**Data Mining, ISM 6136, Fall 2021**

**Group 3**

Santosh Nalgonda (U66688529)

Venkata Sai Pavan Guduri (U21254711)

Kaushik Vaka (U38321067)

Surendra Varma Sagi (U91972271)

Nithin Sai Sudulakunta (U56557516)

**Analyzing and Predicting outcomes of IPL Cricket matches**

**ABSTRACT**

Sports Analytics recently resulted in a lot of impeccable results where in the outcome of a match is predicted based on various factors such as the venue, toss, bowler & batsman etc. IPL is one such a game, where 8 teams compete for the title. We have undertaken the exercise, for predicting the outcome of an IPL match, to design and develop predictive model using various analytical processes and some of the ML algorithms such as Gradient boosting and Decision trees.

The IPL dataset has all the matches played in previous 12 seasons. By taking all the factors accounting in winning the match a model is developed where in the accuracy of winning a team is predicted. With the help of Exploratory Data Analysis (EDA) the data is cleaned and all the null values are filled by the median of that column for the categorical data. Feature engineering helped in selecting the parameters that played an important role in deciding the winner of the match. Machine Learning algorithms helped in predicting the accuracies of the model developed. By exploring this dataset, we propose a model to help teams and team mates to decide which team need to get be attended first, who has higher priority in winning the match.

We can demonstrate from the results has greater than 90% of overall accuracy in predicting the Winner of the match by considering the factors such as toss, venue, bowler and batsmen. We used Decision Tree modeling to predict the winner of a match in IPL. As a part of this study, we also identified the most alarming conditions and factors of this winner prediction. We finalized previous IPL matches datasets to train and evaluate our model. We used the confusion matrix to analyze the classifiers in addition to checking the accuracy of the predictions.

**INTRODUCTION**

Cricket is a bat-and-ball sport in which two teams of eleven players compete on a field with a 22-yard (20-meter) pitch in the middle and two wickets at each end, each with two bails balanced on three stumps. When a fielding team member, known as the bowler, "bowls" (propels) the ball from one end of the pitch to the wicket at the other end, the game begins. The batting team scores runs by striking the bowled ball with a bat and sprinting between the wickets, while the bowling team seeks to prevent this by keeping the ball in the field and getting it to either wicket, dismissing each batter (so they are "out"). Being bowled, which occurs when the ball hits the stumps and dislodges the bails, as well as the fielding side catching a hit ball before it reaches the ground or hitting a wicket with the ball before a batter can cross the crease line in front of the wicket to complete a run, are all methods of dismissal.

The innings finishes when 10 batters have been struck out, and the teams switch roles. In international matches, a third umpire and match referee assist two umpires in adjudicating the game. Cricket is played in a variety of formats, ranging from Twenty20, in which each team bats for a single innings of 20 overs and the game lasts about three hours, to Test matches, which last five days.

The Indian Premier League (IPL) is a professional Twenty20 cricket league in which ten teams compete from ten different cities across India. The BCCI (Board of Control for Cricket in India) established the league in 2007. It takes place every year between March and May and has a special slot in the ICC Future Tours Program.

The IPL is the most popular cricket league in the world, ranking sixth among all sports leagues in terms of average attendance in 2014. The IPL was the first sporting event in the world to be streamed live on YouTube in 2010. According to Duff & Phelps, the IPL's brand value in 2019 was Rs 47,500 crore (US$6.3 billion). The 2015 IPL season provided Rs. 1,150 crore (US$150 million) to India's GDP, according to the BCCI. With 31.57 million average impressions and a 23 percent increase in overall consumption over the 2019 season, the 2020 IPL season broke a tremendous viewership record.

Since the dawn of the IPL in 2008, it has attracted viewers all around the globe. High level of uncertainty and last moment nail biters has urged fans to watch the matches. Within a short period, IPL has become the highest revenue generating league of cricket

**OBJECTIVE:**

Data Analytics has been a part of sports entertainment for a long time. In a cricket match, we might have seen the score-line showing the probability of the team winning based on the current match situation.

For the prediction of match winner

* We wanted to leverage the datasets of Kaggle of every match. This was a crucial information, with which multiple predictions and conclusions were possible.
* As we had data about the team matches played, it was possible to predict the winner based on past matches. Many features given in the data set like venue, bowler to batsman ball by ball performance, can be used to somewhat predict the winner of the match.
* We are also striving the best optimized model like gradient boost which will efficiently fit features of our dataset
* Append different datasets to deliver intelligent results for our prediction

**Importing deliveries Kaggle data set:**

A picture containing text

Description automatically generated

**Importing matches Kaggle data set:**

A screenshot of a computer

Description automatically generated with medium confidence

**Plot of team with most wins:**

Chart, bar chart

Description automatically generated

**Plot of Batsman with most runs:**

Chart

Description automatically generated

**Plot of bowler with most wickets:**

Chart

Description automatically generated

**Pie chart showing the teams winning percentage batting first:**

Chart, pie chart

Description automatically generated

**Pie chart showing the teams winning percentage bowling first:**

Chart, pie chart

Description automatically generated

**METHODS**

**A. DATASET**

We researched more than hundreds of cricket datasets available over the internet which has been published as public material and could be utilized for doing our study on predictive modeling. We finalized a dataset obtained from Kaggle for this exercise.

We used two datasets from Kaggle, one consists of information about every delivery since the beginning of IPL and other data set consists of information about every single match result. These datasets contain every trivial information like Toss winner, Venue, Non-striker batsman and even Umpires standing for the match. The data is clean, so it makes our work easier.

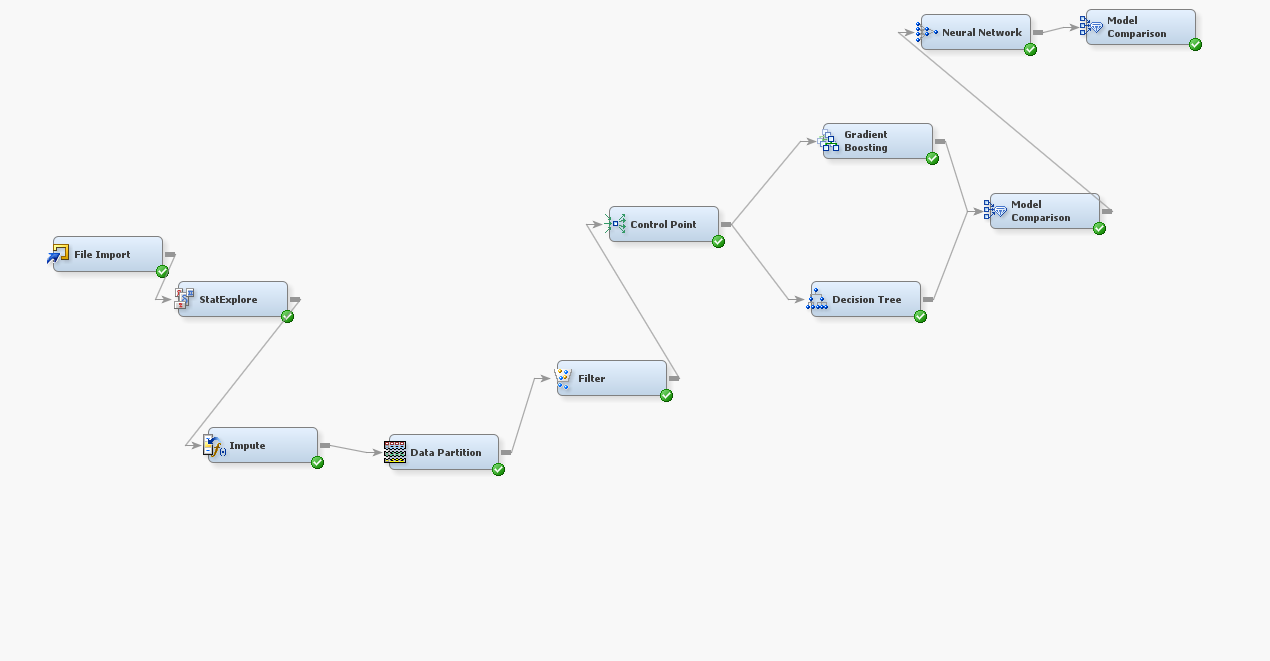
**High-level architecture diagram**

Diagram

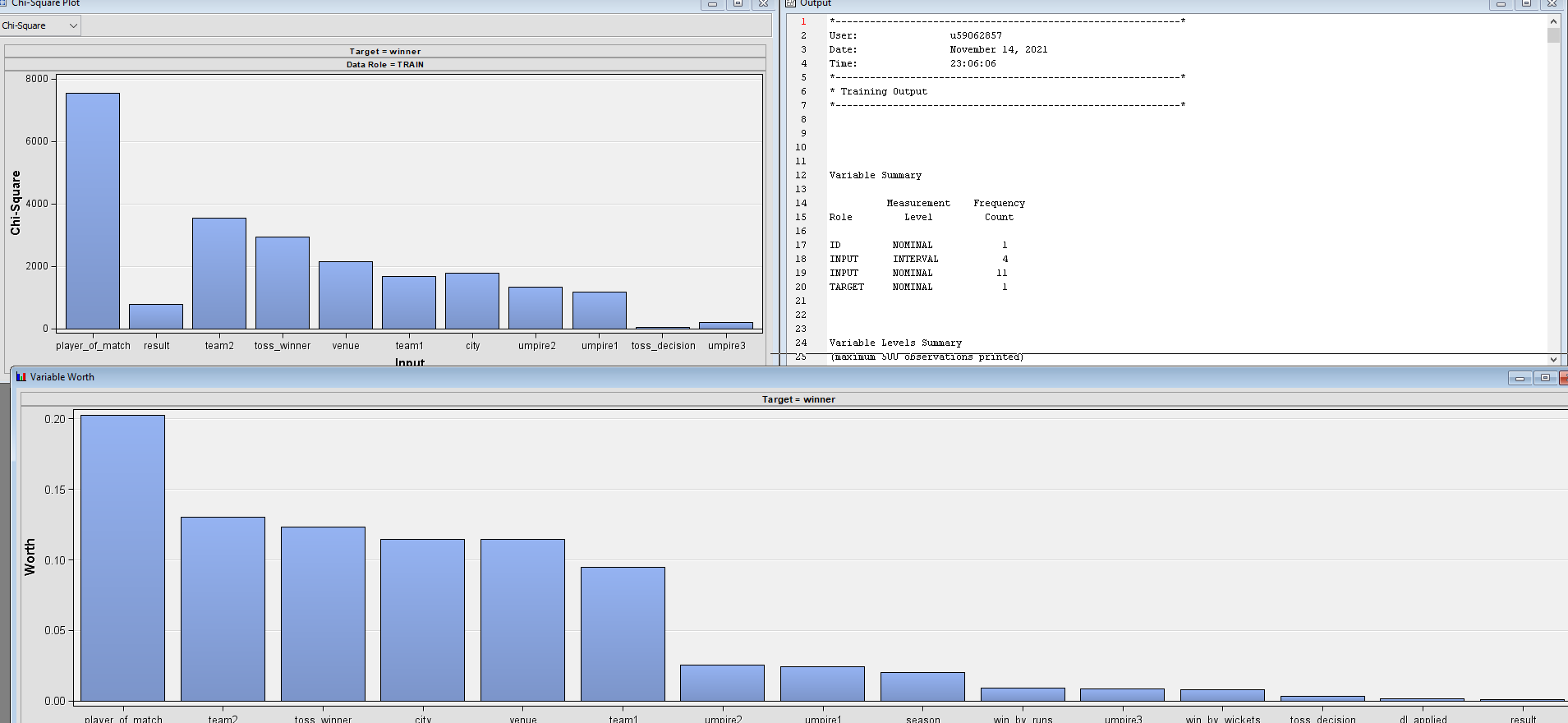
Description automatically generated

**B. EVALUATION**

As mentioned, we have used SAS miner to create and run the model. Below is the snippet of the model.



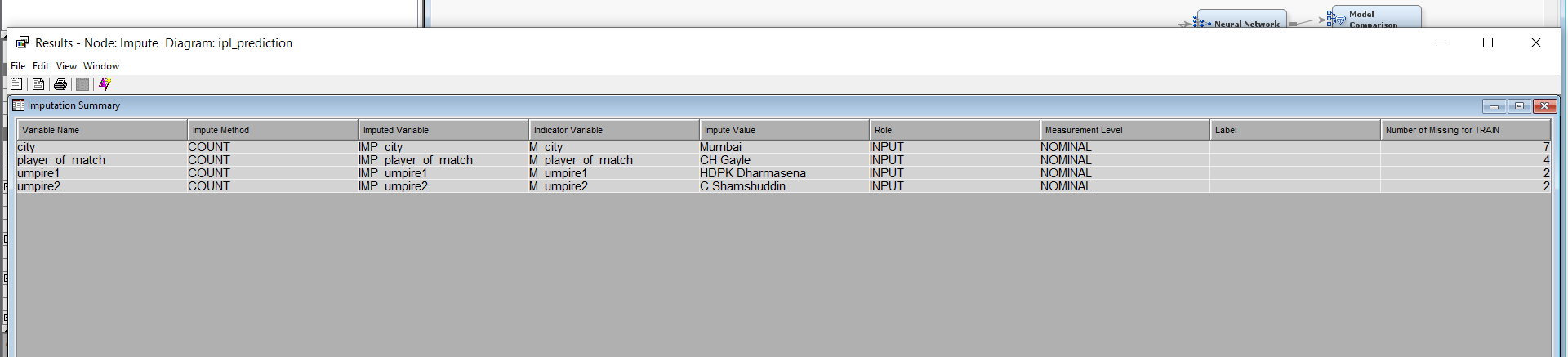
Stat Explore function tells us about the dataset. We can understand different statistics on feature variables in our dataset



Dataset has player of feature as most important factor in contribution for the prediction of target variable

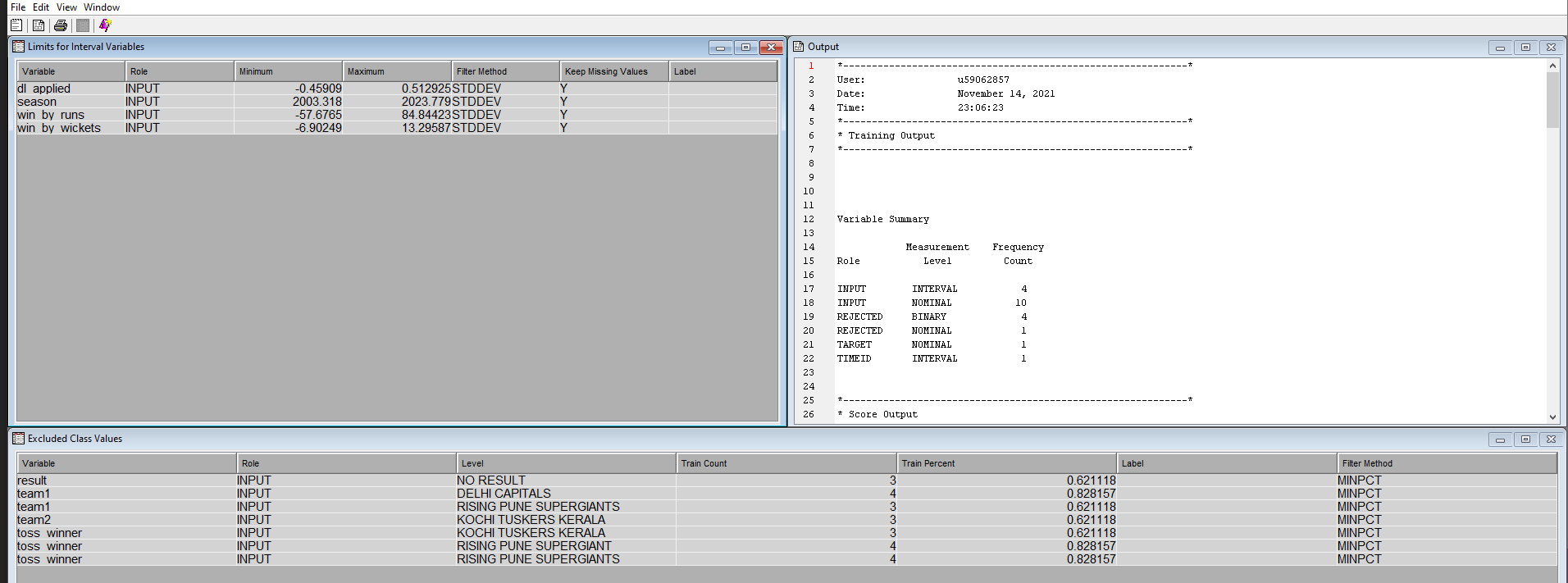
Impute function is used to impute the missing values of few variables.

Following values have been imputed for few of the features in the dataset



Then data has been partitioned such that 65% of the data is utilized in training the model, 25% in validation and 10% in testing the model.

We have also used filter method to remove some unused class values



We created both decision tree, Gradient boosting and Neural Network model to compare which model is doing better and pick the best one for data.

Model comparison function summarizes both the results and tells us which is better among the models we choosen

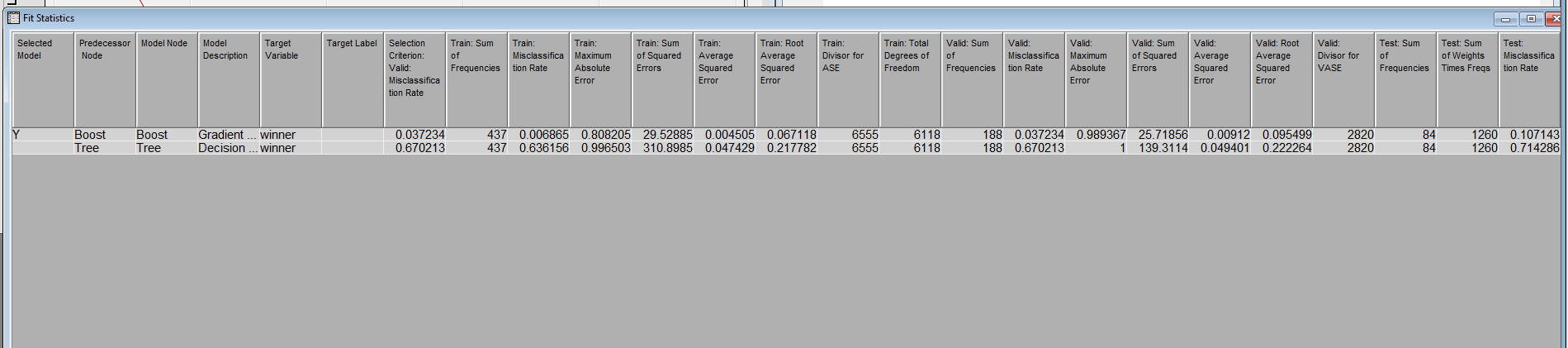
**RESULTS, CONCLUSION AND RECOMMENDATIONS**

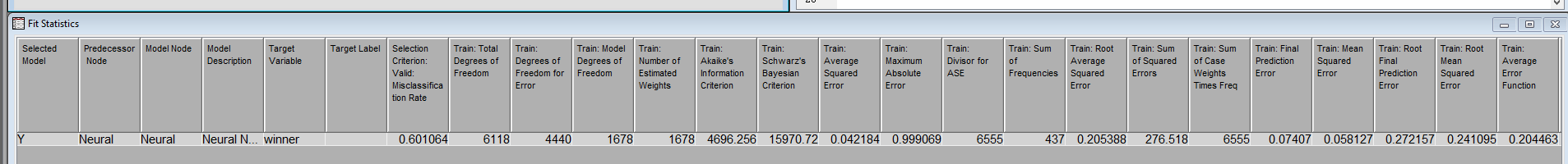
As part of this exercise, we used a predictive algorithm to help predict match winner and this can be used by any commercial market vendors. Currently we have worked on match level data and in future we will try to utilize ball level data and swing of the delivery for accurate predictions on winners of the match as well.

Below is the future work we are planning:

* Efficient Automated system for loading and connecting multiple datasets from multiple sources for model and results efficiency
* Exactly Predict the outcome of next ball based on past features of swing, bowler, speed, accuracy of batter
* Real time models for efficiently learning stats using amazon sage maker with xg boost and random forests whenever new data arrives (Incremental data loads on a particular frequency)

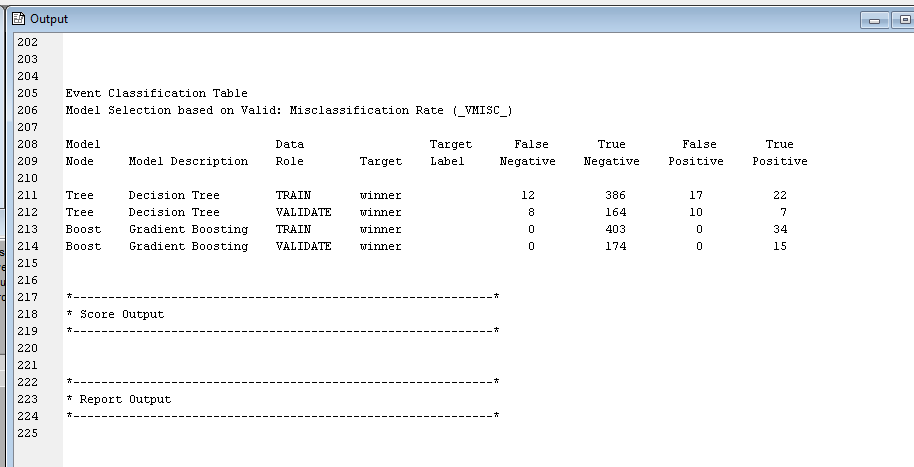
**Fit Statistics:**

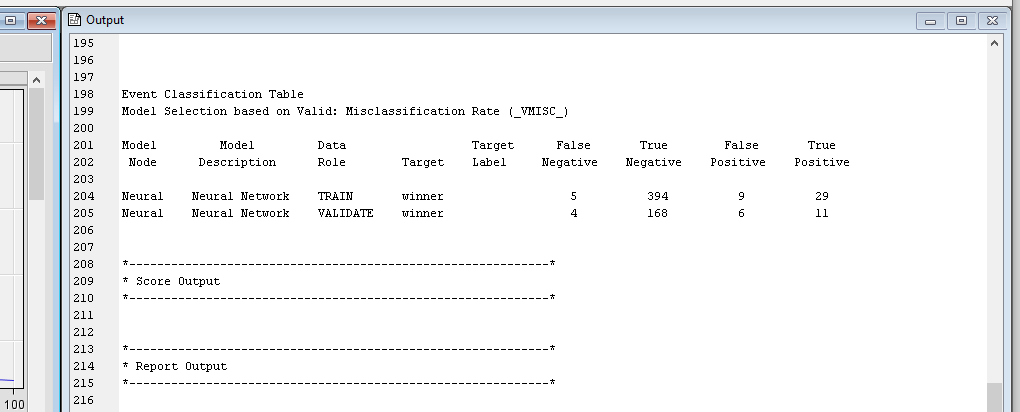




In the above Gradient boosting has less Misclassification rate and has nice prediction rate and hence better model compared to neural networks and decision tree algorithm

**Classification Table:**





Gradient is unable to predict the Negative values here. This also shows gradient boosting is better compared to neural networks and decision tree modelling. We can do predictive modeling using SVM, random forest, logistic regression, k nearest neighbor and do more detailed comparisons.

This model is successful in predicting the winner of a match given based on venue, toss, run\_margin and other factors

To make this model more reliable and realistic it should be trained with large amounts of data from real world and additional factors on different data attributes like deliveries level data and some other key attributes of a match can be added

REFERENCES

1. <https://cricsheet.org/downloads/>
2. <https://towardsdatascience.com/predicting-ipl-match-winner-fc9e89f583ce>
3. <https://www.kaggle.com/nowke9/ipldata>
4. <https://www.kaggle.com/ash316/let-s-play-cricket>