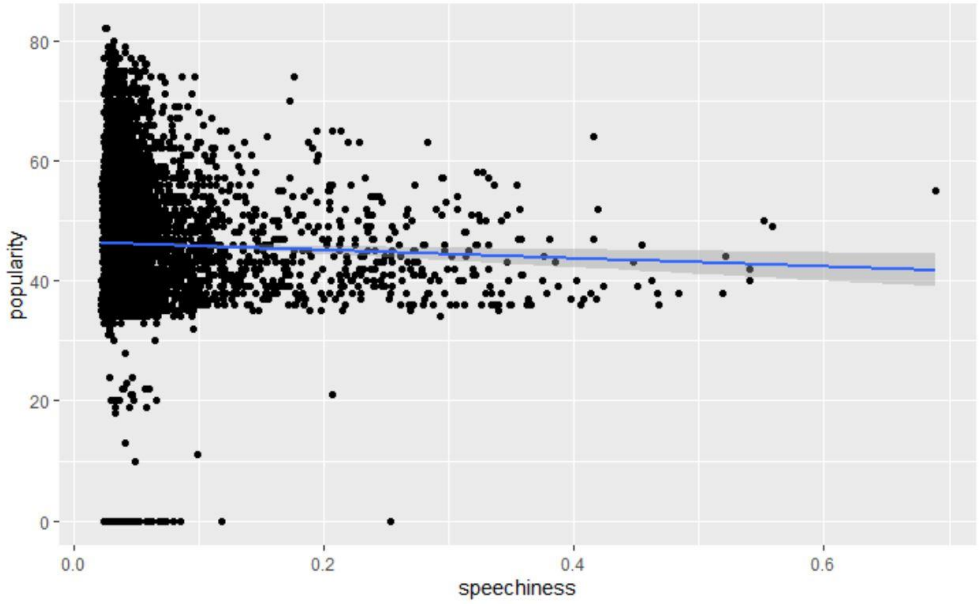
Danceability Pop (Scatter Plot): Suggests popular Pop songs tend to have greater danceability score



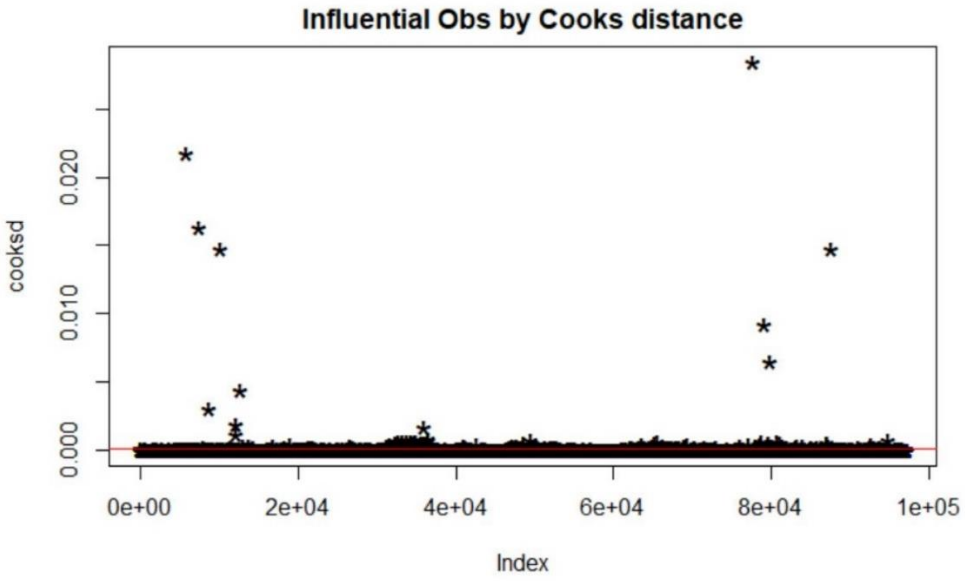
Danceability Country (Scatter Plot): Suggests most songs in Country genre have a moderate danceability score



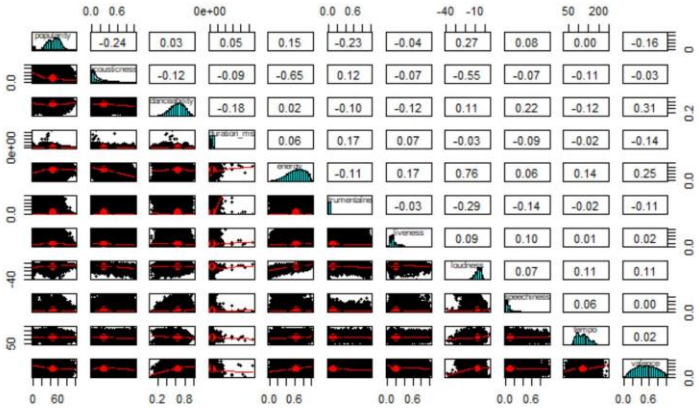
Speechiness Pop (Scatter Plot): Greater presence of spoken words in Pop songs



Speechiness Country (Scatter Plot): Comparatively lesser spoken words in popular Country songs



Outlier Detection – Cooks Distance



Correlation Analysis

a. Positive correlation between tempo, loudness, energy, danceability v/s valence – Joyful songs tend to score higher on these audio features

b. Negative correlation between popularity, energy v/s acousticness

c. Moderate negative correlation between popularity v/s valence – Sad songs might tend to be more popular than happy ones

MODEL 1: LOGISTIC REGRESSION CLASSIFIER

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.986e-01  1.543e-02  51.770  < 2e-16 ***
acousticness       -1.468e-01  6.991e-03 -20.998  < 2e-16 ***
danceability        7.089e-02  1.103e-02   6.427 1.31e-10 ***
duration_ms        -1.044e-07  1.699e-08  -6.142 8.16e-10 ***
energy             -2.642e-01  1.317e-02 -20.067  < 2e-16 ***
instrumentalness   -2.977e-01  6.455e-03 -46.127  < 2e-16 ***
key.A.              1.554e-02  5.865e-03   2.649 0.008074 **
key.B               1.697e-02  5.594e-03   3.033 0.002424 **
key.C.              3.581e-02  4.799e-03   7.462 8.62e-14 ***
key.D               1.138e-02  5.110e-03   2.228 0.025886 *
key.D.              2.457e-02  8.970e-03   2.740 0.006154 **
key.F.              3.585e-02  6.012e-03   5.964 2.47e-09 ***
key.G.              3.090e-02  5.822e-03   5.308 1.11e-07 ***
liveness           -1.253e-01  9.224e-03 -13.580  < 2e-16 ***
loudness            2.260e-02  6.589e-04  34.307  < 2e-16 ***
mode.Minor          1.132e-02  3.095e-03   3.658 0.000254 ***
speechiness         1.604e-01  1.350e-02  11.888  < 2e-16 ***
time_signature.1.4 -4.307e-02  2.074e-02  -2.077 0.037796 *
time_signature.3.4 -5.345e-02  6.435e-03  -8.305  < 2e-16 ***
valence            -1.771e-01  6.978e-03 -25.373  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Dimensionality Reduction – Logistic Regression (Backward Elimination)

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21057  9748
         1   710   822

               Accuracy : 0.6766
                 95% CI : (0.6715, 0.6817)
    No Information Rate : 0.6731
    P-Value [Acc > NIR] : 0.09302

                  Kappa : 0.0579

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

            Sensitivity : 0.96738
            Specificity : 0.07777
         Pos Pred Value : 0.68356
         Neg Pred Value : 0.53655
             Prevalence : 0.67313
         Detection Rate : 0.65117
   Detection Prevalence : 0.95262
      Balanced Accuracy : 0.52257

       'Positive' Class : 0
```

Confusion Matrix – Logistic Regression

MODEL 2: NAÏVE BAYES CLASSIFIER

```
Confusion Matrix and Statistics


predict_nb2     0     1
          0  9409  1945
          1 12358  8625

               Accuracy : 0.5577
                 95% CI : (0.5523, 0.5631)
    No Information Rate : 0.6731
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1981

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

            Sensitivity : 0.8160
            Specificity : 0.4323
         Pos Pred Value : 0.4110
         Neg Pred Value : 0.8287
             Prevalence : 0.3269
         Detection Rate : 0.2667
   Detection Prevalence : 0.6489
      Balanced Accuracy : 0.6241

       'Positive' Class : 1
```

Confusion Matrix – Naïve Bayes

```
Cross-Validated (5 fold, repeated 1 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 29.3  6.0
         1 38.1 26.6

 Accuracy (average) : 0.5588
```

Confusion Matrix – Naïve Bayes (5-Fold Repeated Cross Validation)

MODEL 3: DECISION TREE CLASSIFIER

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 18540  6781
         1  3227  3789

               Accuracy : 0.6905
                 95% CI : (0.6854, 0.6955)
    No Information Rate : 0.6731
    P-Value [Acc > NIR] : 1.135e-11

                  Kappa : 0.2301

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8517
            Specificity : 0.3585
         Pos Pred Value : 0.7322
         Neg Pred Value : 0.5401
             Prevalence : 0.6731
         Detection Rate : 0.5733
   Detection Prevalence : 0.7830
      Balanced Accuracy : 0.6051

       'Positive' Class : 0
```

Confusion Matrix – Decision Tree

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19480  7142
         1  2287  3428

               Accuracy : 0.7084
                 95% CI : (0.7034, 0.7134)
    No Information Rate : 0.6731
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2486

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8949
            Specificity : 0.3243
         Pos Pred Value : 0.7317
         Neg Pred Value : 0.5998
             Prevalence : 0.6731
         Detection Rate : 0.6024
   Detection Prevalence : 0.8233
      Balanced Accuracy : 0.6096

       'Positive' Class : 0
```

Confusion Matrix – Decision Tree (Adaptive Boosting)

MODEL 4: RANDOM FOREST CLASSIFIER

```
Cross-Validated (5 fold, repeated 1 times) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 64.1 15.3
         1  3.3 17.3

 Accuracy (average) : 0.8135
```

Confusion Matrix – Random Forest (5-Fold Cross Validation)

```
Confusion Matrix and Statistics


rf_pred     0     1
      0 20902  4325
      1   865  6245

               Accuracy : 0.8395
                 95% CI : (0.8355, 0.8435)
    No Information Rate : 0.6731
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6017

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9603
            Specificity : 0.5908
         Pos Pred Value : 0.8286
         Neg Pred Value : 0.8783
             Prevalence : 0.6731
         Detection Rate : 0.6464
   Detection Prevalence : 0.7801
      Balanced Accuracy : 0.7755

       'Positive' Class : 0
```

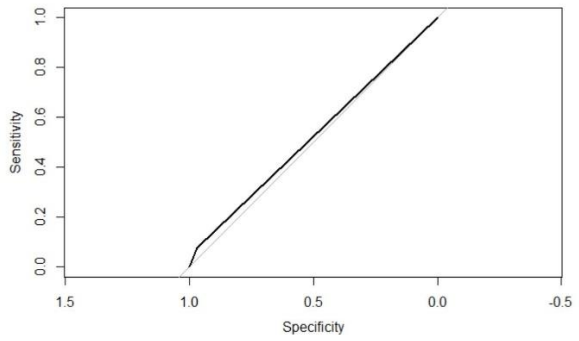
Confusion Matrix – Random Forest

MODEL 5: ENSEMBLE MODEL

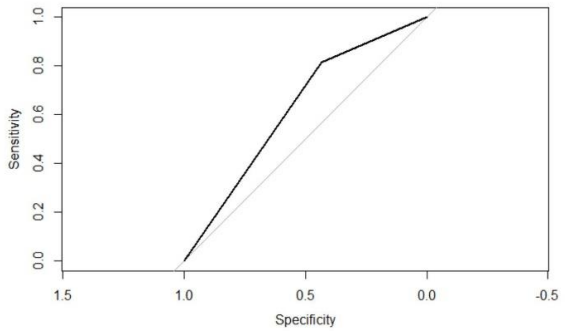
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 21057  9748
         1   710   822

               Accuracy : 0.6766
                 95% CI : (0.6715, 0.6817)
    No Information Rate : 0.6731
    P-Value [Acc > NIR] : 0.09302

                  Kappa : 0.0579

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

            Sensitivity : 0.96738
            Specificity : 0.07777
         Pos Pred Value : 0.68356
         Neg Pred Value : 0.53655
             Prevalence : 0.67313
         Detection Rate : 0.65117
   Detection Prevalence : 0.95262
      Balanced Accuracy : 0.52257

       'Positive' Class : 0
```

Confusion Matrix – Ensemble Model

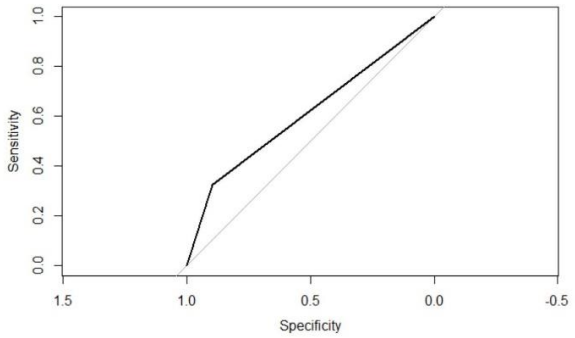
MODEL EVALUATION: AREA UNDER CURVE



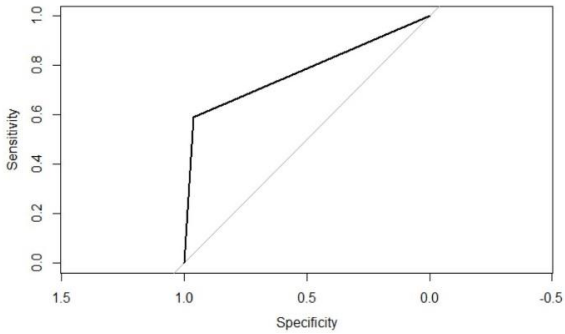
AUC – LogReg



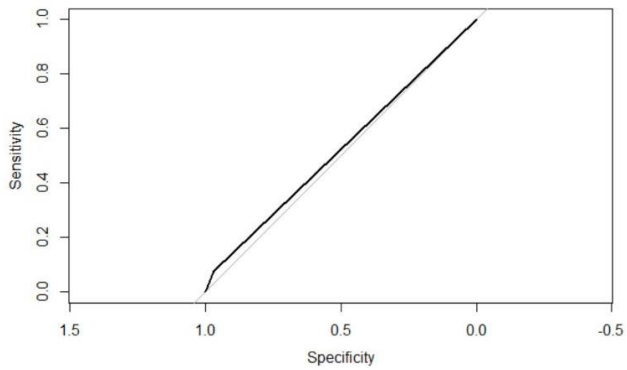
AUC – Naïve Bayes



AUC – Decision Tree



AUC – Random Forest



AUC – Ensemble Model

MODEL EVALUATION: SUMMARY

Model <fctr>	Accuracy <dbl>	AUC <dbl>
Logistic Regression	0.6765934	0.5225745
Naive Bayes	0.5576893	0.6241242
Decision Tree	0.7084145	0.6096234
Random Forest	0.8395027	0.7755420
Ensemble Model	0.6765934	0.5225745

Accuracy & AUC Model Comparison

Features <fctr>	Popular_Mean <dbl>	Drake_Mean <dbl>
Danceability	0.64736715	0.662675052
Energy	0.64696776	0.559371069
Instrumentalness	0.02337607	0.006497928
Loudness	-6.78369380	-8.084322851
Tempo	121.15440178	120.606354298

Favorite Artist (Drake) – Statistical Comparison