## STAT 515 FINAL PROJECT REPORT

***SOCCER ANALYTICS – FIFA 19***

STATISTICAL ANALYSIS AND VISUALIZATION OF SOCCER PLAYER’S DATA

By

Vishnu Deepak Sadayampatti Udaykumar

# INTRODUCTION

## ABOUT THE GAME:

FIFA (**Fédération Internationale de Football Association**) is a non-profit international governing body of association soccer, futsal, beach soccer and e-soccer. FIFA 19 is a licensed soccer simulation video game developed by EA Vancouver as a part of Electronic Arts’ (EA) FIFA series.

## DATASET:

The game contains detailed attributes of 18,207 real life players registered across various continents such as America, Europe, Asia and Africa in the database. The dataset was obtained from Kaggle and contains approximately 90 attributes extracted from the FIFA database.

The

# RESEARCH QUESTIONS

There has always been a debate among the gaming community as to how the gaming franchise assigns the “Overall” rating of a player in the game. To understand the above scenario, can a Model be built that predicts the variable – “Overall” rating of a particular player based on the numerous traits/attributes associated with the player? What are the most important variables in the predicting the Player’s Rating? How accurate are these Models? Can an alternate Model predict the response variable with a better accuracy?

The managers of a sports club (soccer club in this case) need to identify/scout players for replacement in case of an injured player and to build a team respectively. This requires the Managers to identify players of a certain skillset based on the playing position in the field. Can a model be able to classify the players based on a similar skillset so as to make it easy for the Managers to go out in the transfer market and buy the player.

# MULTIPLE LINEAR REGRESSION

The first step in building a Linear Regression model was to perform the normality test of the response variable “Overall” by both graphical (Histogram) and statistical method. The skewness measure turned out to be 0.084 which is close to zero. As a general rule of thumb, if the skewness is between -0.5 to + 0.5, the distribution is approximately symmetric. Based on the Histogram below for the response variable and the skewness measure, it can be concluded that the response variable is normally distributed

A picture containing text

Description automatically generated

Figure 1: Histogram for the Response Variable - Overall

As a next step, the correlation between the response and predicted variables were plotted. The correlation plot is included in Appendix B for reference. It can be seen that there is a significant correlation between the predictor and response variables. Post this the dataset was split into training and test data. Random sampling split 80% of the data into the training set with 20% in the testing set. It is important to divide the data because a model should not train and test on the same data. If the data was not split, the model would fit on all of the data and any measurement of accuracy would only pertain to the data with no indication on how well the model does on new data.

Fitting the Model:

Using the lm() or linear model function in R on our training dataset, we created a multiple linear regression model with “Overall” as the response variable and 35 independent variables as denoted in Appendix A. Based on the results, we were able to determine that several variables did not have a statistically significant relationship with our response variable.

A close up of text on a white background

Description automatically generated

Figure 2: Summary of Linear Regression Model

It is seen from the above Figure that there are 4 variables namely Weak Foot, Dribbling, Curve, Free Kick accuracy and Long Passing which are not statistically significant based on the p-value. As a next step the aforementioned variables were removed, and the Linear Model was built on the train dataset and the Model was used to predict the response variable in the Test dataset.

Evaluating the fit:

The R-squared for the Linear Model was 0.88 which means that the model explains about 88% of the variability of the response data around its mean. The root mean square error (RMSE) of the Model is 2.406 indicating how far from the regression line the data points are. To further evaluate the model, the summary diagnostic plots were used, which did not indicate any assumption had been violated

![A close up of a map

Description automatically generated]()

Figure 3: Summary Plots for the Linear Regression Model

Residuals vs Fitted: The residuals vs fitted plot is intended to identify residuals with non-linear patterns, such as a non-linear relationship between predictor variables and the outcome variable. As seen from the image, there are random fluctuations of the residual values out zero indicating linearity

Normal Q-Q: Based on the QQ plot, the residuals are normally distributed. Data Points closely follow the straight line at a 45% angle upwards

Scale Location: This plot checks for homoscedasticity (uniform variance), showing if residuals are spread equally along the ranges of predictors. We would expect to see a horizontal line with randomly spread points roughly equal around the line. There are fewer points at the high and low ranges of our data, but our variance is fairly constant.

# RANDOM FOREST REGRESSION

The accuracy of the Linear model based on the R-squared value was 88%. On top of the Linear Model, Random Forest regression model was built to predict the same Response Variable “Overall” rating of the player to improve the accuracy of prediction and find the important predictors. The variables considered for the model in addition to the predictors used in the Linear model are Categorical variables such as Preferred foot, Position, Work Rate, Nationality and Club.

The dataset was split into training and test data. Random sampling split 50% of the data into the training set with 50% in the testing set. Two versions of the Random Forest Model were built on the training dataset based on different values of mtry (Number of Variables available for splitting at each tree node). The number of tress built were 500 in number

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.NO | mtry | Number of Trees | MSE (Train) | MSE (Test) | % Variance Explained |
| 1 | 6 | 500 | 0.6791805 | 0.8033 | 98.62 % |
| 2 | 12 | 500 | 0.4647257 | 0.6602 | 99.05 % |

Table 1: Summary Random Forest Model Performance

It is seen from the above table that the best of the two models based on the accuracy and Mean square error is the 2nd Model (i.e. mtry =6). The model has an accuracy of 99% in predicting the response variable with a Mean square error of 0.6602.

![A close up of a map

Description automatically generated]()Figure 4: Important Variable for Predicting Price

Figure 4 shows that how much the MSE would increase is a variable was permuted, or randomly shuffled. The higher the %IncMSE, the more important the variable. The Random Forest Model determined the most important variables for predicting the response variable “Overall” as Value of the Player (In million Euros), Age, Release Cause of the Player (In million Euros).

# K -MEANS CLUSTERING:

K Means algorithm is one of the simplest unsupervised Machine Learning algorithms. It aims to cluster a collection of data points based on certain similarities. The task here is to group players of similar skillset based on the individual attributes of the player. The dataset contains 27 different playing positions in the field. In order to simplify the exercise, the numerous playing positions were grouped to four categories based on the table below

|  |  |
| --- | --- |
| Existing Playing Position (in the dataset) | Modified Playing Position |
| GK | GOALKEEPER |
| CB, LB, LCB, LWB, RB, RCB, RWB | DEFENDER |
| RM, RDM, RCM, RAM, LM, LDM, LCM, LAM, CM, CAM, CDM | MIDFIELDER |
| ST, RW, RS, RF, LW, LS, LF, CF | FORWARD |

Table 2: Categorization of Playing Position

STEPS: First, the Goalkeeper position was removed along with players without a position listed. The next step was to select only numeric player attributes – Player Value, Wage and Overall Rating have been removed from the dataset so that these variables don’t sway the groupings and allow for the clusters to contain like-for-like players based of their skillset alone. The number of clusters were set to 8 based on trial and error method.

|  |  |  |  |
| --- | --- | --- | --- |
| **CLUSTER** | **DEFENDER** | **FORWARD** | **MIDFEILDER** |
| 1 | 1489 | 0 | 20 |
| 2 | 8 | 460 | 1336 |
| 3 | 1186 | 4 | 319 |
| 4 | 0 | 1433 | 116 |
| 5 | 1105 | 59 | 1480 |
| 6 | 1184 | 13 | 657 |
| 7 | 461 | 34 | 1186 |
| 8 | 5 | 1031 | 1147 |

Table 3: Cluster Data