PREDICTING AND CLASSIFYING MARKET VALUES OF SOCCER PLAYERS.

SUMMARY

This report is on a project that focuses on the transfer/Market values of soccer players by predicting the market value of each individual players and also classify such players value as high or low with respect to the median average of the market value. The report concludes that there are several factors that contributes to the market values of players and those factors are known as the variables, some of which are employed in this project.

INTRODUCTION

In recent years, sports data has been a popular topic for statistician because of the many important aspect it can be categorized into and how they have been able to draw many conclusions from these data through analysis. One of the most important time in soccer is during the transfer period where players are transferred from one club to the other through monetary transaction.

DESCRIPTION OF DATA

These data were collected from https://www.sofifa.com. This is a dynamic website where each feature need to be selected for it to appear on the page. Web-scrapping was used to extract this dataset from the website using selenium driver which works well with dynamic site. This driver then passes the page url to a beautiful soup object before scrapping.

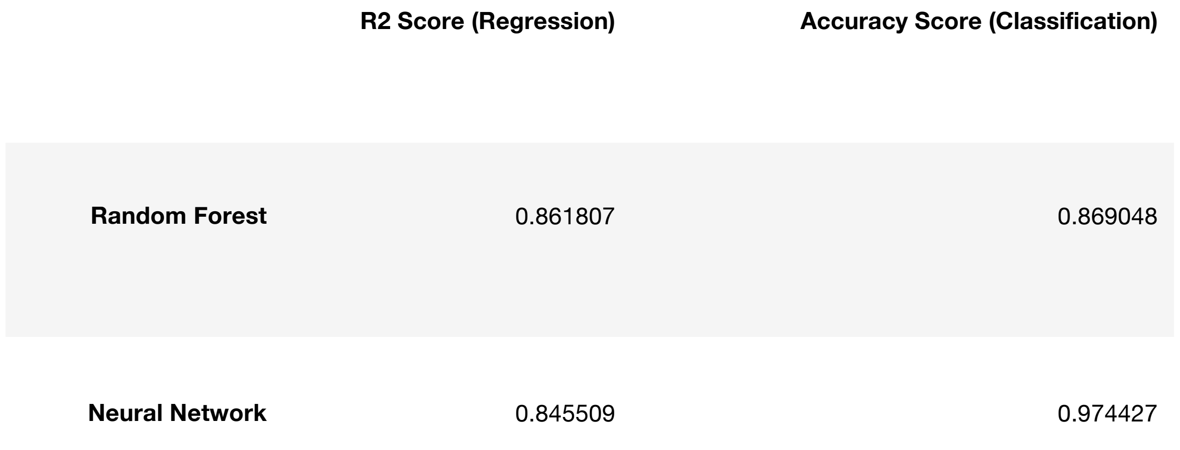
MODEL SELECTION

Two modeling techniques were employed in this project. Random Forest Regressor and Neural Network Keras Regressor were used for the regression problem while Random Forest Classifier and Keras Classifier of the Neural Network were also employed to the classification problem. A dictionary of hyper-parameters was passed to both of the Random Forest model and a grid search was used to iterate through the different parameters in other to choose the best. These parameters include a list of estimators, criterion and n_jobs for random forest, while a list of epochs, batch size, and number of neurons for each hidden layer was used for neural network. The selected variables are chosen based on predictors with a statistical significance of p value less than 0.05. The list of the variables that was used for modeling includes

['Age', 'Height', 'Preffered_foot', 'Overall Rating', 'Potential', 'Growth', 'Value', 'Wage', 'Tot_Attack', 'Crossing', 'Finishing', 'Head_Accu', 'Short_pass', 'Total_skill', 'Dribbling', 'Curve', 'FK_Accu', 'Long_pass', 'Ball_control', 'Total_move', 'Acceleration', 'Sprint_speed', 'Agility', 'Reactions', 'Balance', 'Total_power', 'Shot_power', 'Stamina', 'Strength', 'Long_shots', ‘'Total_mentality', 'Aggression', 'Interceptions', 'Positioning', 'Vision', 'Penalties', 'Composure', 'Total_defend', 'Marking', 'Standing_tackle', 'Sliding_tackle', 'Total_GK', 'Gk_diving', 'GK_handling', 'GK_kicking', 'GK_position', 'GK_reflexes'].

PREDICTION AND DISCUSSION

As stated in the last section that two models were employed in this problem, which are random forest and neural network. In other to obtain these scores, the dataset was split into a training set and a testing set with a 0.6 and 0.4 ratios respectively. The training set was used to train our model while the trained model was used to make a prediction on the test set. The figure below shows the summary of their respective prediction scores.

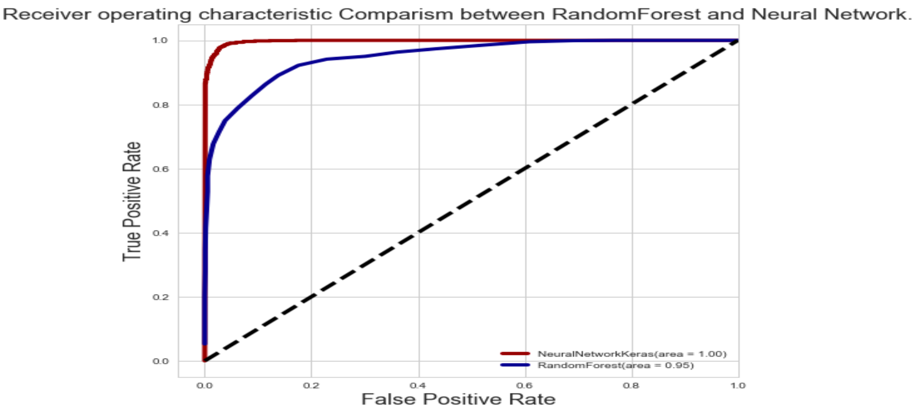


	R2 Score (Regression)	Accuracy Score (Classification)
Random Forest	0.861807	0.869048
Neural Network	0.845509	0.974427

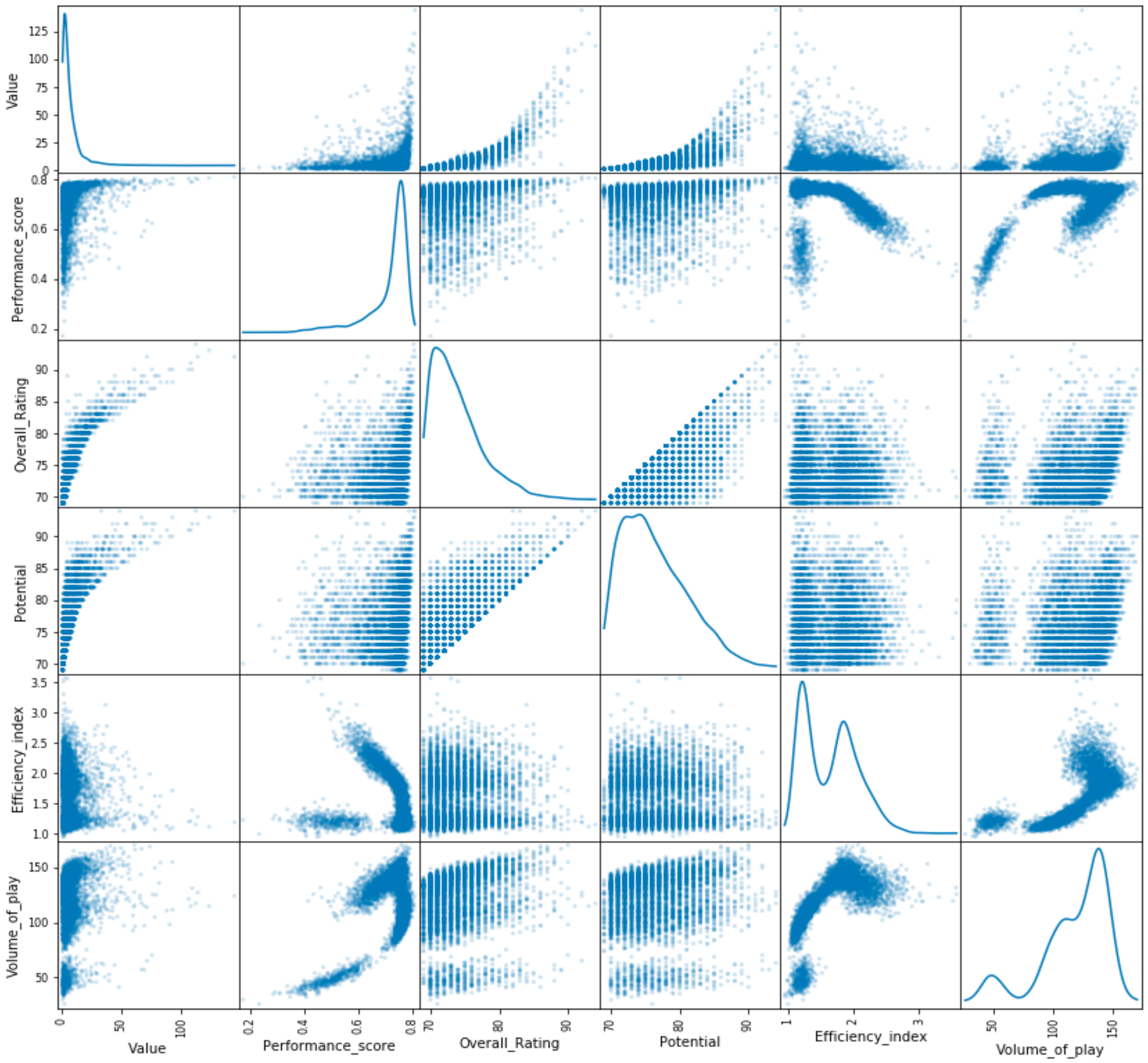
As we can see that the random forest did better than the neural network in the regression and vice versa for the classification problem. These results were based on the hyper-parameters that was used in the modeling process. This means that the hyper-parameters can be tweaked to produce better results. The figure below shows the confusion matrix for both models in the classification problem.

	Predicted Below medium (RandForest)	Predicted Above medium (RandForest)	Predicted Below medium (Keras)	Predicted Above medium (Keras)
Below Medium	1027	99	1095	31
Above medium	198	944	27	1115

This figure shows a merged data frame of the confusion matrices of each model. The table shows the **true positive** of 944 which is the number of predictions that the random forest model got right that were above the median of the market value of soccer players while 99 was the number of predictions that the model predicted as above the median value that were wrong (**false positive**). Also, the random forest model predicted correctly the number of **true negative** as 1027 which is the number of the market values of players that are below the median value but also predicted 198 values wrongly as the number of market values that are below the median value when they are actually the numbers above the median value. These wrong predictions are called the **false negative**. As for the neural network model, 1115 true positive, 31 false positives, 1095 true negatives and 27 false negatives predictions were made. This shows that the neural network model performed better than its counterpart. The figure below shows the Receiving operating characteristic curve (ROC) of both model where the neural network did better than the random forest model.

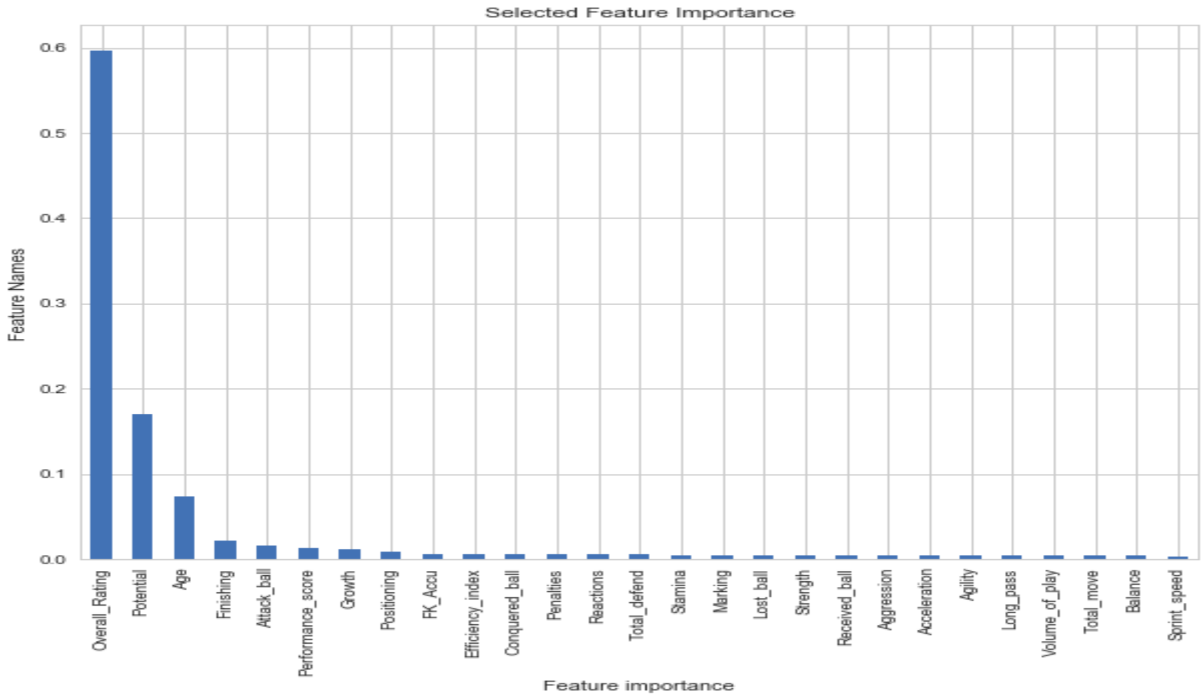


The red line as shown from the roc curve is that of the neural network with an area under the curve (AUC) of 1 and this shows how well the neural network performed, while the blue line represents the AUC of the random forest model with 0.95. Finally, the figure below is the pair plot of the true values of the test set against the predictions of both model.



CONCLUSION AND SHORTCOMINGS

The graph below shows the bar plot of the predictor variables based on feature importance of the Random Forest Regression model.



While, the predictor variables in this project contributes to the determination of the market values of soccer players, it will be great if,

1. more data points are added because of numerous number of soccer players with unique attributes. Adding more data points will help inform our model better.

2. The variables that were used did not totally reflect all possible factors that could contribute to a player’s market value. For instance, the number of goals scored by a player is a good variable that can affect the market value.

3. Players from different positions should have different criteria for judging their performance. For example, goals should be a fair assessment of a striker's ability, but not so fair for defenders or goalkeepers. If we could include data such as number of assists, number of clean sheets, tackles per game, etc., it would be much better for the evaluation of midfielders (assists), defenders (tackles) and goalkeepers (clean sheets). Again, this was not possible since detailed performance data could only be found for current season.