National College of Ireland

ANALYSIS OF PLAYER RATINGS BASED ON INTRINSIC FACTORS TO SUPPORT TEAM SELECTION

MSc Research Project
Data Analytics

Sidharthan Selvaraj
x15009114

School of Computing
National College of Ireland

Supervisor: Dr.Paul Hayes

National College of Ireland
Project Submission Sheet – 2015/2016
School of Computing

Student Name:	Sidharthan Selvaraj
Student ID:	x15009114
Programme:	Data Analytics
Year:	2016
Module:	MSc Research Project
Lecturer:	Dr.Paul Hayes
Submission Due Date:	22/08/2016
Project Title:	ANALYSIS OF PLAYER RATINGS BASED ON INTRINSIC FACTORS TO SUPPORT TEAM SELECTION
Word Count:	4835

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	11th September 2016

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

List of Figures

ANALYSIS OF PLAYER RATINGS BASED ON INTRINSIC FACTORS TO SUPPORT TEAM SELECTION

Sidharthan Selvaraj
x15009114
MSc Research Project in Data Analytics

11th September 2016

Abstract

Over the recent years, the sports industry has witnessed a rapid growth. A majority of the growth and popularity was contributed by football clubs all over the world. Players are the key factor off-field and on-field for clubs growth and popularity. A large group of football clubs still struggle to manage a player. A right set of players for a given game would change the match results for the clubs. The aim of this project is to predict the player ratings ahead of a given match. This rating predictions gives the scope to the football managers in selecting the right squad for a given competition as supportive element. The top European club players are taken into consideration for this supervised analysis. Individual player intrinsic factors and also the Players team performance for individual matches are recorded. For this regression problem, Random forest and deep neural networks algorithms are utilized. The predictions results proved a variance of 20.6 percentage and 19 percentage from the actual ratings.

1 Introduction

Popularity of the sport football has rapidly increased in recent years all over the world with a huge turnover, a number of fan followings and also in share market. A recent statistic in Germany proved that more than 12 million fans are watching football matches in the stadiums Tiedemann et al. (2011). This rapid growth has increased the individual clubs turnover year on year. A majority of the clubs turnover which are not limited from sponsors, television broadcasting, players and share market. A high percentage of the turnovers are spent on the transfer windows and players wages which is 76 percentages Tunaru et al. (2005). Hence it is crucial for the individual clubs to manage the players efficiently. A players performance is one of the important factors that changes a match result for a particular competition. Limited number of research are carried out on the players performance which motivated in conducting this research. Football clubs are more dependent on their managers and hence it becomes vital for the managers in choosing a right squad. The probability of winning a match also depends on the right set of starting eleven players. There are few more factors apart from players affect the chance of winning which are weather conditions, home support and right set of formations. A detail

analysis on player trends proved that players are more dominant since the players are not owned by a club rather they are contracted Tunaru et al. (2005).

A number of systems publish the football players statistics which are more reliable. All the football players and clubs statistics are extracted from the real time video systems. The players location detection and shot classifications from the video frame which are captured and then converted in to a number of data points. Cheng et al. (2013) proved statistically a number of the shots classification more precisely with the location prediction system. In the current trends, a majority of the football forecastings are done by the sports expert, betting market and the statistical models. The individual forecasters have a number of information on the players intrinsic factors and the extrinsic factors. Sports expert have more details on the players extrinsic factors such as the player injury, current skillsets, players form. The betting market purely focuses on the earlier match and player statistics in determining the odds. In this research the data are collected from a number of various reliable web sources .The web sources obtain the data from opta sports which captures the real time video streaming and delivers a number of match and individual players statistics. Various data mining algorithms have been applied for football predictions. Hucaljuk and Rakipović (2011) implemented nave bayes and regression models for determining the match outcomes. In this research, we will be implementing and comparing deep learning neural networks and random forest integrated with H20 system and without using H20 systems.

We will be discussing the following sections Literature review, Methodology, Implementation, Evaluation, Conclusion and Future work.

2 Literature Review

Various stochastic models and data mining models were inefficient in accurately classifying the sports outcomes. In the earlier research, a maximum of 80 percentage of accuracy was achieved in predicting the football match outcomes. A number of software tools are developed for the club managers to manage the players, one of them was Soccerscope2 softwareAbreu et al. (2010). The motivation behind this research is to find the football players ratings using a real time player intrinsic factors. Tunaru et al. (2005) in his research advocated that a players success rate would have a major impact on a clubs success. Predictions are widely used in predicting the game outcomes in other sports too. Orchard and Powell (2003) predicted the basketball match outcomes using neural network algorithm. In a real time scenario, the players efficiency is computed based on the number of intrinsic factors such as player match statistics, age and player positions. Tiedemann et al. (2011) in his research analysed the individual player statistics per match by considering factors like minutes played, total goals, tackle ratio and pass completion ratio. In this research we have also considered some of the factors for the predictions. Arnason et al. (2004) has classified the player intrinsic factors upon the various injuries, severity of the previous injuries, number of dribbles, sprint speed and body posture. Dvorak and Junge (2000) determine the players efficiency based on the long passes, short passes, headers and dribbles. Some of the other intrinsic factors that we have considered for this research are Total assists, Number of yellow cards, number of red cards, fouls suffered and fouls committed. Age, fitness level, injury

history and gender were taken into consideration as intrinsic factors by Taimela et al. (1990). The player position factor which is one of the key contributor in the research in determining the intrinsic index in determining the player ratings. Feess and Muehlheusser (2003) have considered the market price of the players as one of the factors in their research.

Football players are more prone towards injuries which is also one of the serious factor. Arnason et al. (2004) and team have analysed the risk factors for injuries by taking individual match sample on-field and training. In recent years, there has been a number of predictions on various sports. Gumm et al. (2015) in his kaggle competition predicted the outcome of NCA tournament. In the knockout stage predictions, Gumm et al. (2015) predictions were more precise. A multiple data mining techniques were implemented by Caruana and Niculescu-Mizil (2006) which includes support vector machines, decision tress, nave bayes and bagged tress. Williams et al. (2006) in his research has used a number of folds for the training and validation sample to decrease the computational time and increase the accuracy. Two samples of different subsets were considered for a comparison with all data and two subsets with various training and testing levels which increased the overall efficiency in Williams et al. (2006) research. UmaMaheswari and Rajaram (2009) researched on dimensionality reduction technique combined with association mining using a historical cricket dataset. A similar research on cricket ODI match outcome predictions using Bayesian network which had a higher accuracy than the PCA model Kaluarachchi and Aparna (2010). In Caruana and Niculescu-Mizil (2006) research the dimensionality reduction techniques were not implemented which could have reduced a larger number of least importance factors for the predictor class. Analysis of online players interaction and social diversity was analysed in Shim et al. (2011) research. In the research, the player statistics are collected from various servers upon which linear regression and bagging tree algorithms were applied Shim et al. (2011). In Huang and Chang (2010) research the author predicted the world cup 2006 outcomes at each individual knockout. A multilayer perceptron achieved a higher accuracy of 76.9 percentage with a variations in neuron till 40Huang and Chang (2010).

In a different research Joseph et al. (2006) researched on predicting the football results for the club Tottenham hotspur. A detailed study which analysis various splits and implemented a multi-classifier for the predictions Joseph et al. (2006). Miljković et al. (2010) in his research predicted the match outcomes for the basketball game using the earlier match statistics. Naive baiyes was utilized for this classification problem which under fitted the model Miljković et al. (2010). In the research, we are analysing the test results visually through as effective visualization model. Healey (1996) in his research analysed various effective colour based visualization. Three techniques are introduced which are colour distance, colour category and linearly separation. In a different research by Segel and Heer (2010) the effective visualization contradicted Colour based approach. Segel and Heer (2010) advocated an effective visualization based on the context of the subject. A number of case studies were dealt from a different industry perspective in segels research.

3 Methodology

This section discusses about the process flow of the research and software tools used for this research.

3.1 Process Flow

A six stage process flow model has been developed for this supervised learning. The process flow follows a knowledge discovery and data model. In the initial data extraction stage, the players statistics are collected from a number of top web sources which are ESPN, BBC Sport and who scored. The player statistics has been recorded from the ESPN and BBC Sports. Individual player statistics per match data are collected from the HTML tables using R Studio. The player statistics data are collected per team for the year 2015-2016 matches. A threshold value of more than 8 matches a player should have appeared for a given club is set for this research. The predictor variable Ratings data are extracted from the web source who scored.com. The player statistics data and the player rating variable values are merged in to a single data repository using MySQL joins. The second stage of the process is the data pre-processing which is done using the google refine and R Studio. Post pre-processing stage, the players data are transformed in according to the real player names, player position. In this stage, the cleaned data set is loaded in to IBM SPSS to check the descriptive statistics and the correlation among the various intrinsic factors. Data splits are done for individual players records from the pre-processed stage to building the model. The final three stages are performed in R studio post the data splits. Building the data modelling from the splits, validating the model and predicting the player ratings which are three stages. Random Forest and Deep neural network algorithms are applied to solve this regression problem. H20 system is integrated with both the algorithm for a better performance in terms of computational complexity. In the data modelling stage, the data is normalised prior to developing a training model using the training split. In the validation stage, the trained model is validated with a set of player validation records and then tuned using a number of folds. A less variance fold model is applied to the test data for predicting the variance. The predicted results are again denormalized to obtain the predicted ratings. The six stage process flow is shown in Figure1

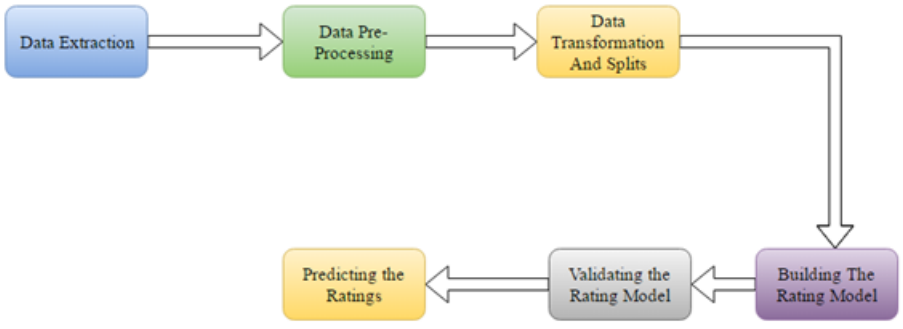

Figure 1: Process Flow Diagram

3.2 Software Tools Used

In this section, a list of software that are used for developing the individual artefacts for this research are listed below

1. IBM SPSS Statistics
2. R Studio 64 Bit
3. Google Refine
4. MYSQL
5. Tableau Desktop 9.2

4 Implementation:

This section discusses in detail about the dataset, pre-processing of data and the implementation of the algorithms used for this research.

4.1 Dataset

In this analysis, we are utilizing the data that is collected from ESPN, BBC Sport and who scored.com. The datasets from individual web sources are joined using individual players instance which is the player names. The dataset consists of 17413 observations, 13 attributes which are related to individual player match statistics records. The player match statistics factors include total goals, total minutes played, total assists, Appearance status as substituted or started, total shots, shots on target, fouls committed, fouls suffered, number of yellow cards, number of red cards ,teams result on home and away and player rating .The appearance factor is spitted in to two categories as substituted-in at a time and substituted-out at a given time .There are 3 more general players and club records which are players opponent team, competition type, players team. All the players statistics record are categorical values that varies between 1 to 10. Seasons are classified from the date range values. The predictor variable, the player ratings varies between the ranges of 5 to 10. The dataset has a combination of multivariate attributes. Figure 2describes the dataset

Figure 2: Data Description Diagram

4.2 Pre-processing of data

The raw data for the players are initially extracted individually for each clubs and merged into a single source file for player factors. Similarly, the raw data for players ratings are extracted with respective to individual clubs and then merged in to a single source file for player ratings. Both the source files are joined using player name as the reference key in MYSQL. At this stage, we have all the raw data as a single raw data source in MySQL.

The single raw data source is then loaded in IBM Spss statistics to check the descriptive statistics and correlations of the players intrinsic factors with the predictor variable ratings. A bi-variate correlation analysis proved a small number of factors are highly correlated. Few players and club records are excluded from this stage due to least importance in terms of correlation with the predictor variable. Post the analysis, the factors that were highly correlated with the predictor variable ratings are total shots, total assists, shots on target, fouls committed and total goals

4.3 Data transformation and splits

In this stage, a clean dataset with highly correlated variables with the predictor variable are taken into consideration. The entire cleaned dataset is loaded into google refine again where the players names, players team names and players opponent team names are modified and clustered to refine the right name. The text facet changes and clustering are the two actions that are performed in google refine. Post this transformation, the data is loaded into R studio where the data splits are performed. .The data splits were based on the date attribute and the number of matches appeared. For example, a player has appeared in 38 games between August 2015 to May 2016 where the training and validation splits are done for the 70 percent of 38 matches which is 26 matches and the remaining 12 matches for the test splits. The splits based on the date was utilized for the players with lower appearances. For example, a player has appeared in 10 games between August 2015 to May 2016 where the training and validation splits are done considering matches played between August to February which is more than a season and the remaining date range which is February to May matches for the test splits.

4.4 Implementation of the Algorithms

In this section, we will discuss about the algorithms that are implemented for this research. A deep neural network and random forest algorithms are applied to solve the regression problem. Post our initial algorithms performance, it is decided to integrate the H2O system which reduced the computational time in predicting the player ratings. A detailed description on individual algorithms performance in predicting the players ratings class are discussed.

4.4.1 Deep Learning Neural Network

A deep neural network was chosen for this regression problem in predicting the predictor variable Player rating. In R Studio, the packages that are utilized are neuralnet and H2O for the computation. Prior to applying the model on the training data, the data matrix is created and then the data is normalized in the scale of 0 to 1 in order to reduce a large variance in the dataset. The train dataset, validation dataset and test dataset are loaded in to the H2O. Prior to applying it is ensured

that all the three datasets attribute values are having a similar data types. A deep learning which is integrated with H20 package is applied in training the model. The number of hidden layers chosen for the feed forward network is two. A trial and error approach is followed in choosing the right set of neurons for the hidden layers. In the initial stage 30 neurons is chosen for the two hidden layers and then gradually increased the neurons to 40. The activation function that was used for building the model is Tanh. The computational time for training the model was less than 1 minute using the H2O system. The training model is then applied for validation dataset to see any overfitting and under fitting in the training dataset. Training model is applied on the test dataset for predicting the player ratings. Root mean square value of 0.19 was observed for the normalized scale ranges from 0 to 1. A plot was plotted to observe the variance from the actual to the predicted values post denormalizing the test data.

4.4.2 Random Forest

Random forest algorithm is one of the classification and regression techniques which builds a number of decision trees in the form of forest. In R Studio, the packages that are utilized are the random forest, caret and H2O for the computation. The three datasets are normalized prior to building the model. It is ensured that all the data types are in synchronous prior to building the model. The scaled data frame for the train, validation and test is then loaded into H2O system. The model is trained by increasing the number of trees gradually. Initially, a tree size of 200 is chosen and then gradually increases to 250. A saturation point for root mean square value was achieved when the tree size was increased more than 750 and hence a tree size was limited to 750. The trained mode l is applied on the validated set which is similar check that was performed on deep neural network. Predictions are done from the trained model and applied on the test data. The actual and predicted results are plotted to view the variance between the ratings. Root mean square value of 0.20 was observed for the normalized scale ranges from 0 to 1. The predicted players ratings are then denormalized in order to see the actual variance.

5 Evaluation

In this section, we will be evaluating the individual algorithms results and performance analysis of the models.

5.1 Deep Learning Neural Network

Initial results of neural networks with less number of neurons produced a larger root mean squared value. A number of various splits were attempted in the trial and error method to minimise the root mean squared value. 30:10 split produced a minimal error rates when compared to the other hidden layer splits. A minimal input dropout threshold value was set to .2 which reduced the error rates from 0.24 to 0.21. Apart from the trial and error approach a reason for choosing a 30:10 split was based on the variance in the number of independent variables and also size of the training dataset. An error rate of 19 percentages was achieved after tuning the model which was a better classifier in predicting the player rating class.

The observations proved that deep learning neural net classifier responded upon the adaptive rate and the momentum rate using H2O. The computational time that was taken for each of the tuned iterations models were still less than 1 minute.

When the deep neural network model was trained without using the H2O system, the computational time was more than 60 minutes with an error rate of 24 percentages. A downside of the deep learning neural network model is the difficulties in analysing the output in terms knowing the independent variable importance values. The computational time and accuracy of the deep neural network using H2O and without using H2O system is depicted in the Figure 3

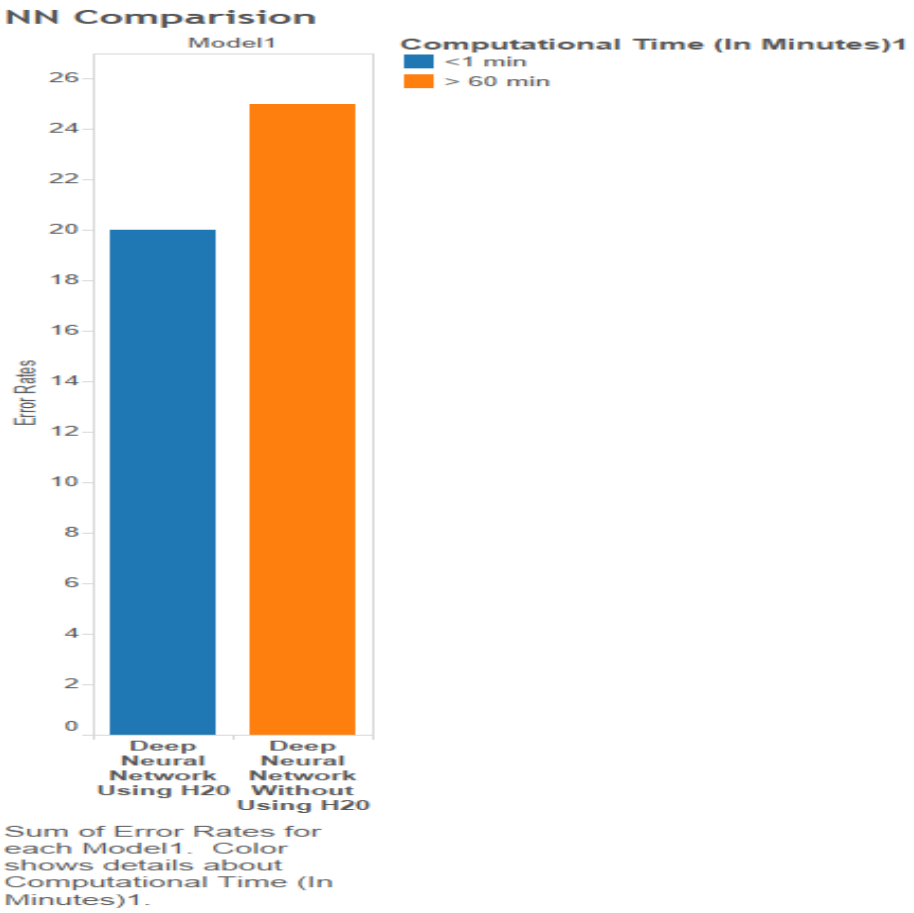

Figure 3: Comparison Of Two different Learning Models

The predicted and the actual results are compared for the top 20 highest predicted players ratings and top 20 for least predicted player rating. From the Figure4, we can infer that players with highest rating based on the accuracy .The model has predicted better for players Aaron cress well,Adrian,Adam Lallana ,Alexis Sanchez. Figure5 Infers the poorly rated players by the deep neural network model .On an average players Tiago, Thomas Vermalaen, Robert Hilbert are few players who are poorly predicted. The deep learning model has precisely predicted top 20 for high rated and low rated players showing a difference in variations between the two different sets of players.

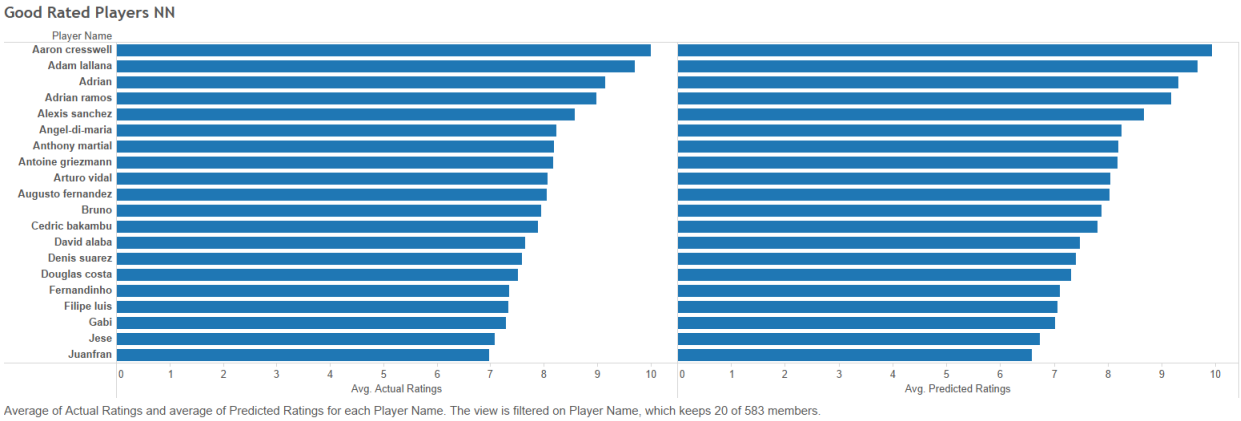

Figure 4: Good Rated Players Trained Using Deep Learning

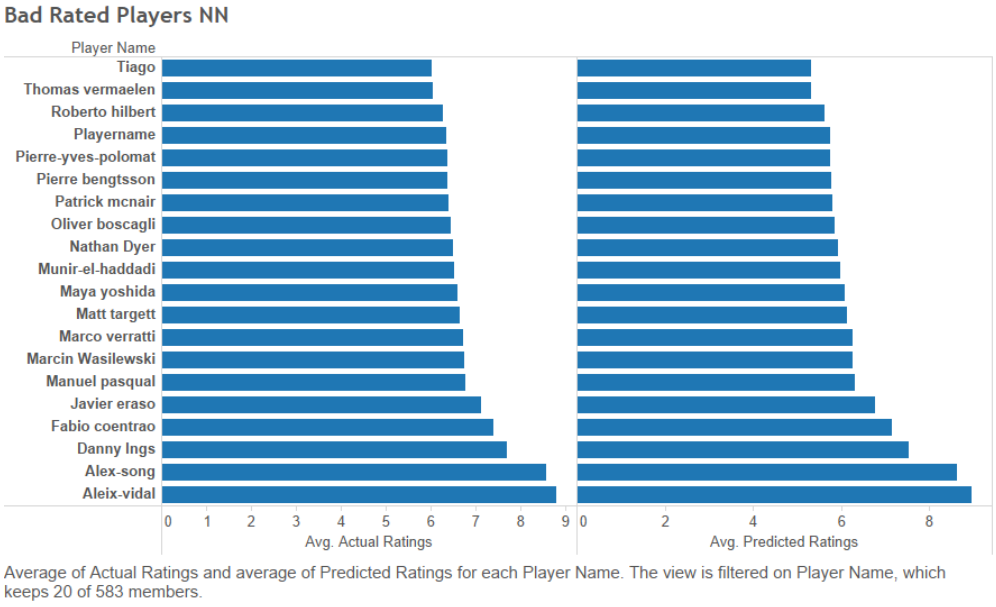

Figure 5: Bad Rated Players Trained Using Deep Learning

5.2 Random Forest Algorithm

Random forest algorithm is tested using the H2O system and without the H20 for predicting the player ratings. The initial performance without using performed poorly in terms of the error rates, tuning and the computational time. When the model was tested with the H2O system the performance was better in terms of response time, less error rates and flexibility in tuning the model. Similar to the random forest package, H2O package depicts the variable importance which is a key factor in this supervised analysis that differentiates from deep learning neural network. Figure6 infers the contributions of the variables in predicting the predictor variable Player Ratings. Variables Opponent team, Total Shots, Fouls suffered are the highest contributors in predicting the predictor variable and Red Cards, Yellow cards, Assists are some of the lowest contributors in predicting the player rating. In terms of the response time, the model initially took 40 minutes to run without

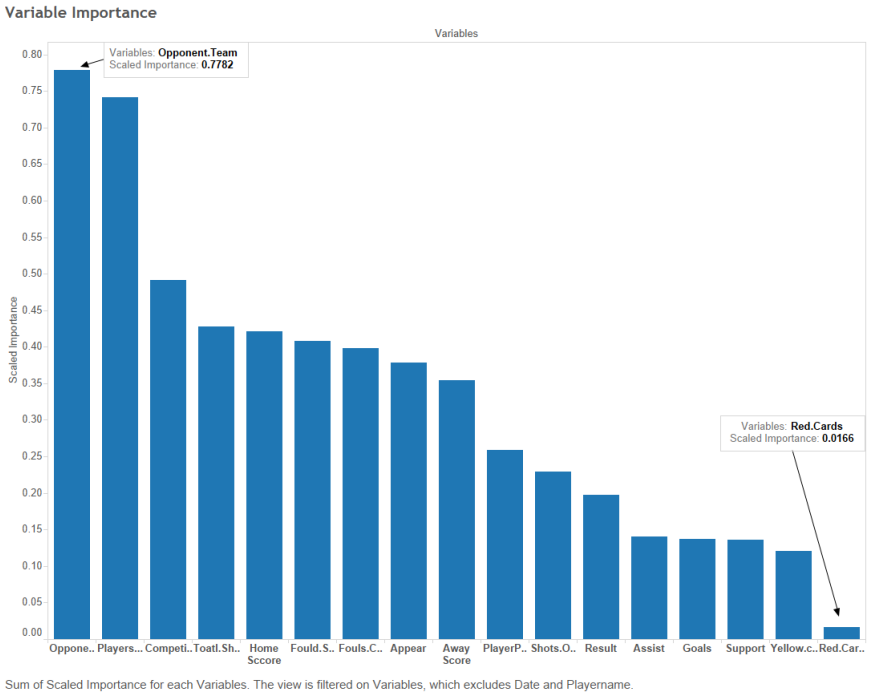

Figure 6: Variable Importance

using the H2O System with a tree size of 750. In the second analysis, when the model was integrated with the H2O system it took 5 minutes for a tree size of 750. A number of tree size was varied to initially decrease the root mean square value. In order to decrease the error rates, the entire dataset was cross validated using a number of splits by date and matches appeared. Low error rates were achieved for the 70:30 split. The predicted results proved that the good and bad predicted results were of a mid-ranged players .The predicted results variations for all the players were less when compared to the actual ratings. Figure7 depicts the random forest good rated players which ranged from 0 to 7.5. A stable value ranged from 7.0 to 7.2 was achieved for all the top rated players in this prediction. Similarly, Figure8 depicts the random forest bad rated players which also ranges from 6.9 to 7.2 which is a stable predictive range. This limitation is because the mid-ranged players have appeared in the less number of matches and there are less number of match records for the mid-ranged players.

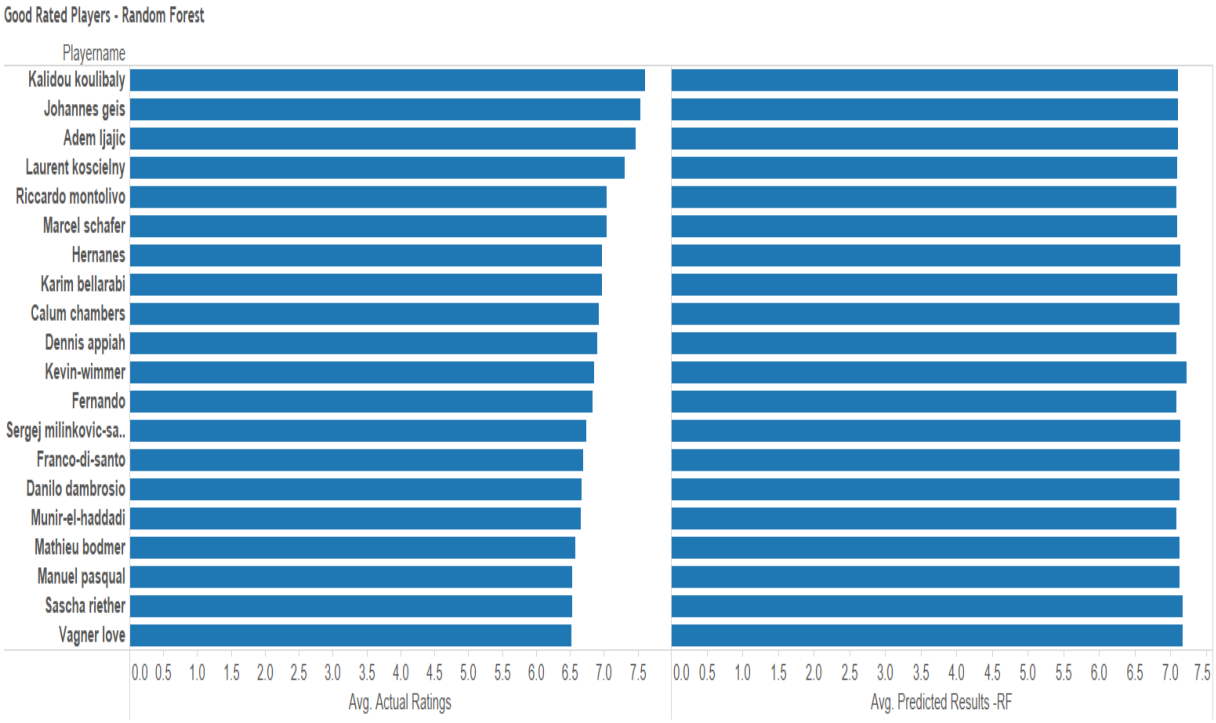

Figure 7: Good Rated Players: Random forest

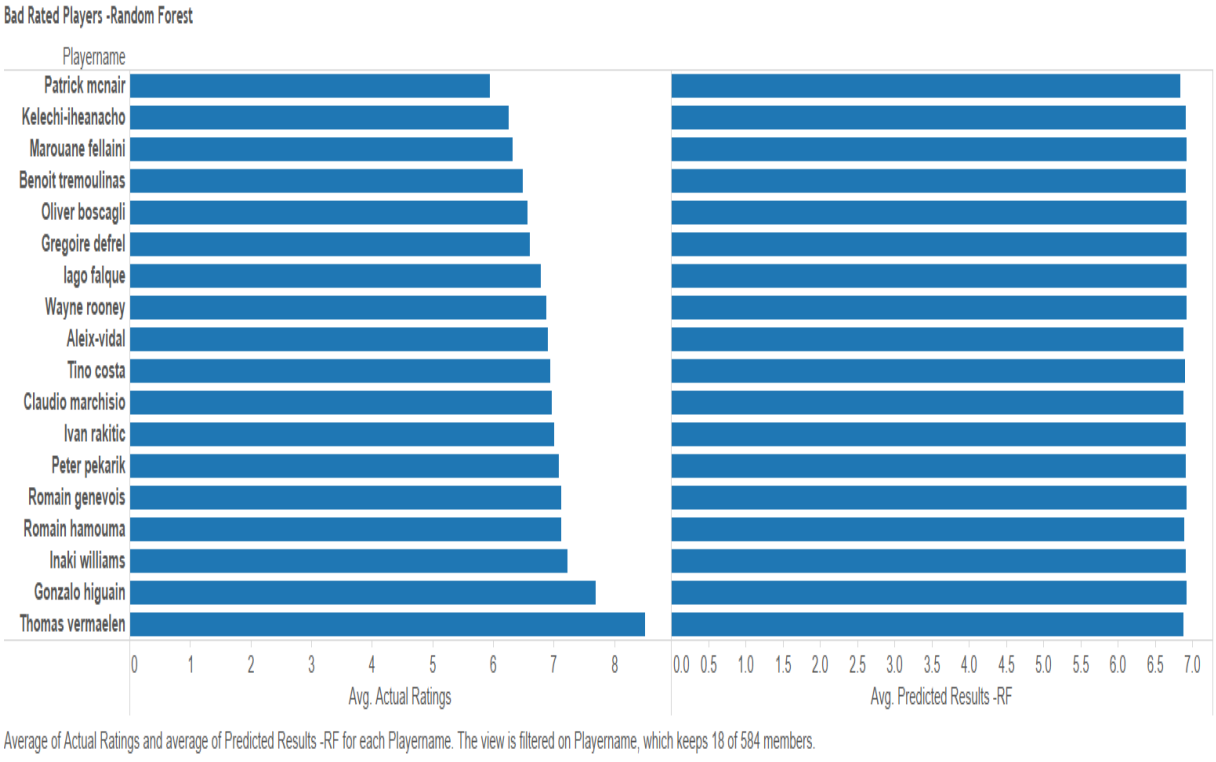

Figure 8: Bad Rated Players: Random forest

5.3 Performance Analysis of the Models

This section evaluates performance comparison in terms of the computation time and the error rates. In terms of the computation time, H20 system using the deep learning artificial feed forward network performed very well when compared to the random forest algorithm. Figure9 depicts the models performance in terms of computation time which clearly proves the model run time for the train data set. The performance testing was done by increasing the hidden layers and the number of trees for a similar splits respectively in deep learning neural network and random forest. The second performance evaluation is based on the error rates on the predicted results. Deep neural networks precisely were able to produce the results upon a number variations in the hidden layers and the threshold input dropout values and the hidden dropout values. The random forest algorithm couldn't replicate similar results precisely upon a minimal change in the number of trees and the tree depth value. Figure10 infers the errors rates for both the models on the predicted results

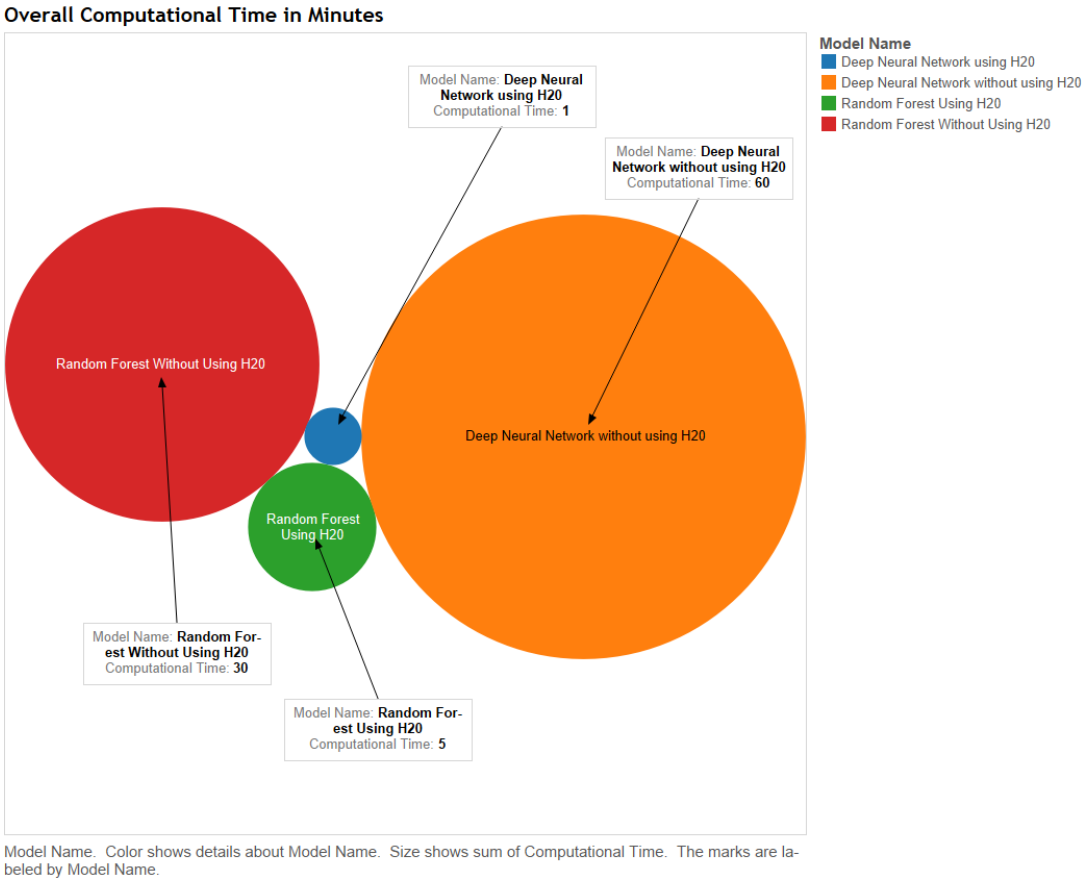

Figure 9: Overall Computation In Minutes

An error rate of 20.60 and 19 are achieved for random forest and deep learning neural network. Though the variations are less, the predictor variable values differ a lot for both the algorithms. As stated earlier, Random Forest algorithm has predicted a number of players between a range of values whereas the deep neural network has responded to a large variety of the rating values from a low range to a higher range which makes a better model for this regression problem.

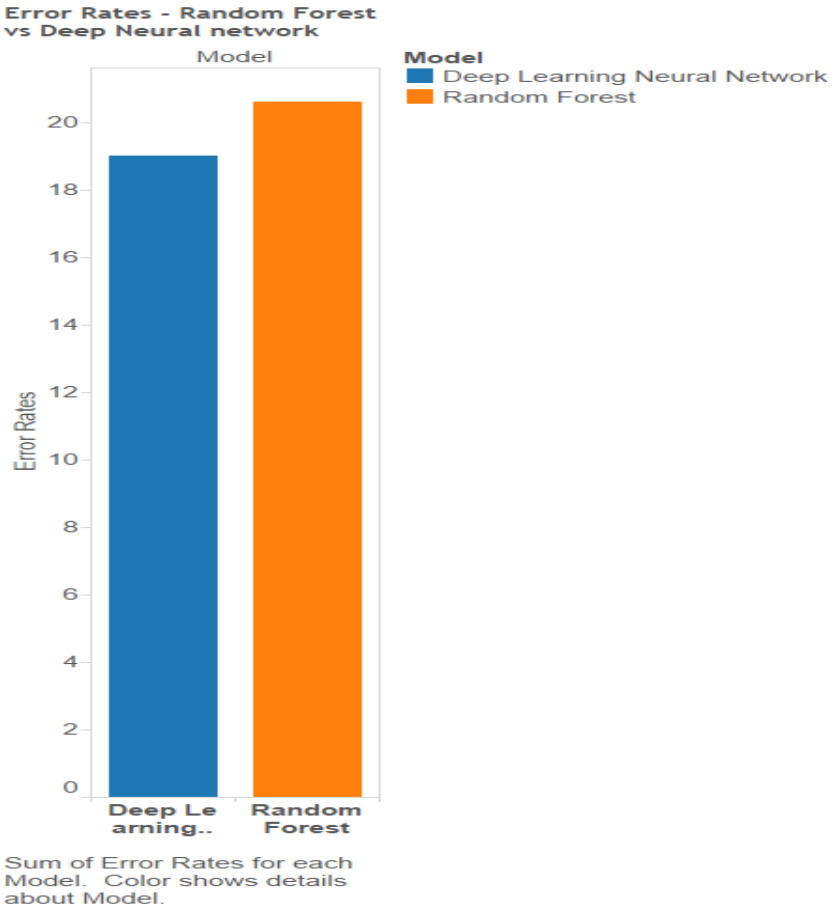

Figure 10: Error Rates - Random Forest Vs Neural Network

6 Conclusion and Future Work

Performance of the players are one of the vital factors in deciding the outcome of the match. Prediction of player ratings is very important for choosing the right squad ahead of the match which can support the club manager in deciding the best playing XI. This research focussed on predicting player ratings based on the key factors like players position, opponent team, total number of assists and total shots on target. Only a very few research has been done on the players performance by considering age, gender, injury, fitness levels, match statistics but this research has used the players intrinsic factors per game to correctly assess the players ratings using deep learning ANN and Random forest integrated with H2O. Two factors which are taken into consideration for performance evaluation of the models are computational time and error rates. Computational time taken by the models when integrated with H2O is very less when compared to running models without H2O integration. ANN and Random forest models using H2O have measured an RMSE of 19 and 20.6 respectively.

Since the number of low rated players have appeared in the minimal number of matches, the data for these players are very less in the dataset which did not help the model training for low rated players. One of the difficulties faced was splitting the dataset for individual players per match which can be automated in the further research. The large variance in the prediction results could have been avoided if the dataset have more number of player records for a given year. One of the future works is the automation of players dataset which can be fed directly

into the machine learning algorithms to suit a real time environment. Due to time constraint, regression testing is not performed on various software and hardware platforms.

Acknowledgements

I would like to thank my project supervisor Dr. Paul Hayes - School Of Computing for this valuable time and guidance. His continues guidance help me in tuning my research in right path. I coul not having has imagined a better ad visor and mentor for my research project. I would also like to thank my family- My parents and brother for supporting me spiritually thought all faces of my life.

References

Abreu, P., Moura, J., Silva, D. C., Reis, L. P. and Garganta, J. (2010). Football scientia-an automated tool for professional soccer coaches, *2010 IEEE Conference on Cybernetics and Intelligent Systems*, IEEE, pp. 126–131.

Arnason, A., Sigurdsson, S. B., Gudmundsson, A., Holme, I., Engebretsen, L. and Bahr, R. (2004). Risk factors for injuries in football, *The American journal of sports medicine* **32**(1 suppl): 5S–16S.

Caruana, R. and Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 161–168.

Cheng, Q., Agrafiotis, D., Achim, A. M. and Bull, D. R. (2013). Gaze location prediction for broadcast football video, *IEEE transactions on image processing* **22**(12): 4918–4929.

Dvorak, J. and Junge, A. (2000). Football injuries and physical symptoms a review of the literature, *The American Journal of Sports Medicine* **28**(suppl 5): S–3.

Feess, E. and Muehlheusser, G. (2003). Transfer fee regulations in european football, *European Economic Review* **47**(4): 645–668.

Gumm, J., Barrett, A. and Hu, G. (2015). A machine learning strategy for predicting march madness winners, *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on*, IEEE, pp. 1–6.

Healey, C. G. (1996). Choosing effective colours for data visualization, *Visualization'96. Proceedings.*, IEEE, pp. 263–270.

Huang, K.-Y. and Chang, W.-L. (2010). A neural network method for prediction of 2006 world cup football game, *The 2010 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8.

Hucaljuk, J. and Rakipović, A. (2011). Predicting football scores using machine learning techniques, *MIPRO, 2011 Proceedings of the 34th International Convention*, IEEE, pp. 1623–1627.

Joseph, A., Fenton, N. E. and Neil, M. (2006). Predicting football results using bayesian nets and other machine learning techniques, *Knowledge-Based Systems* **19**(7): 544–553.

Kaluarachchi, A. and Aparna, S. V. (2010). Cricai: A classification based tool to predict the outcome in odi cricket, *2010 Fifth International Conference on Information and Automation for Sustainability*, IEEE, pp. 250–255.

Miljković, D., Gajić, L., Kovačević, A. and Konjović, Z. (2010). The use of data mining for basketball matches outcomes prediction, *IEEE 8th International Symposium on Intelligent Systems and Informatics*, IEEE, pp. 309–312.

Orchard, J. W. and Powell, J. W. (2003). Risk of knee and ankle sprains under various weather conditions in american football., *Medicine and science in sports and exercise* **35**(7): 1118–1123.

Segel, E. and Heer, J. (2010). Narrative visualization: Telling stories with data, *IEEE transactions on visualization and computer graphics* **16**(6): 1139–1148.

Shim, K. J., Hsu, K.-W. and Srivastava, J. (2011). Modeling player performance in massively multiplayer online role-playing games: The effects of diversity in mentoring network, *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, IEEE, pp. 438–442.

Taimela, S., Kujala, U. M. and Osterman, K. (1990). Intrinsic risk factors and athletic injuries, *Sports Medicine* **9**(4): 205–215.

Tiedemann, T., Francksen, T. and Latacz-Lohmann, U. (2011). Assessing the performance of german bundesliga football players: a non-parametric metafrontier approach, *Central European Journal of Operations Research* **19**(4): 571–587.

Tunaru, R., Clark, E. and Viney, H. (2005). An option pricing framework for valuation of football players, *Review of financial economics* **14**(3): 281–295.

UmaMaheswari, P. and Rajaram, M. (2009). A novel approach for mining association rules on sports data using principal component analysis: For cricket match perspective, *Advance Computing Conference, 2009. IACC 2009. IEEE International*, IEEE, pp. 1074–1080.

Williams, N., Zander, S. and Armitage, G. (2006). A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification, *ACM SIGCOMM Computer Communication Review* **36**(5): 5–16.