Group 4

SONG POPULARITY PREDICTION

Asha Gutlapalli, Diwash Bajracharya,
Shruthi Laya Hariharan

MOTIVATION

- Music is the most effective form of art that has the capacity to make humans feel all types of emotions like love, anger, fear, and sorrow

- Research shows that music can heal people of various serious physical and mental illnesses.

- Music online platforms like Spotify pop up train their algorithms to recommend music based on the user's preferences.

PROBLEM STATEMENT

- In this project, we attempt to answer the following questions:

  - Goal: Will the song become popular?
  - Explainability: Which features have the most influence on the popularity of a song?
  - Predictability: Which model achieves the most accuracy compared to alternative methods?

DATASET

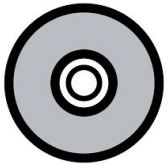
- The dataset contains the top 2000 songs collected through 1956 to 2019.

- It was sourced from Kaggle, an online community of data scientists and machine learning practitioners where users publish datasets, explore and build models.

- It consists of 13 features namely Title, Artist, Top Genre, Year, Beats Per Minute, Energy, Danceability, Loudness, Liveness, Valence, Length, Acousticness, Speechiness, and the response variable popularity.

DATASET

Title



Artist

Top Genre

Year

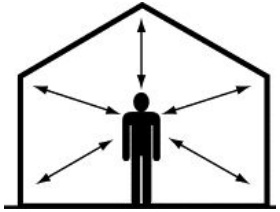
Beats Per Minute

Energy

Danceability

Loudness

Liveness



Valence

Length

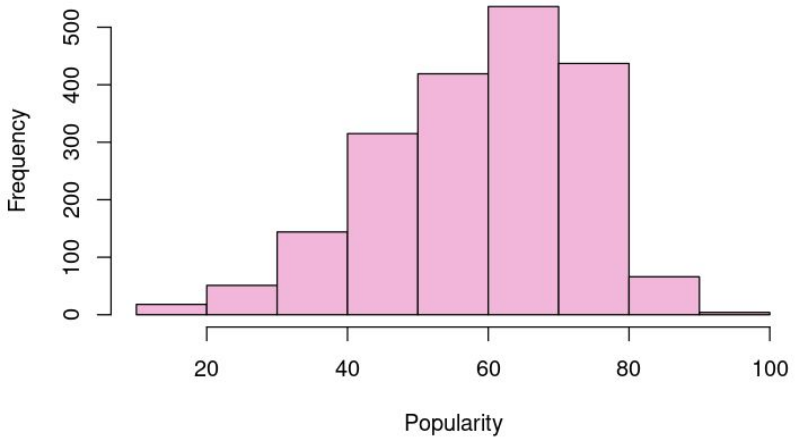
Acouticness

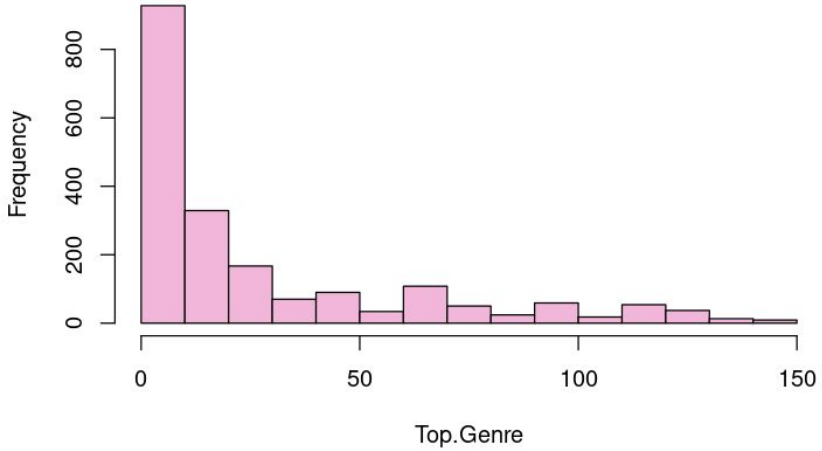
Speechiness

Popularity

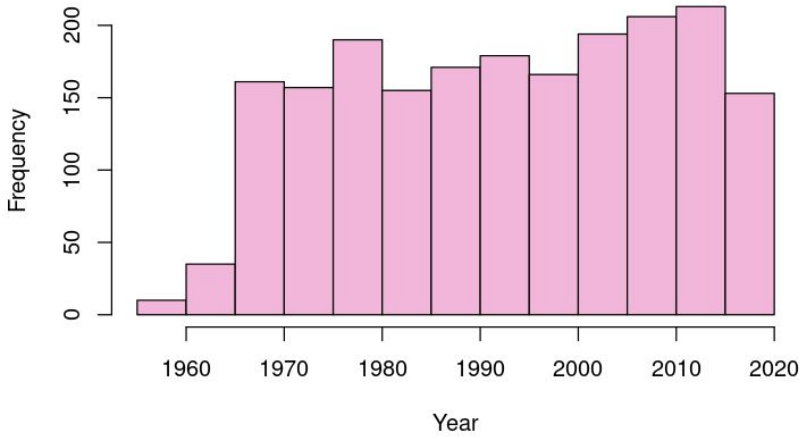
EXPLORATORY DATA ANALYSIS

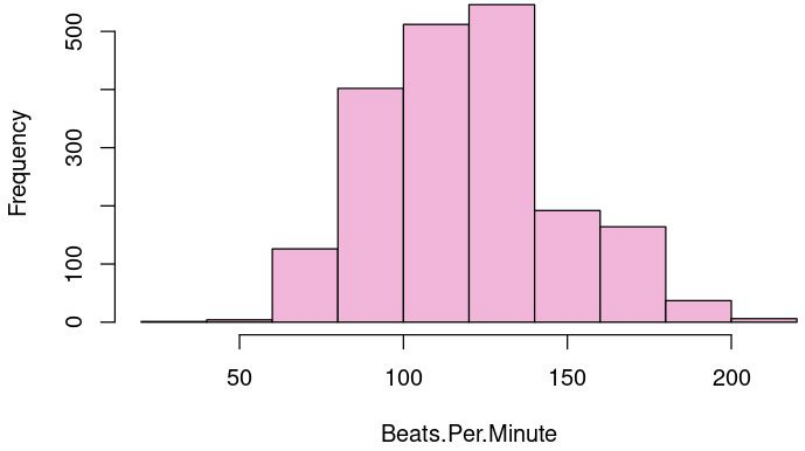
The distributions of all the attributes and the response variable are visualized below:



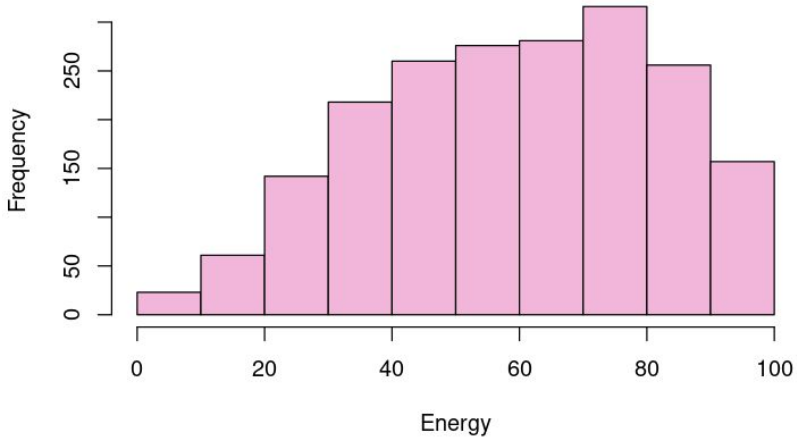


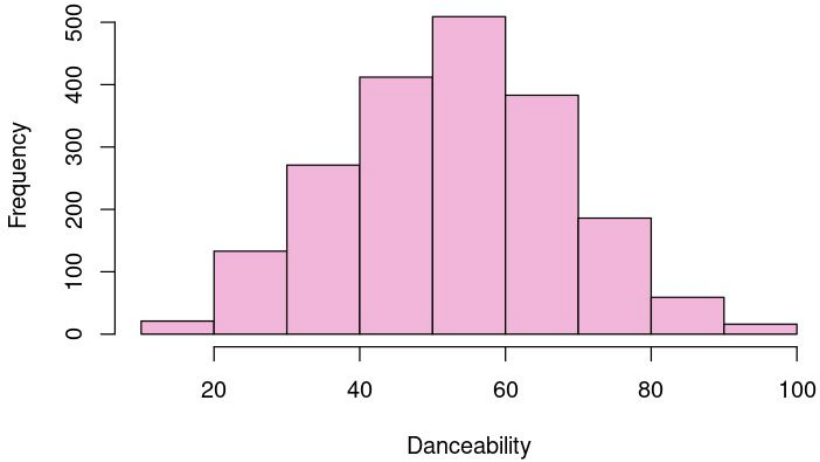
EXPLORATORY DATA ANALYSIS



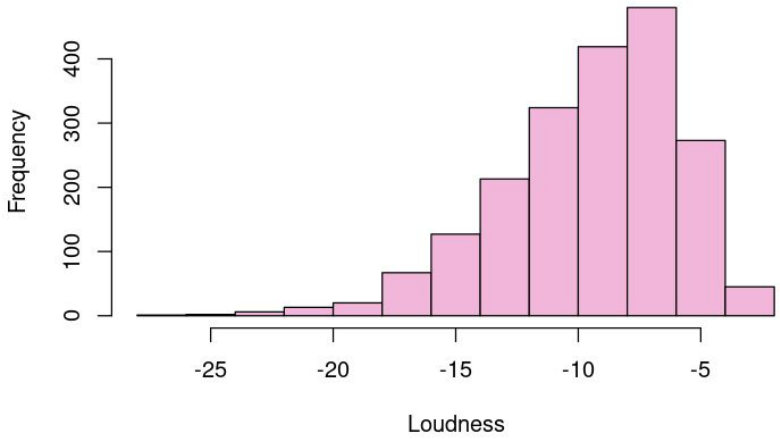


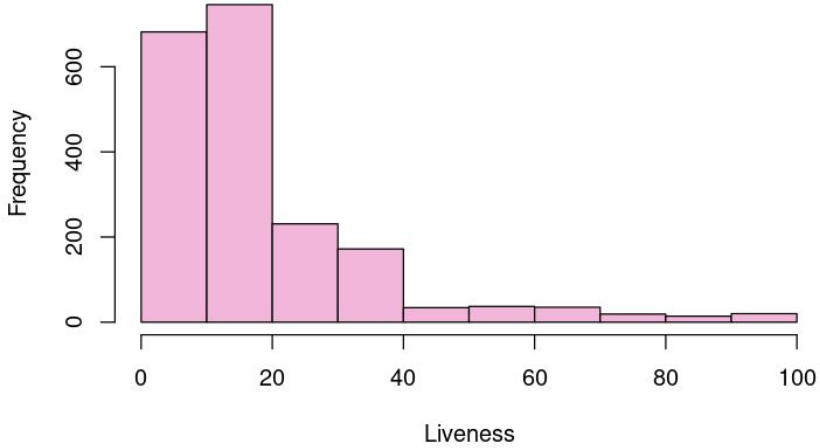
EXPLORATORY DATA ANALYSIS



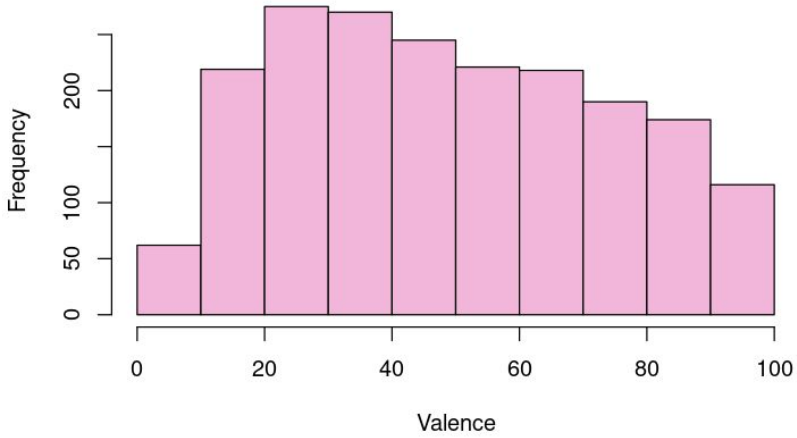


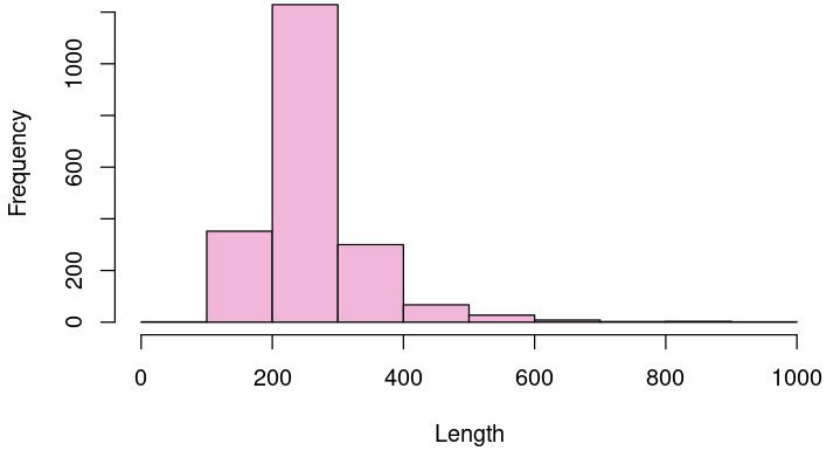
EXPLORATORY DATA ANALYSIS



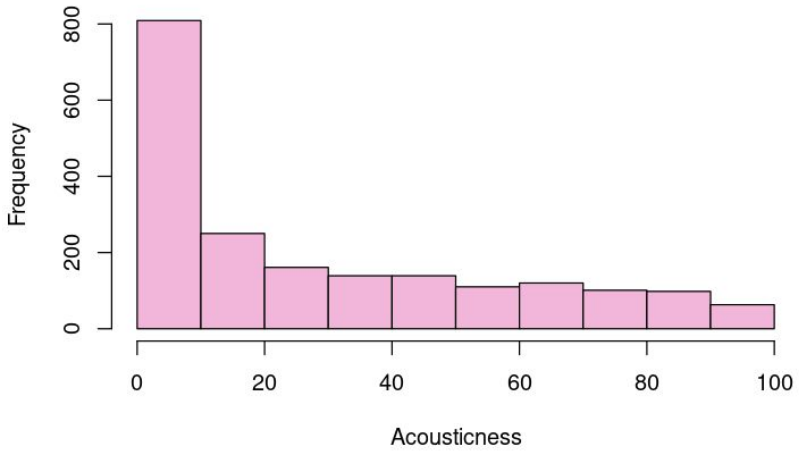


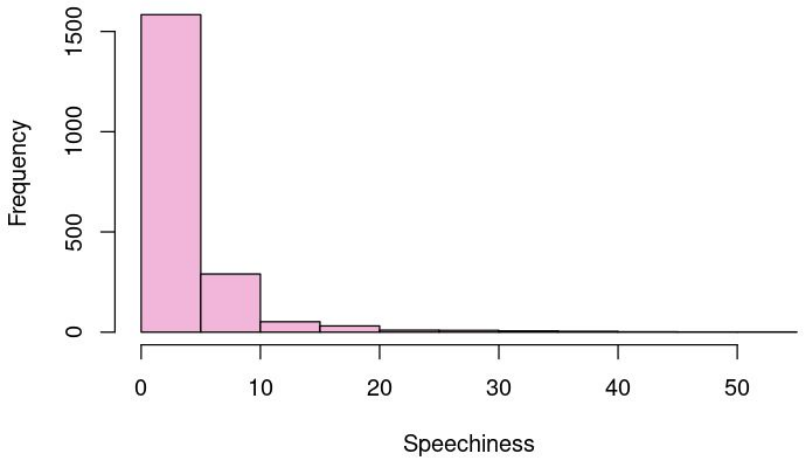
EXPLORATORY DATA ANALYSIS





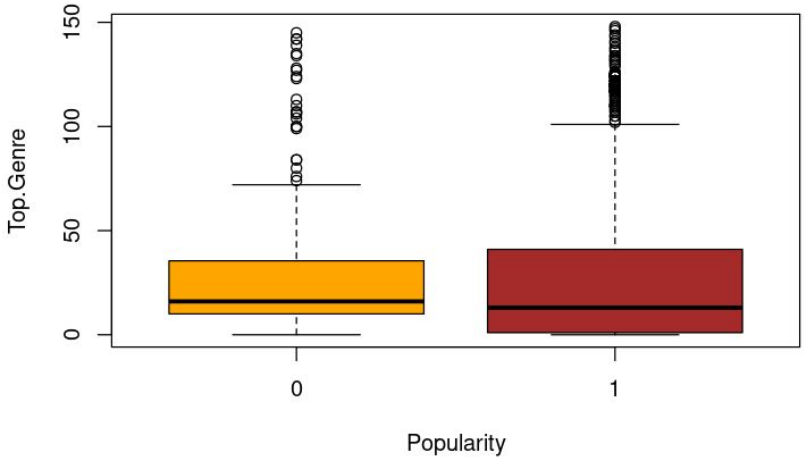
EXPLORATORY DATA ANALYSIS

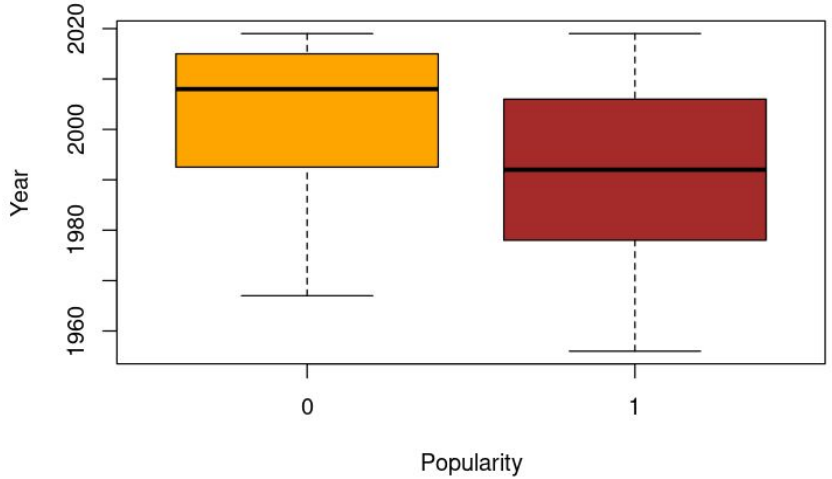




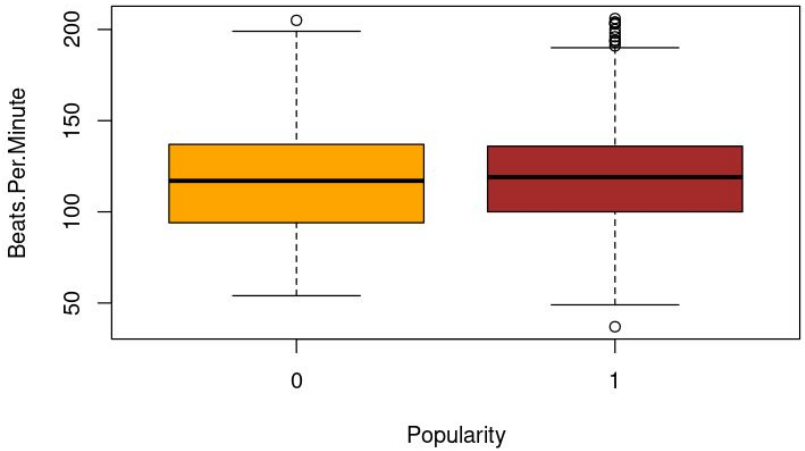
EXPLORATORY DATA ANALYSIS

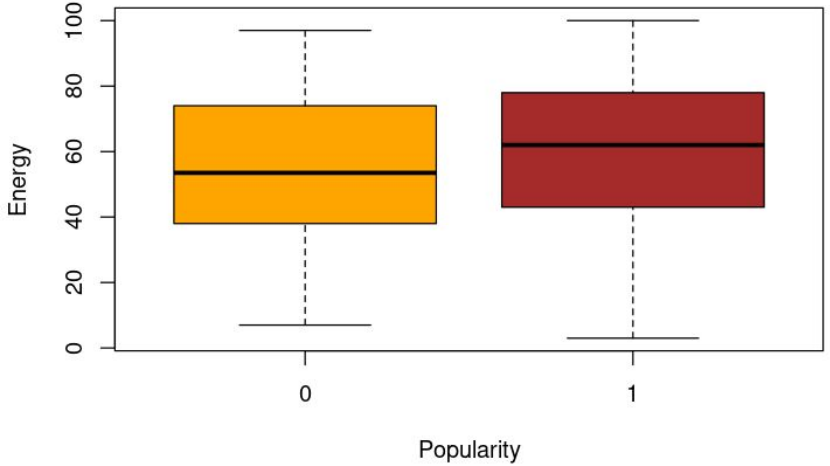
The relationships between all the attributes and the response variable are visualized below:



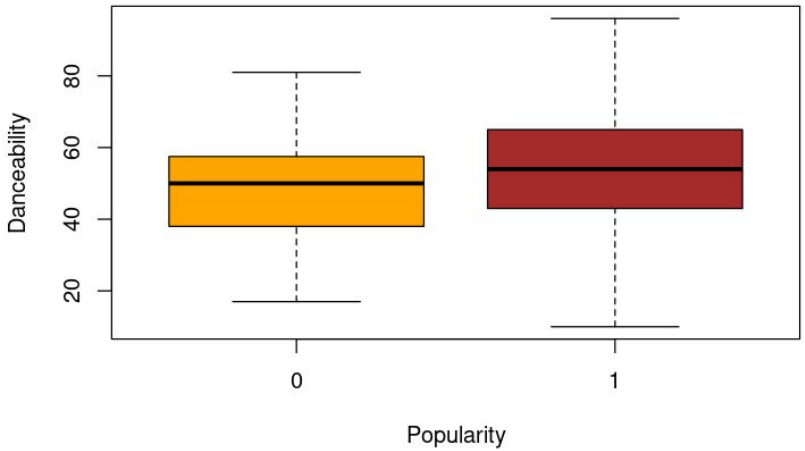


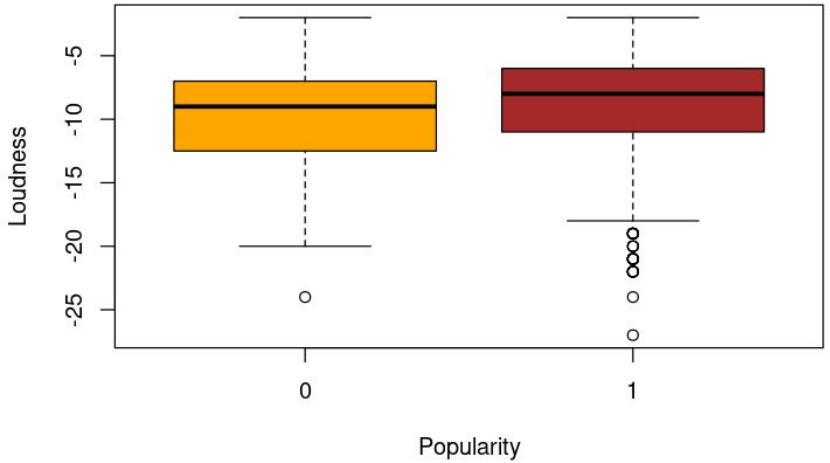
EXPLORATORY DATA ANALYSIS



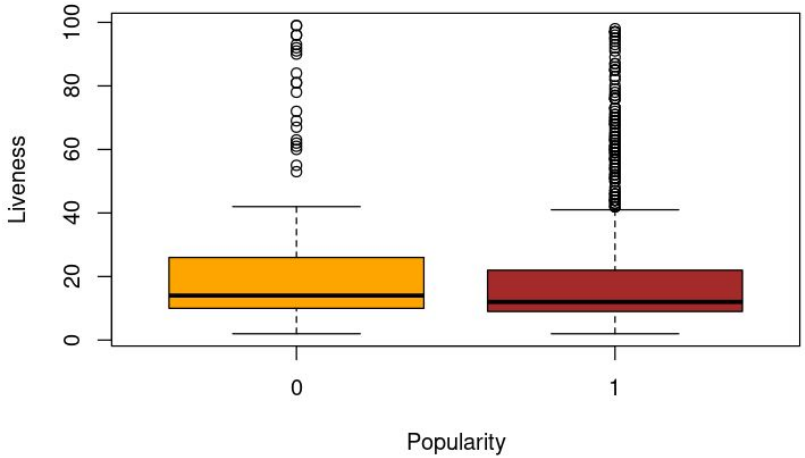


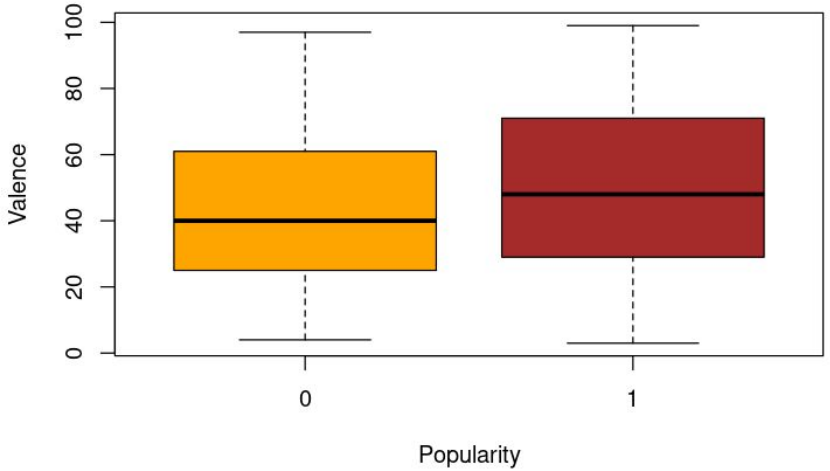
EXPLORATORY DATA ANALYSIS



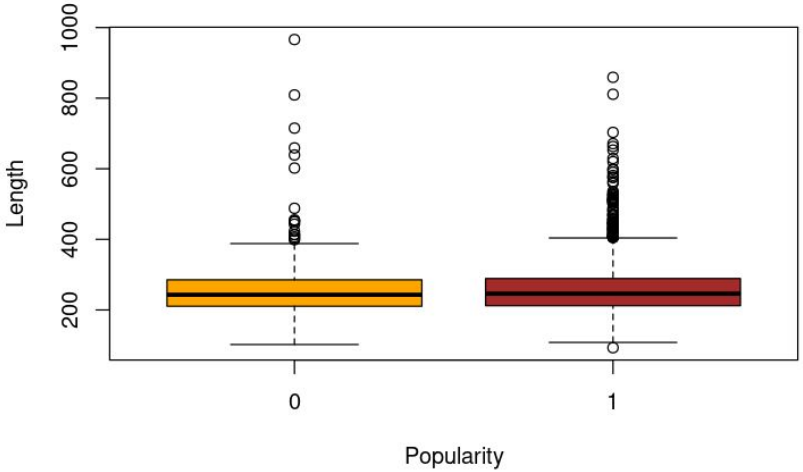


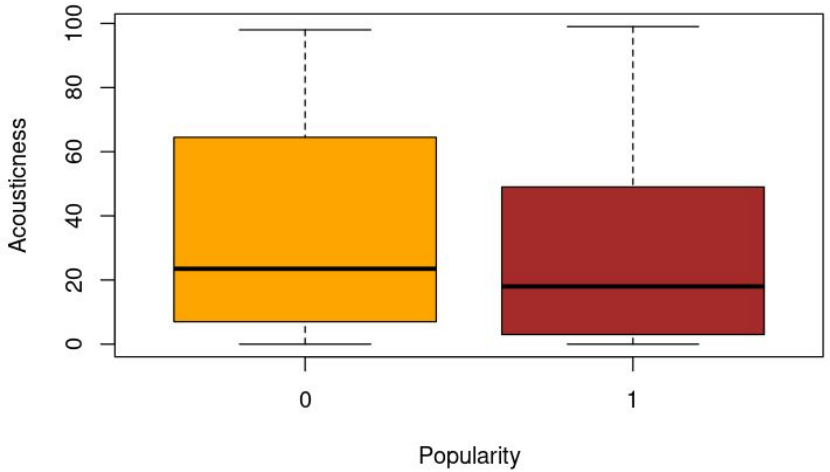
EXPLORATORY DATA ANALYSIS



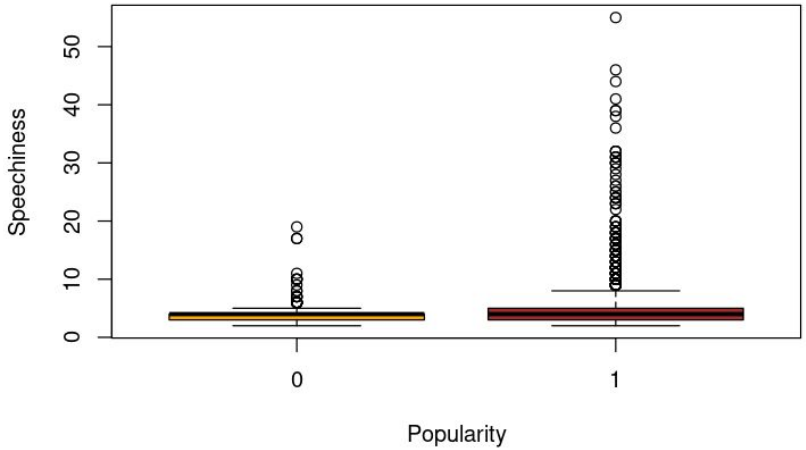


EXPLORATORY DATA ANALYSIS



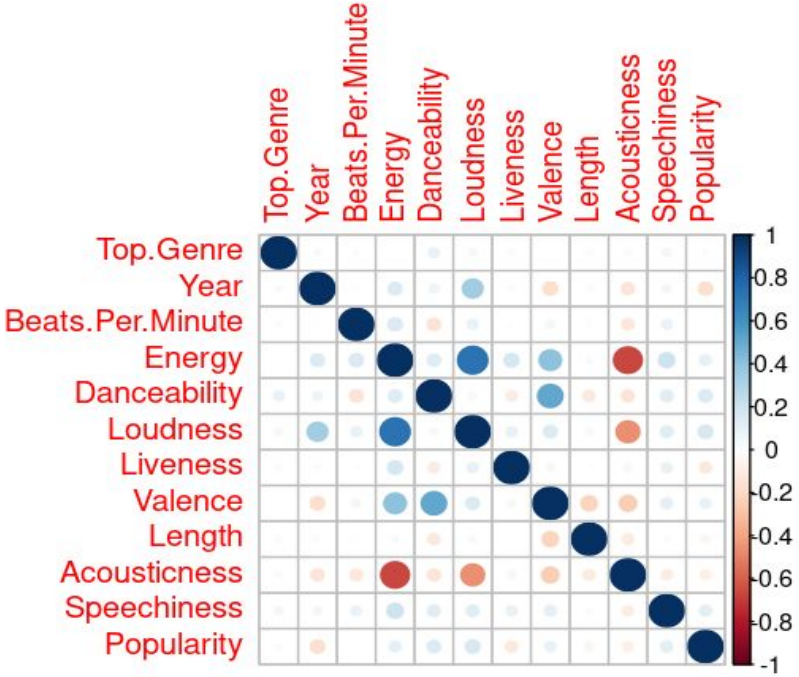


EXPLORATORY DATA ANALYSIS



EXPLORATORY DATA ANALYSIS

The correlation between the attributes and the response variable is visualized below:

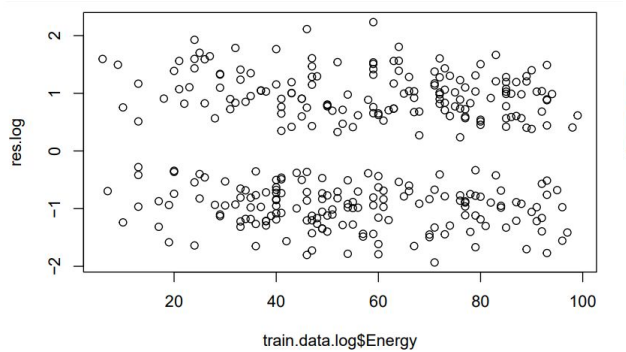


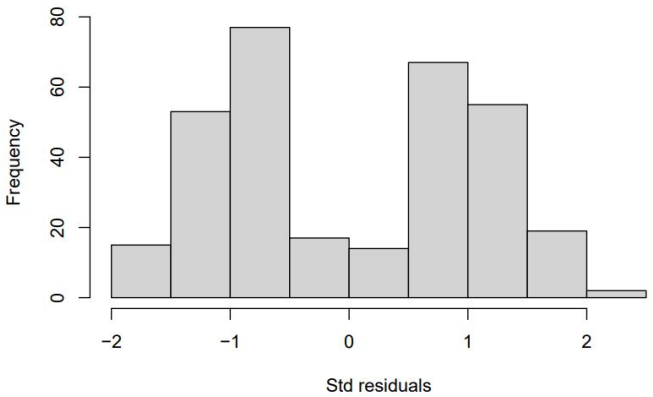
APPROACH

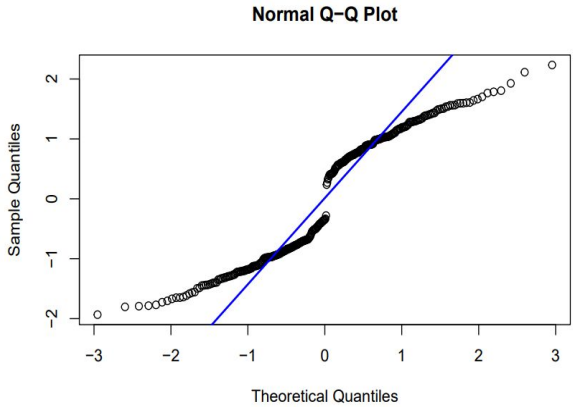
- The Song Popularity was considered as the response variable for the model.
- We considered shifting the Song Popularity Index from a numerical to a binary variable, where the threshold was set as k = 0.4
- The data set was split as 85% training data and 15% testing data
- After this transformation, we modelled the data using the following models
    - Logistic Regression
    - LASSO Regression
    - Elastic Net Regression
    - Decision Tree
    - Stepwise Regression
    - Ridge Regression
    - K-Nearest Neighbour
    - RandomForest

LOGISTIC REGRESSION

- Logistic regression is used to predict a binary outcome and is useful to classify response variable into a category








```
## Logistic Model

##    Accuracy Sensitivity Specificity
##   0.8070175   0.8000000   0.8125000
```

```
round(c(pearson.log, 1-pchisq(pearson.log, 307)), 2)

## [1] 308.65   0.46
```

STEPWISE REGRESSION

Stepwise regression iteratively constructs a regression model step-by-step by adding or removing one independent variable at a time to the final model. We perform both forward and backward stepwise regression..

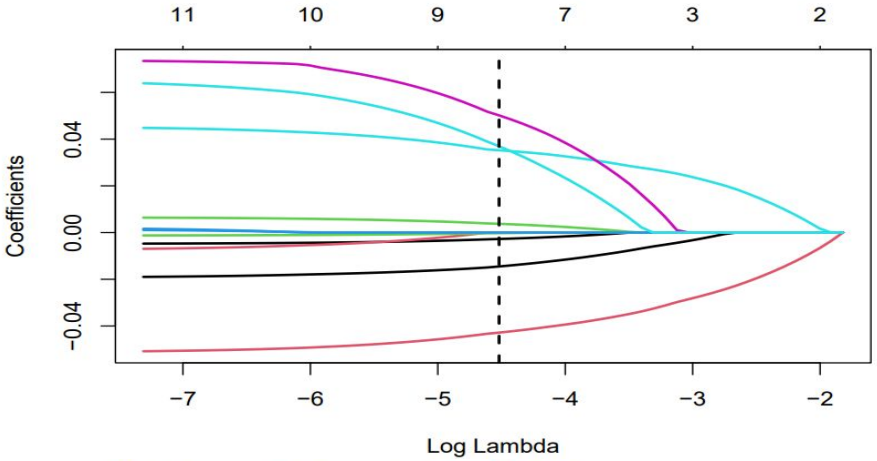
```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9913  -0.9561  -0.3836   0.9708   2.0938
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  95.275223  17.329204   5.498 3.84e-08 ***
## Year         -0.048301   0.008633  -5.595 2.21e-08 ***
## Danceability  0.035956   0.009066   3.966 7.31e-05 ***
## Liveness     -0.020412   0.008438  -2.419   0.0156 *
## Loudness      0.070569   0.035346   1.997   0.0459 *
## Speechiness   0.071820   0.049310   1.456   0.1453
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 442.15  on 318  degrees of freedom
## Residual deviance: 369.72  on 313  degrees of freedom
## AIC: 381.72
```

```
## Stepwise Regression

##    Accuracy Sensitivity Specificity
##   0.8245614   0.7307692   0.9032258
```

LASSO REGRESSION

Lasso regression, or the Least Absolute Shrinkage and Selection Operator, is a modification of linear regression. In lasso, the loss function is modified to minimize the complexity of the model by limiting the sum of the absolute values of the model coefficients (also called the L1-regularization).

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)       83.9190880464
## Top.Genre         -0.0027501889
## Year              -0.0428256336
## Beats.Per.Minute   0.0037381507
## Energy             .
## Danceability       0.0352363167
## Loudness           0.0501544287
## Liveness          -0.0145192863
## Valence            .
## Length            -0.0002564901
## Acousticness       .
## Speechiness        0.0369238388
```

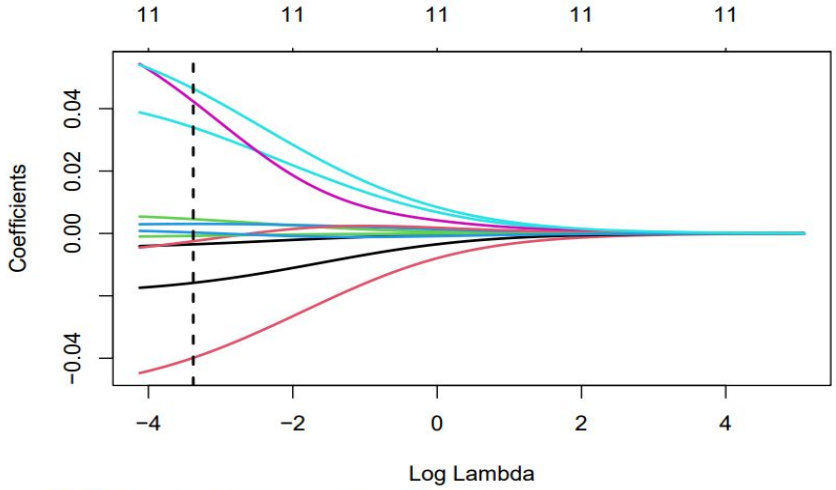



```
## Lasso Regression Model

##     Accuracy Sensitivity Specificity
##    0.7719298   0.6969697   0.8750000
```

RIDGE REGRESSION

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization.

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)       77.9782306359
## Top.Genre         -0.0034914179
## Year              -0.0398689898
## Beats.Per.Minute   0.0045884310
## Energy             0.0030483618
## Danceability       0.0339685072
## Loudness           0.0423200080
## Liveness          -0.0158440391
## Valence           -0.0024905395
## Length            -0.0007740176
## Acousticness       0.0003019409
## Speechiness        0.0464087120
```

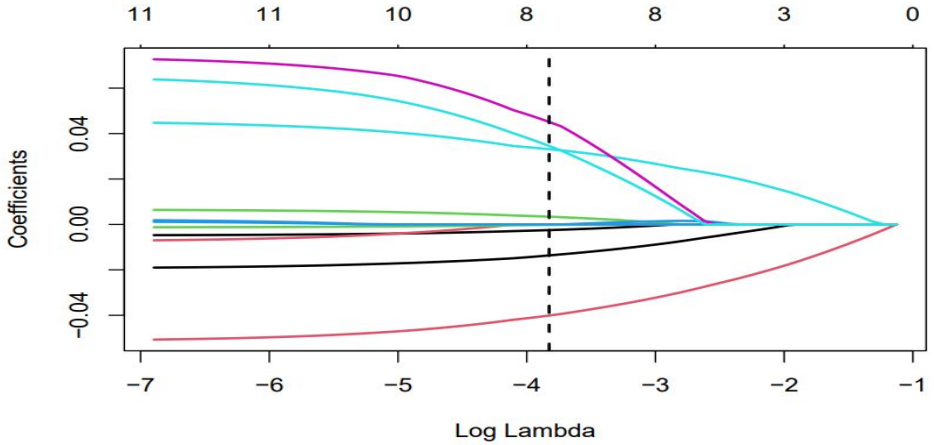



```
## Ridge Regression Model

##     Accuracy Sensitivity Specificity
##    0.7894737   0.7500000   0.8275862
```

ELASTIC NET REGRESSION

Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training. L1 regularization penalizes the sum of absolute values of the weights, whereas L2 regularization penalizes the sum of squares of the weights.

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)       78.5761625214
## Top.Genre         -0.0025019416
## Year              -0.0401089606
## Beats.Per.Minute   0.0034033634
## Energy             .
## Danceability       0.0331366649
## Loudness           0.0451387682
## Liveness          -0.0136482973
## Valence            .
## Length            -0.0002265023
## Acousticness       .
## Speechiness        0.0345353198
```



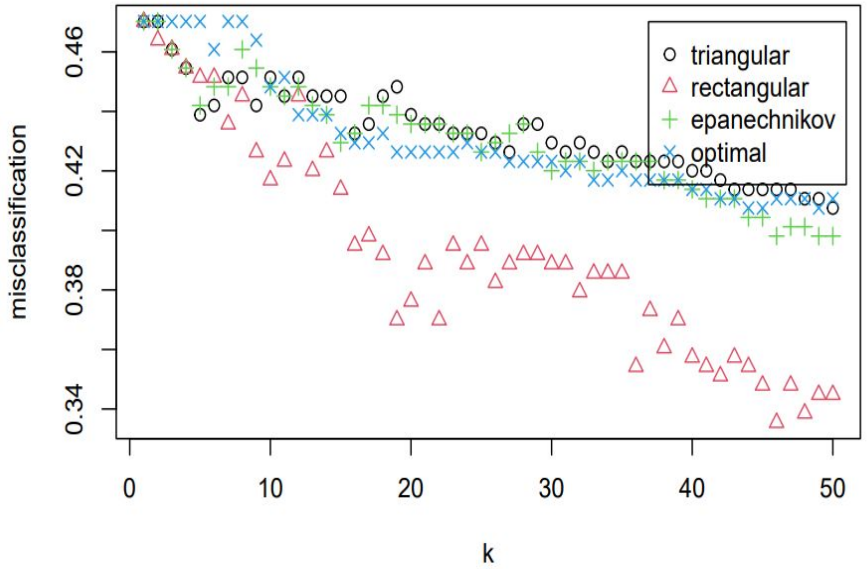

```
## Elastic Net Regression Model

##    Accuracy Sensitivity Specificity
##   0.7719298   0.7096774   0.8461538
```

K-NEAREST NEIGHBOR

K Nearest Neighbor algorithm falls under the Supervised Learning category and is used for classification and regression. K Nearest Neighbor suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

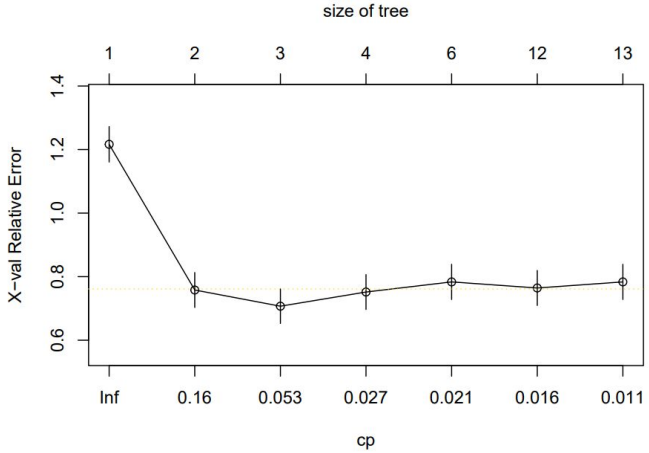
```
##
## Call:
## train.kknn(formula = Class ~ ., data = train.data.log, kmax = 50
##
## Type of response variable: nominal
## Minimal misclassification: 0.3354232
## Best kernel: rectangular
## Best k: 46
```

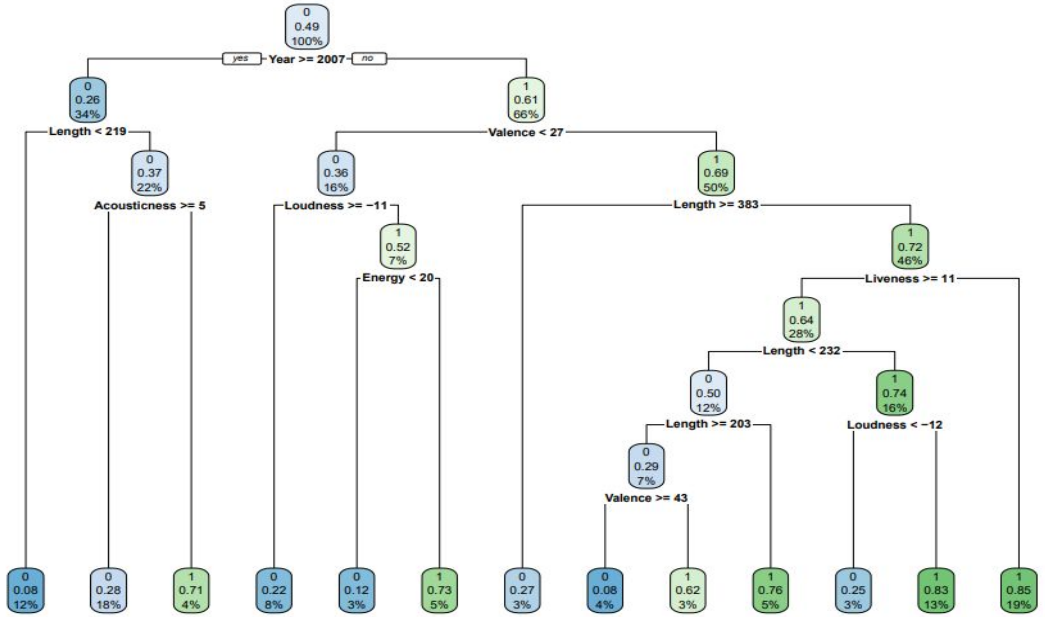
```
## KNN Model

##    Accuracy Sensitivity Specificity
##   0.7894737   0.7500000   0.8275862
```



DECISION TREE

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.






```
## Decision Tree

##    Accuracy Sensitivity Specificity
##   0.6315789   0.6923077   0.5806452
```

RANDOM FOREST

Random forests are an ensemble learning method for classification & regression that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees.

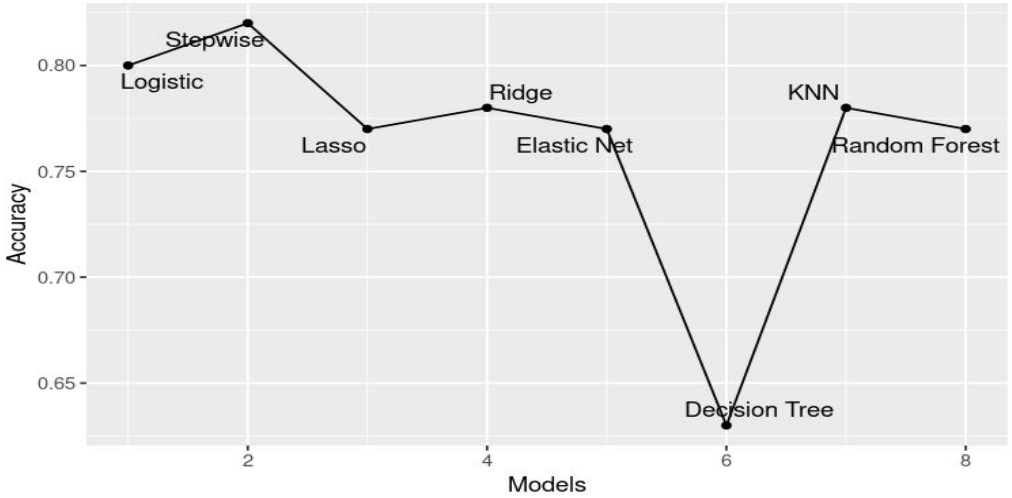
```
##                 Length Class  Mode
## call               3   -none- call
## type               1   -none- character
## predicted        319   factor numeric
## err.rate        1500   -none- numeric
## confusion          6   -none- numeric
## votes            638   matrix numeric
## oob.times        319   -none- numeric
## classes            2   -none- character
## importance        11   -none- numeric
## importanceSD       0   -none- NULL
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            14   -none- list
## y                319   factor numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

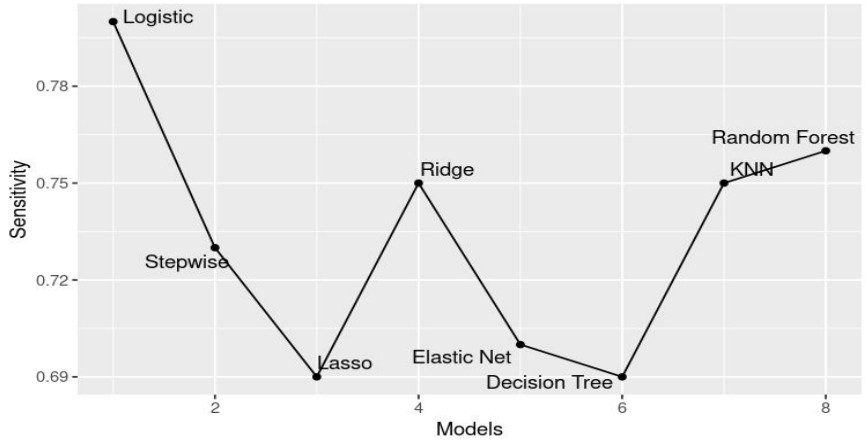
```
## Random Forest

##    Accuracy Sensitivity Specificity
##   0.7719298   0.7692308   0.7741935
```

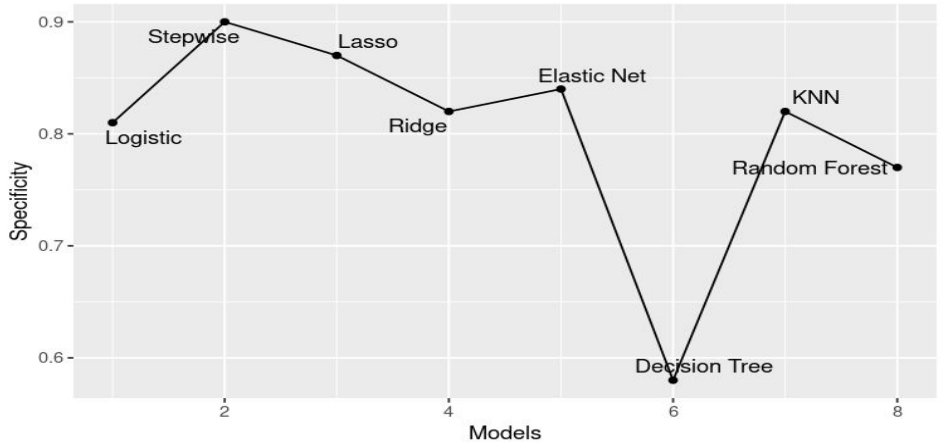
DISCUSSION AND RESULTS

Name of the Model	Accuracy	Sensitivity	Specificity
Logistic Regression	0.8070	0.8000	0.8125
Stepwise Regression	0.8245	0.7307	0.9032
LASSO Regression	0.7719	0.6969	0.8750
Ridge Regression	0.7894	0.7500	0.8275
Elastic Net Regression	0.7719	0.7096	0.8461
Decision Tree	0.6315	0.6923	0.5806
K-Nearest Neighbor	0.7894	0.7500	0.8275
Random Forest	0.7712	0.7692	0.7741

DISCUSSION AND RESULTS







DISCUSSION AND RESULTS

Model	Top Genre	Year	Beats Per Minute	Energy	Danceability	Loudness	Liveness	Valence	Length	Acousticness	Speechiness
LASSO	-0.0027501889	-0.04282563 36	0.0037381507		0.0352363167	0.0501544287	-0.0145192863		-0.0002564901		0.0369238388
Ridge	-0.0025019416	-0.0401089600	0.0034033634	0.0030483618	0.0331366649	0.0451387682	-0.0136482973	-0.0024905395	-0.0002265023	0.0003019409	0.0345353198
Elastic Net	-0.0025019416	-0.0401089606	0.0034033634		0.0331366649	0.0451387682	-0.0136482973		-0.0002265023		0.0345353198
Stepwise		-0.063075	0.006856		0.034709	0.126734			-0.003235		

CONCLUSION

After modelling the dataset with the discussed regression models, it is found that Stepwise Regression model yields the highest accuracy and it is concluded that the following variables contribute the most to a song popularity, according to Stepwise Regression -

- Year
- Danceability
- Liveness
- Loudness
- Speechiness

THANK YOU