*IE 7275: Data Mining in Engineering*

**Project Report**

Shushmath Koniki

Manikanta Borusu

koniki.s@husky.neu.edu

Borusu.s@husky.neu.edu

**Percentage of Effort Contributed by Student 1: 50%**

**Percentage of Effort Contributed by Student 2: 50%**

**Signature of Student 1:Shushmath Koniki Signature of Student 2: Manikanta Borusu**

**Submission Date: 21 April. 2019**

**Report**

**Prediction of Match Result using Half Time Statistics**

**Introduction**

The Premier League is the top level of the English football league system. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League. The Premier League is a corporation in which the member clubs act as shareholders.

A football match is played at home of one team and another team will be playing away, there are 11 players in each team and each team will be allowed to make 3 substitutions in the entire 90 minutes. Our goal here is to predict if the home team is going to win the match are not based on half time stats. There are several predictors that are considered like shots by home and away teams, number of corners taken by home and away teams, number of corners taken by the home and away teams, number of shots on target and number of goals scored by the end of half time.

**Problem statement**

Predicting if the home team won the match based on all the predictors would be key factor, this will help the team to change strategies on how to approach the second half and play better, it will also help the management with the marketing strategies. Premier league being one of top 5 leagues in Europe it is key to predict the score lines to play at the top level.

**Proposed solution**

Each team in the Premier league can use one of the following predicting models to predict if the home team is going to win the match or not: 1. Logistic Regression

2. Random Forests

3. Neural Networks

**Procedure**

The Premier league 2018’s season dataset was taken from bet365.com. Initial pre-processing involved dealing with some missing values and we took the average as there weren’t a lot of missing values.

After that as there were several more attributes like the name of the stadium, the referee etc., were dropped as they don’t have any effect on the prediction and the data related to full time score were also dropped.

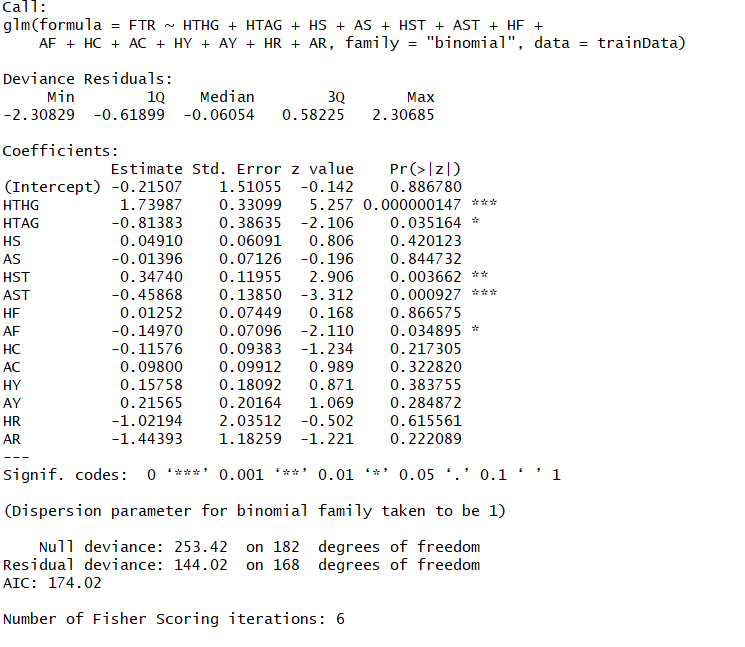
The response variable fulltime results were converted into a binary categorical variable to align with the project goals

Data cleaning involved checking in the data needed resampling, but both the classes had equal records, so we went on with it.

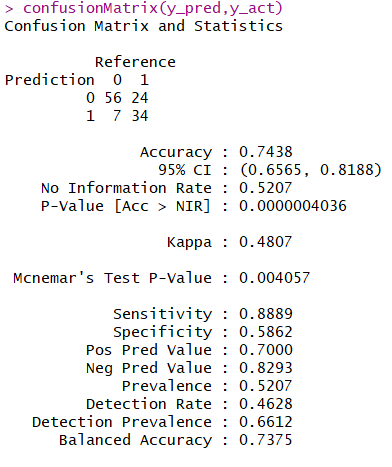
So, the final dataset had 304 observations and 15 variables.

The dataset was divided into 60% training set and 40% test set

The summary for the fitted model shows that the variables HTHG, HTAG, HST,AST,AF have a statistically significant influence on the dependent variable

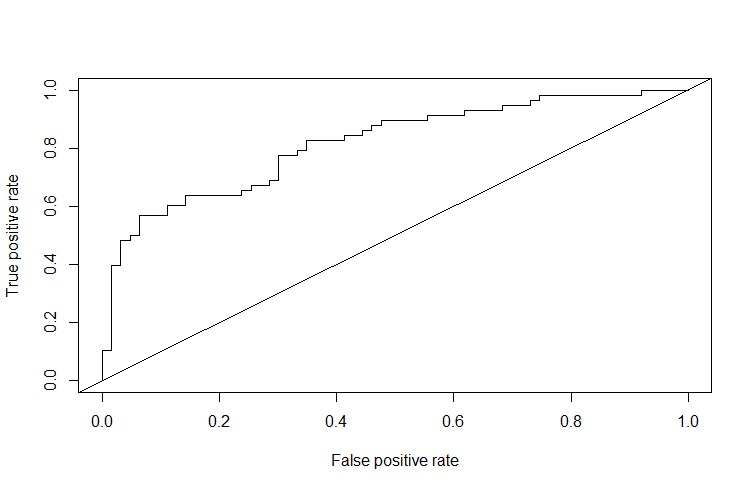


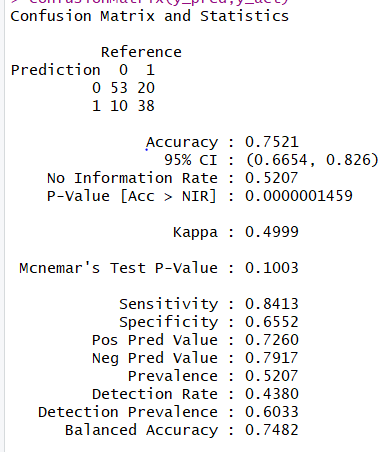
With 0.5 cut off, predicted the probability of the test dataset and created a confusion matrix



Achieved an accuracy of 74.3%

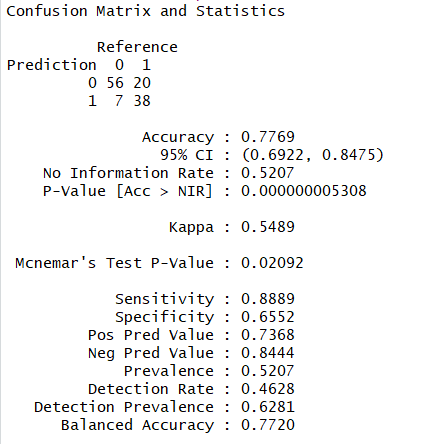
Constructed a ROC curve, which looks as follows:





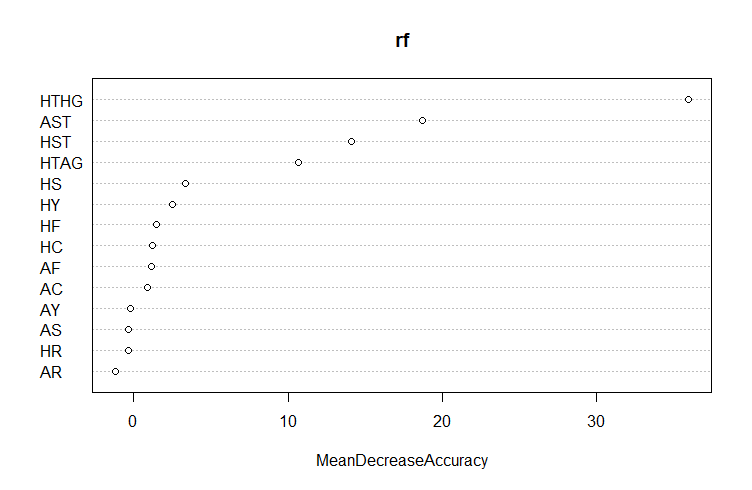
**Random Forests**

Again, we started with 60 and 40 split and implemented Random Forests model using RandomForests package and created a confusion matrix



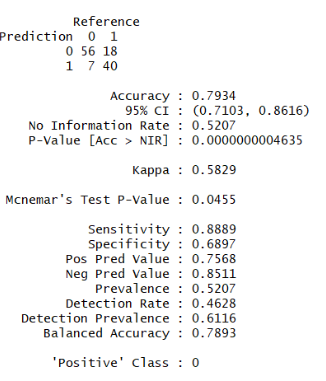
An accuracy of 77% was achieved with this model

Next, we plotted a variable importance chart and noticed that variables HTHG, AST, HST, HTAG have a significant influence on decision variable



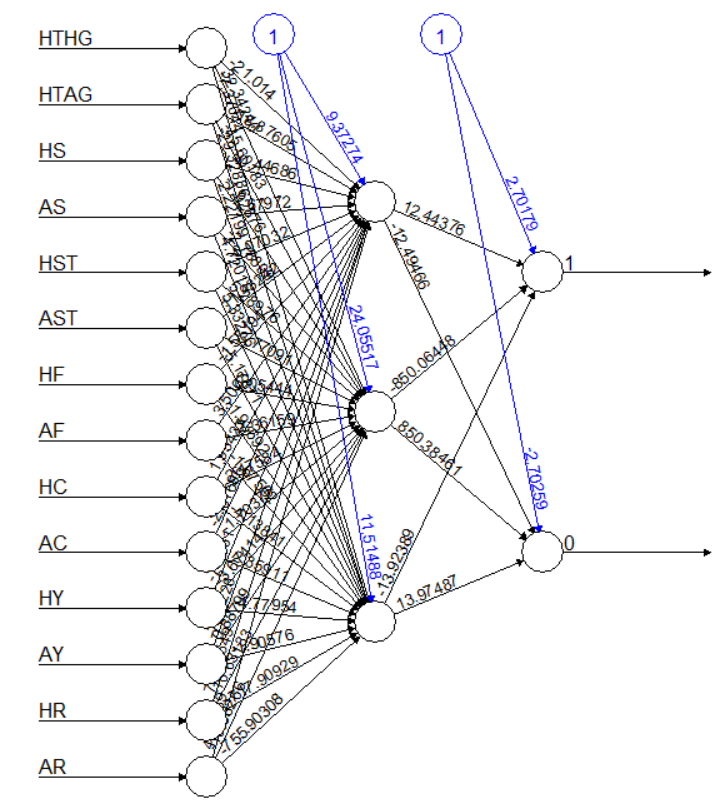
Later on, the default no of trees value of 500 was changed to 300 and implemented the model

The accuracy of the score was improved to 79.3 % with this step.

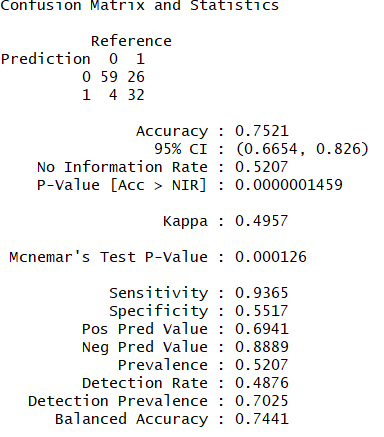


**Neural Networks**

For this model we followed the split as 60% training and 40% test as we did for the above two models and implemented Neural Network model with 3 neurons and a single layer

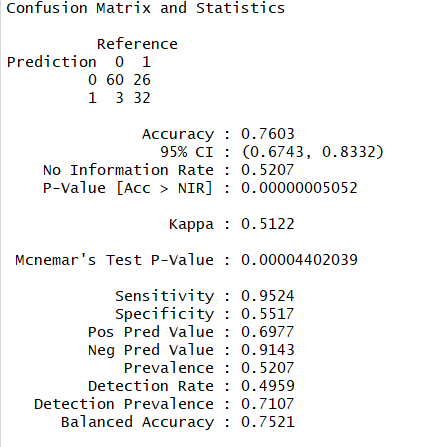


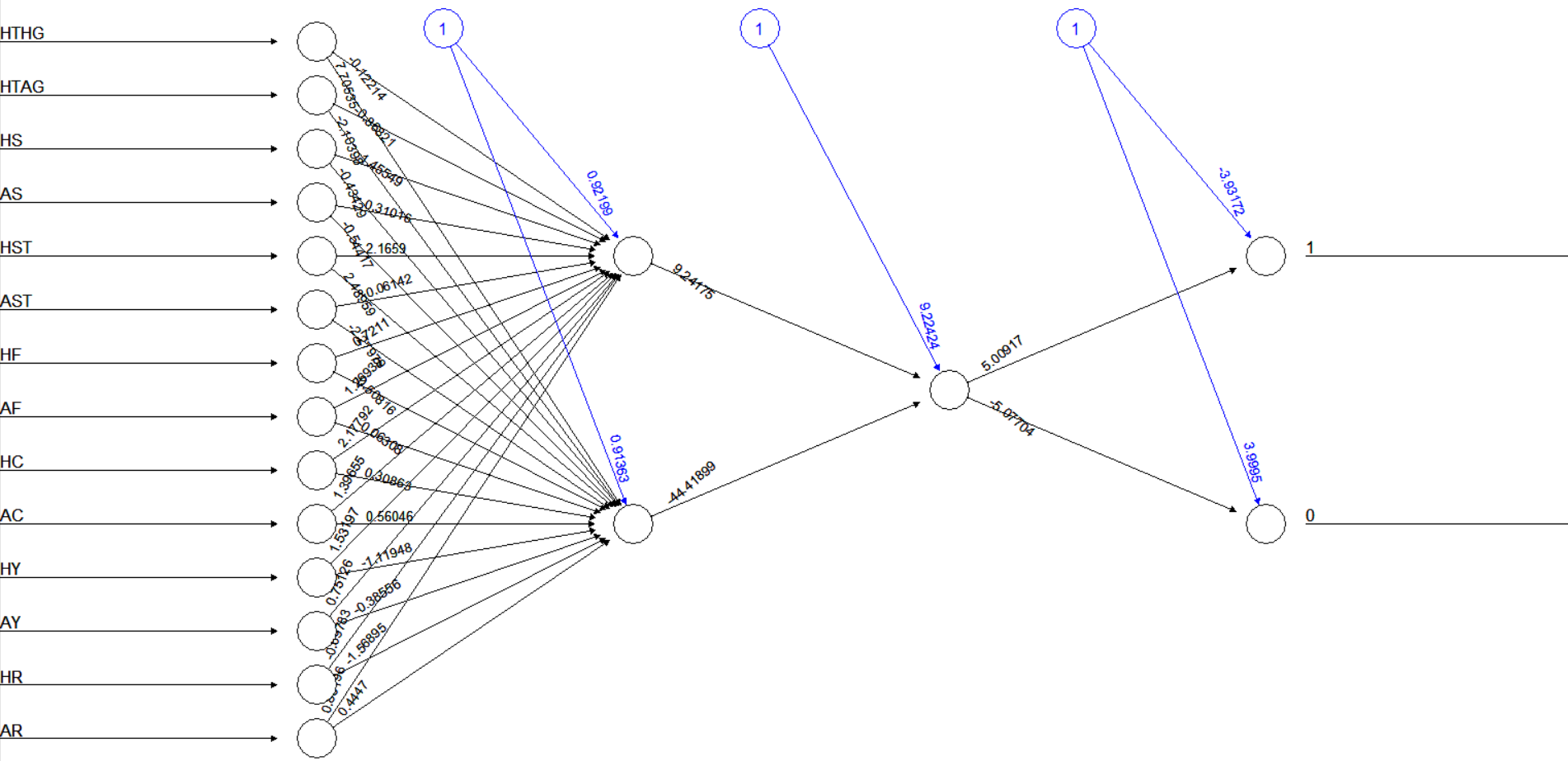
Using a cut off value of 0.5 a confusion matrix was created and obtained an accuracy of 75%



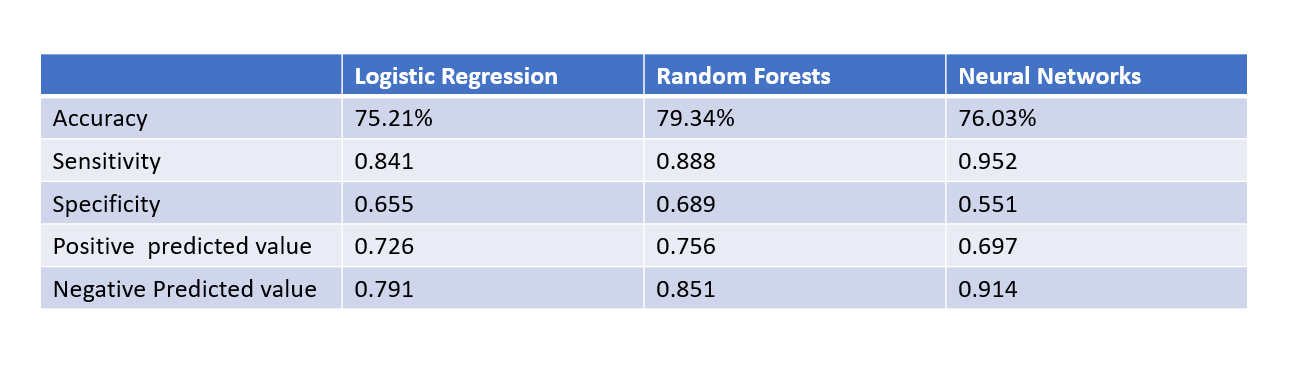
Implemented a neural network model with 3 neurons arranged in 2 layers (2 and 1 ) and computed a confusion matrix using a cut off of 0.5

This helped to achieve an accuracy of 76%





**Comparing the three models**



We can see that out of all the models Random Forests performs better for the Dataset

**Conclusion**

The Model is tested on live data of the ongoing soccer matches and has given similar accuracy scores.

This project has helped us understand what are the major factor that influence the final result of a soccer match,

This model can be used to understand what is the general trend in which the game is progressing and re-evaluate the strategies based on it at half time.

This Model can also be used for placing safe bets on home team based on half time statistics.