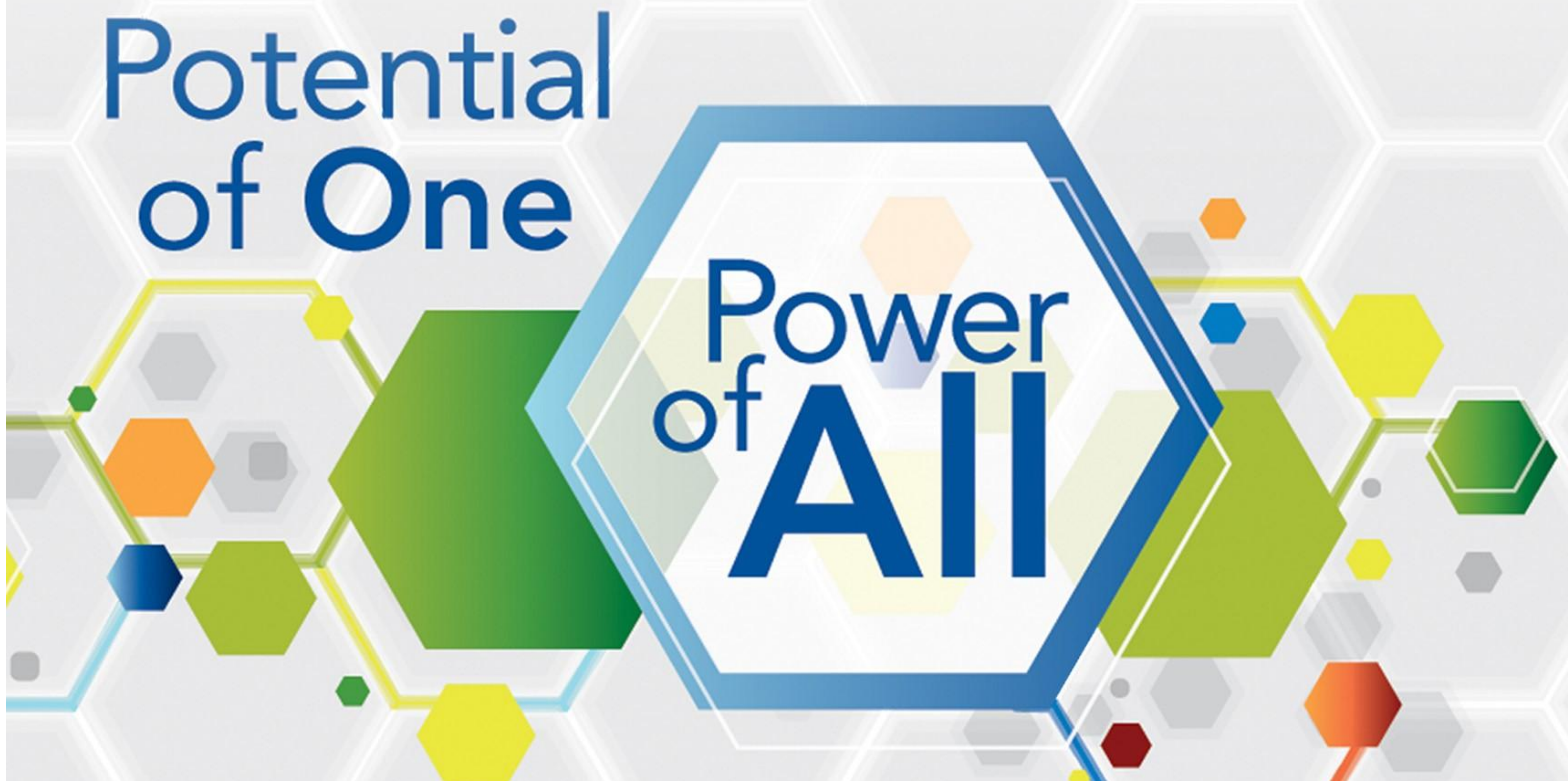
SAS GLOBAL FORUM

The Soccer Oracle: Predicting Soccer Game Outcomes Using SAS® Enterprise Miner ™



The Soccer Oracle: Predicting Soccer Game Outcomes Using SAS® Enterprise Miner™

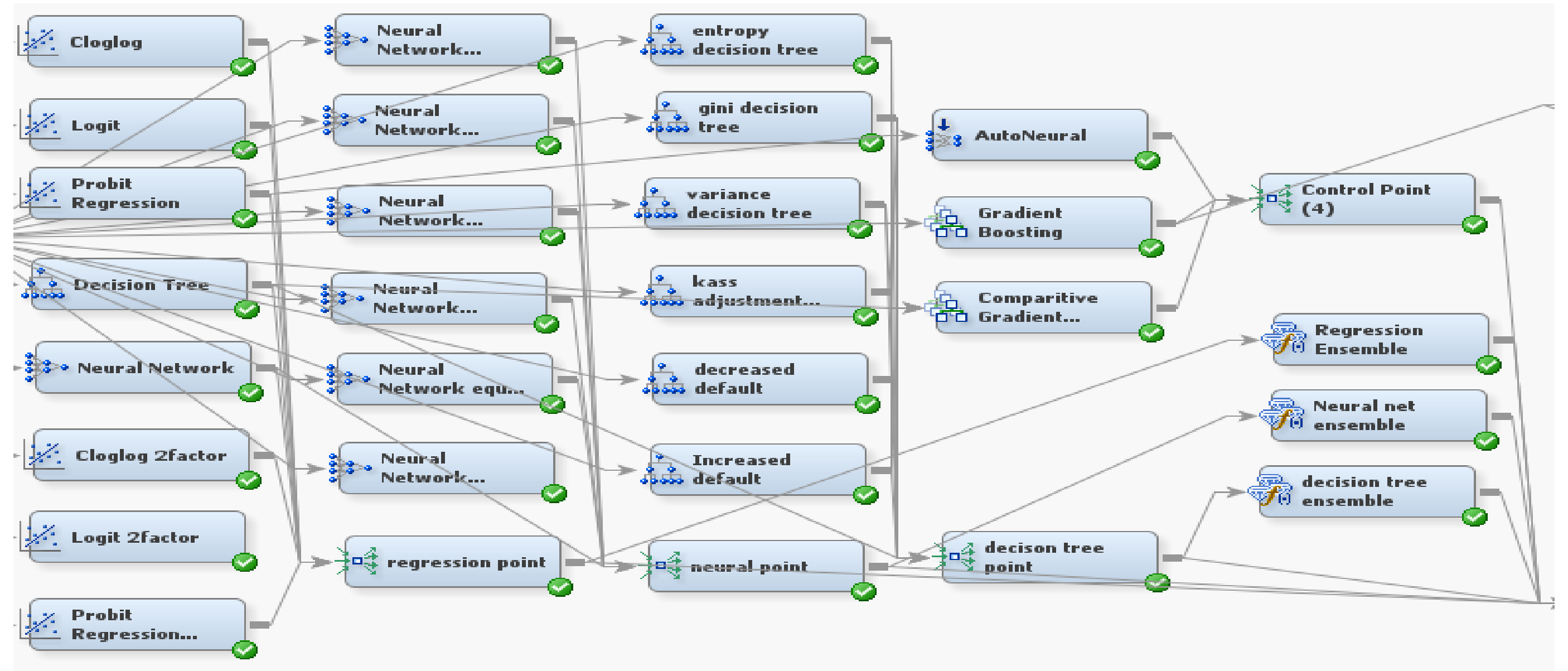
Vandana Reddy & Sai Vijay Kishore Movva
Department of MSIS, SAS and OSU Data Mining Certificate Program, Oklahoma State University
Faculty Advisor – Dr.Goutam Chakraborty, Professor (Marketing)

Abstract

The game of soccer is considered to be the greatest show on earth. Applying models to analyze sports data has always been done by teams across the globe. The film "Moneyball" has generated much hype about how a sports team can use data and statistics to build a winning team. The objective of this poster is to use the model comparison algorithm of SAS® Enterprise Miner™ to pick the best model that can predict the outcome of a soccer game. It is hence important to determine which factors influence the results of a game. The data set used contains input variables about a team's offensive and defensive abilities and the outcome of a game is modeled as a target variable. Using SAS® Enterprise Miner™, multinomial regression, neural networks, decision trees, ensemble models and gradient boosting models are built. Over 100 different versions of these models are run. The data contains statistics from the 2012-13 English premier league season. The competition has 20 teams playing each other in a home and away format. The season has a total of 380 games; the first 283 games are used to predict the outcome of the last 97 games. The target variable is treated as both nominal variable and ordinal variable with 3 levels for home win, away win, and tie. The gradient boosting model is the winning model which seems to predict games with 65% accuracy and identifies factors such as goals scored and ball possession

Objective

- To predict the outcome of soccer games using a predictive model algorithm of SAS® Enterprise Miner™ .
- Identify what variables play a significant role in influencing the target.
- Determine what would be the general statistics of a title winning team.



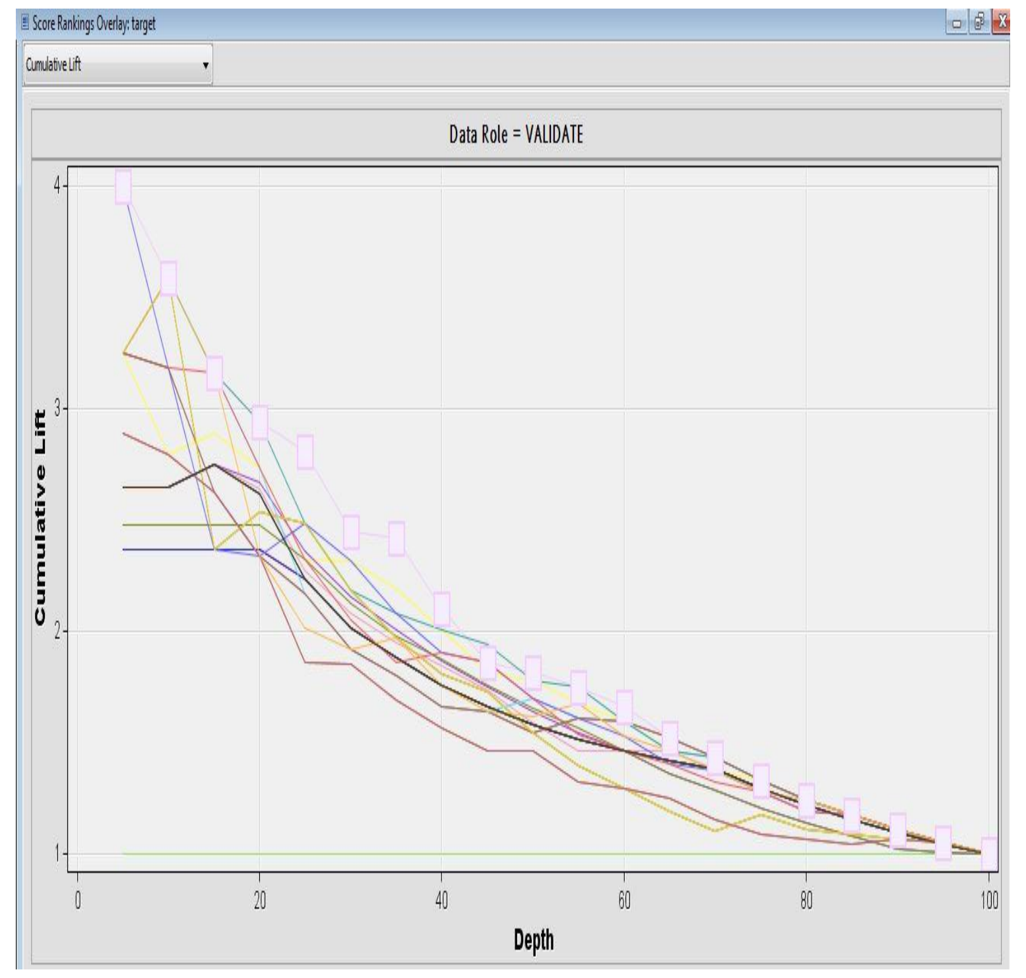
Method

Data Preparation:

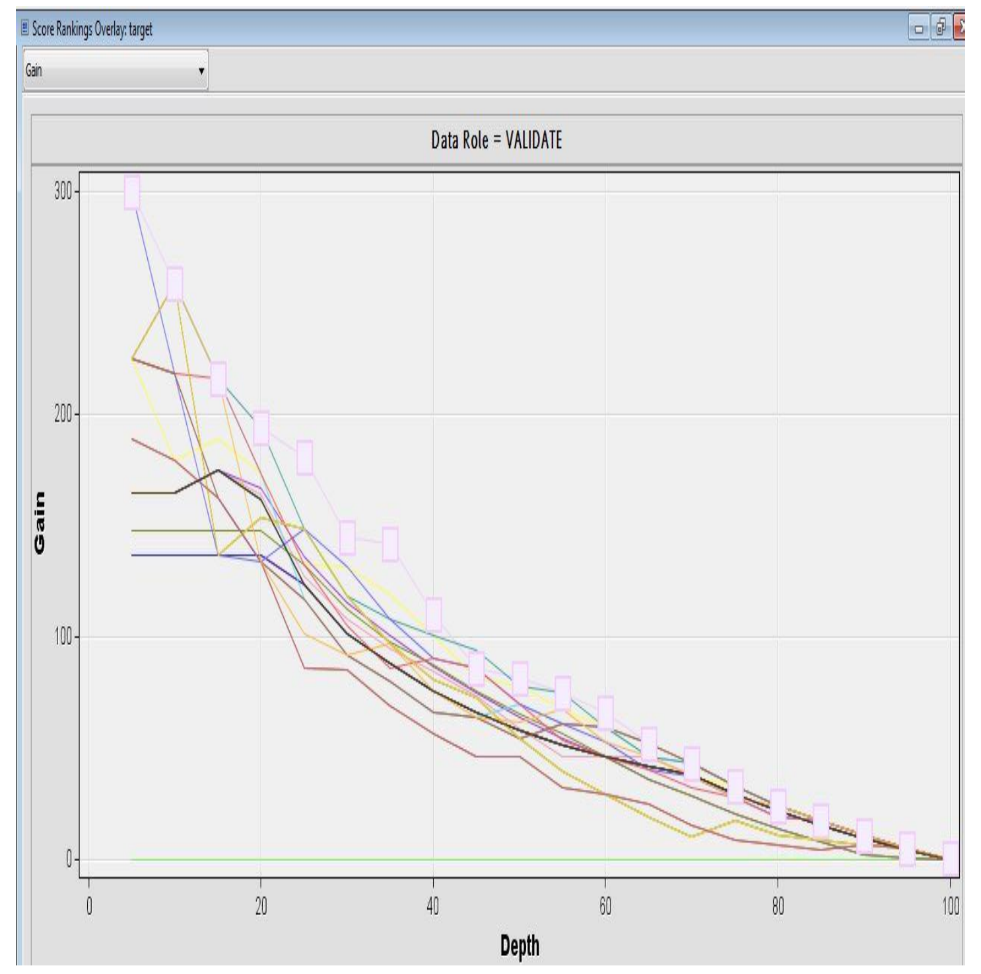
- The data was collected from various websites such as www.soccerstats.com and the data set contains 380 observations with 26 variables for each game,
- There are no missing values in the data set and hence no imputation is required.
- The variables in the data set have skewness and kurtosis which can be deemed normal by connecting a transformation node.
- The target variable is a nominal variable with 3 levels, "0"indicates a draw "1" indicates a home win and "2" indicates an away win.

Model Building:

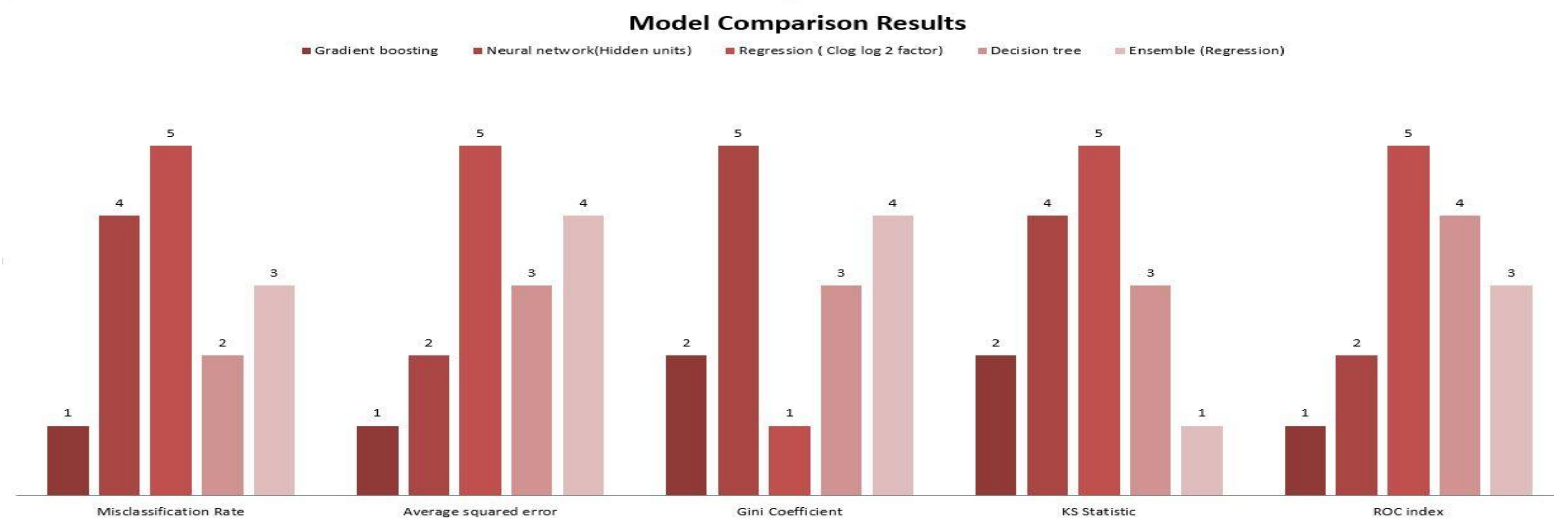
Before a model is built, the dataset is split into training and validation data. The training data is set at 75%, and the validation data is set at 25%. The first set of models run are various logistic regression models. Similarly, a set of neural networks have been run. As neural networks run a class of flexible, nonlinear regression models that are interconnected in a nonlinear dynamic system, they are useful for analyzing increasing volumes of data and identifying patterns in data. Decision trees are also run using various options available through SAS. The idea of using a decision tree is to recognize relationships in data that cannot be in general identified by a regression equation. The gradient boosting model has been used to create a series of decision trees that together form a single predictive model. The auto neural models are also run along with ensemble models, combining various regressions, decision tree, and neural network models together. A total of 26 models are selected and compared using the model comparison node after analyzing over a 100 models.



Lift Curve



Gain Curve



Model Comparison Results

- The models are compared first on their misclassification rate. The gradient boosting model has the least misclassification rate at 0.34.
- Next, the models are compared based on the average squared error, which gives the average of the squares of the difference between the actual observations and those predicted. Again gradient boosting fairs better with an average squared error of 0.164. The gradient boosting model also performs better in terms of other statistics such as the gini coefficient and the Kolmogorov-Smirnov statistic with values 0.79 and 0.39, respectively.

Conclusion

- Among most selection statistics, the gradient boosting model performs better than the rest with a 65% accuracy.
- Variables such as no. of away goals ,no. of red cards, home team position and shots on target play the most significant role in influencing the target.
- The title winning team Manchester United have scored 41 goals away form home, which is considerably higher than the average goals for the whole league, they average 10 shots a game converting 23 percent of the chances and they win 6 corners a game.
- The power of SAS EM is demonstrated with a 65% accuracy and can be successfully used in the field of sports Analytics.

References

- www.optastats.com
- www.soccerstats.com
- D. Karlis and I. Ntzoufras I. Statistical Modelling for Soccer Games.
- http://support.sas.com/publishing/pubcat/chaps/57587.pdf
- https://support.sas.com/edu/schedules.html?ctry=us&id=76
- www.eplindex.com
- www.football-data.co.uk
- D. Dyte and S. R. Clarke. A Ratings Based Poisson Model for World Cup Soccer Simulation

SAS® GLOBAL FORUM

Washington, D.C.

March 23 – 26, 2014