INFSCI 2160 DATA MINING PROJECT REPORT

A PREPRINT

Tshering Sherpa
University of Pittsburgh
School of Computing and Information
Pittsburgh, Pennsylvania
tts12@pitt.edu

Yanru Li
University of Pittsburgh
School of Computing and Information
Pittsburgh, Pennsylvania
yal125@pitt.edu

Yuyan Li
University of Pittsburgh
School of Computing and Information
Pittsburgh, Pennsylvania
yul225@pitt.edu

July 16, 2020

ABSTRACT

This project analyzes the value FIFA players worth. We obtained the data from kaggle dataset which provide many information regarding players' skill. As most of these information related to players' performance and reputation, we assumed that our prediction will be confident. As there are many standard to decide the top valued players, we decide to treat players with over 100 million euros as top valued players. To predict the top valued FIFA players, we trained several models including XGBoost, LightGBM, CatBoost, and Logistic Regression. To see which factors have most important contribution on the value, we use the SHAP value to show the importance of each factor. As we expected, our models all performs good and well-fit, we choose the best model which generate highest AUC to be used to predict top valued FIFA players.

1 Introduction

FIFA stands for the Fédération Internationale de Football Association. FIFA is an association governed by Swiss law founded in 1904 and based in Zurich. It has 211 member associations and its goal, enshrined in its Statutes, is the constant improvement of football [1]. In the United States of America it is popularly known as soccer. Our project is going to analyze the value of FIFA players. There are many factors that can be used to decide a player's value, such as potential, international reputation, agility, work rate, etc. Our goal is to develop a model to select top valued players by predicting their values from these factors. In this project, we've used several models to train the data and the final model is able to predict results well.

2 Dataset

2.1 Data Description

Our data is "FIFA 19 complete player" from kaggle datasets. Our dataset is the latest edition of FIFA, and the data is scraped from https://sofifa.com/. There are 13725 rows and 87 columns in this dataset. There are several data types include integer, object and float. The dataset includes players' name, ID, age, vision, nationality, clubs they belong to, their position, scores they got in each exercise, wage, and value they worth, etc. This dataset has some missing values, and we replaced them with the mean of the attributes (columns).

https://www.overleaf.com/9557785675ngfbdckswjqv

2.2 Data Analysis

Since there are many categorical variables, we transferred them into numerical variables, which is facilitated to analyze. We were trying to find the top valued FIFA players. There are many standards to decide which value of FIFA player is the top. In our research, we decide to take player whose values are more than 100 million euros as top valued FIFA players.

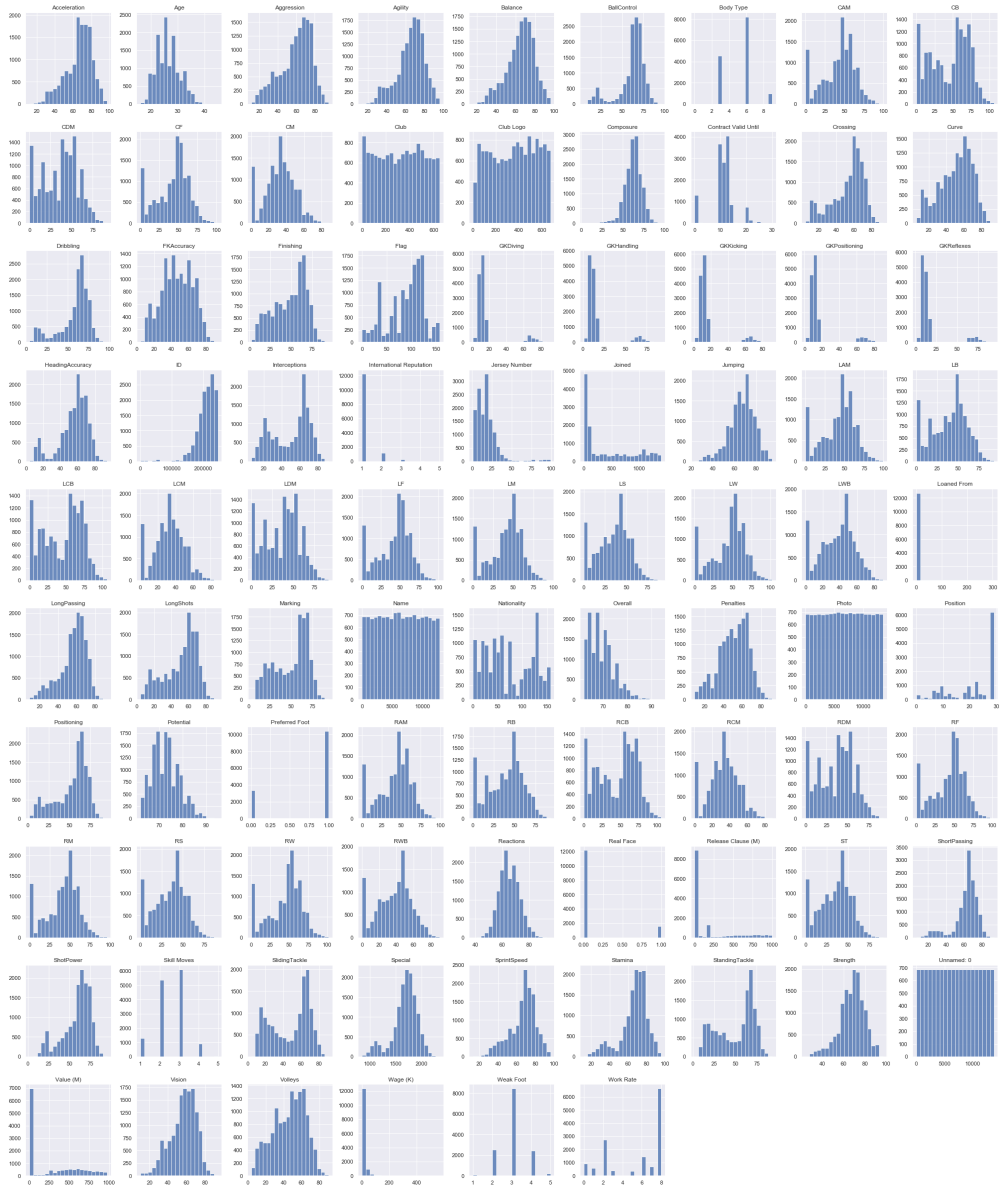

Figure 1: Histogram for All Variables.

Figure 1 shows the histogram for all variables in this dataset. For age, most players are 20 to 30 years old, they are young and with good energy. In addition, we can see most scores for exercise are left skewed, which means most players play well in exercise.

For further analysis, we drop several variables which may not have contributions to our model like photos, club logo, name, their contract valid date, the time they joined club, etc. We would build models on these new selected variables.

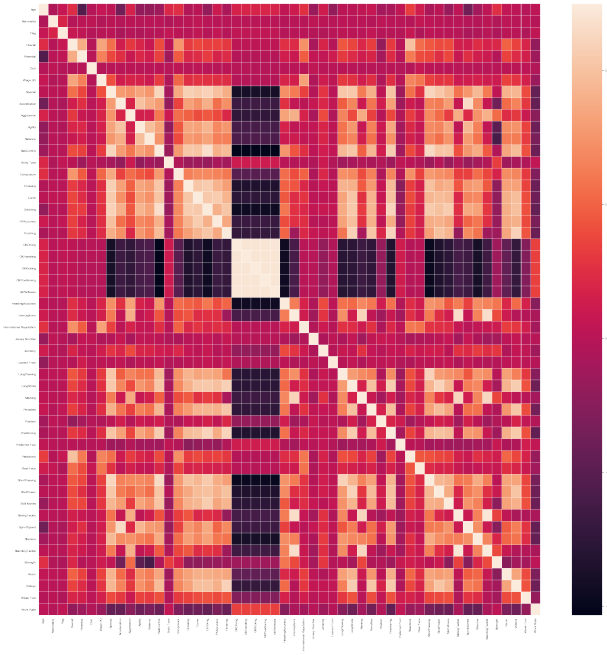

Figure 2: Heatmap for Variables

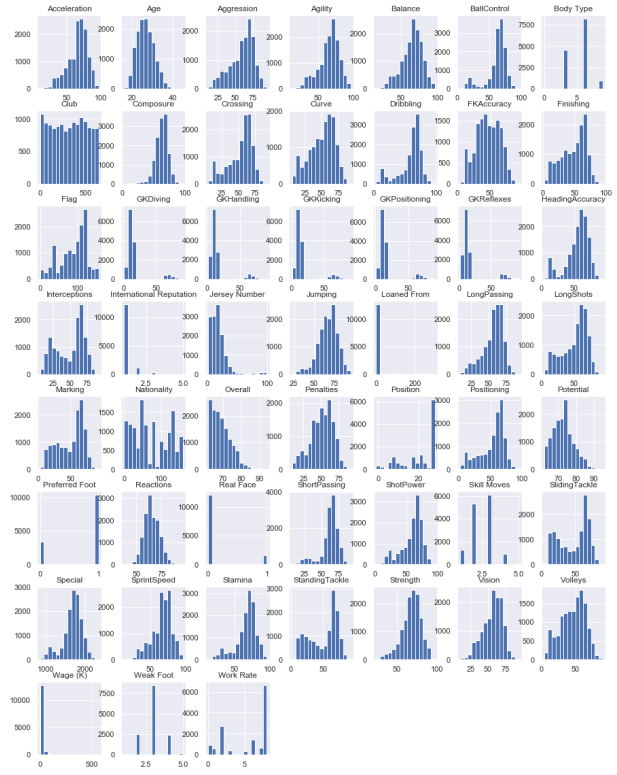

Figure 3: Histogram for Variables

Next, we did the heatmap for the variables that we have used. From Figure 2, we can easily see the correlation between variables in which black means variables have high negative relationship, and white means variables have high positive relationship. It makes sense since the dataset is about the scores the players get in each exercise and players would get a higher score if they are great. Then we did the histogram again, and this time the histogram shows more obvious that most attributes of exercise are left-skewed, which most players earn the good scores in each exercise.

3 Problem Description

FIFA is multi-billion company and if we can predict the value of player then we can let FIFA know what the player is worth and how other clubs can deal on buying or trading him/her with different clubs. The players are often traded or bought from one club to another and the clubs pays almost more than 100 million euros. Therefore, it is crucial that the clubs know what's the actual player worth.

We made the value of the player as a target and changed it to binary so that we can find out if the value of player will be higher than 100 million euros or less, which is around 70 percent. We classified value into two categories 1 and 0. 1 represents value greater to equal to 100 million euros, and 0 represents value less than 100 million euros.

4 Methodology

From this dataset we chose the features that we thought were important for our project. In order to transformed the data we used cat.code to change our data to categorical values to those that has string data. We then split the data in train and test set with .80 and .20 respectively. We then use XGBoost, LightGBM, CatBoost, Logistic Regression algorithm to get our AUC scores. To get the better results we did grid search and hyperopt for the best AUC we got. Finally we did SHAP summary plot and interaction plot to find out the most important features from our data and how it interacts to other features.

5 Evaluation

After running through our code multiple times we got similar good AUC score for XGBoost, lightgbm, and Catboost models. For XGBoost, AUC for train set is 0.9978, AUC for test set is 0.9961. For LightGBM, AUC for train set is 0.9973, AUC for test set is 0.9953. For CatBoost, AUC for train set is 0.9666, AUC for test set is 0.9639. For Logistic Regression, AUC for train set is 0.9620, AUC for test set is 0.9541. Then we did Grid Search for XGBoost, which AUC for train set is 0.9999, AUC for test set is 0.9972. We also did Hyperopt for XGBoost, which AUC for train set is 0.9746, AUC for test set is 0.9732.

Table 1: AUC Comparison

Model	Train AUC	Test AUC
XGBoost	0.9978	0.9961
LightGBM	0.9973	0.9953
CatBoost	0.9666	0.9639
Logistic	0.9620	0.9541
GridSearch for XGBoost	0.9999	0.9972
Hyperopt for XGBoost	0.9746	0.9732

Table 1 shows the AUC for train set and test set for all algorithms we tried.

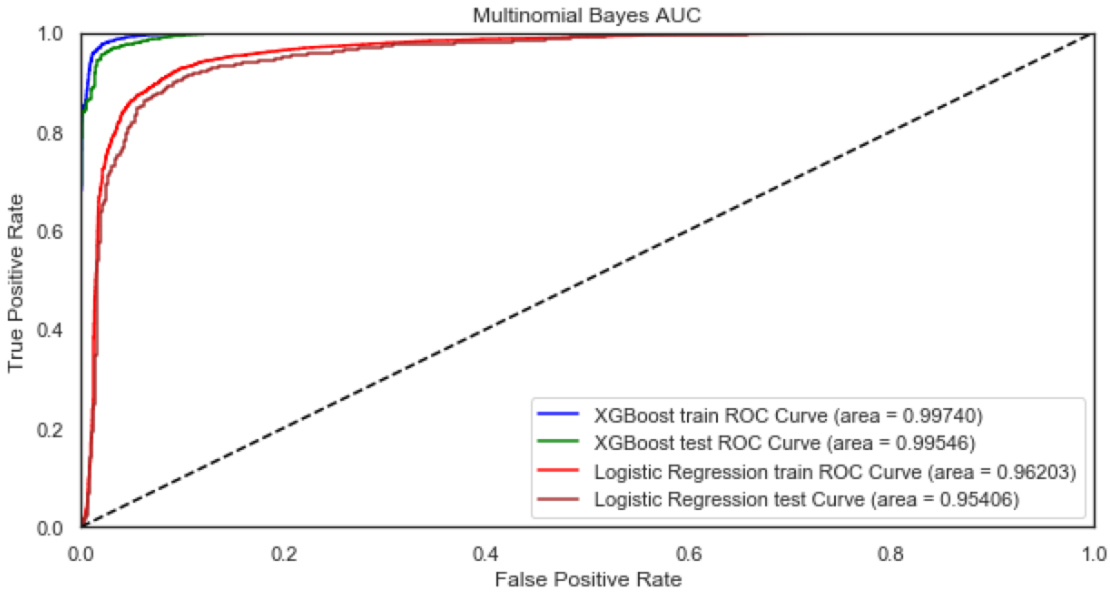

Figure 4: AUC Comparison

Since the difference between XGBoost and LightGBM is very small, and the difference between CatBoost and Logistic Regression is small too. Thus, Figure 4 only shows AUC for XGBoost train and test set, and AUC for logistic regression train and test set.

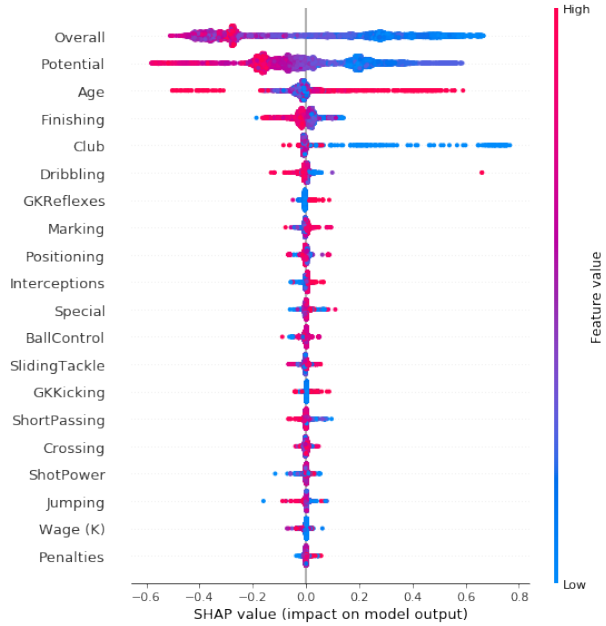

Figure 5: SHAP Summary Plot for Train Variables

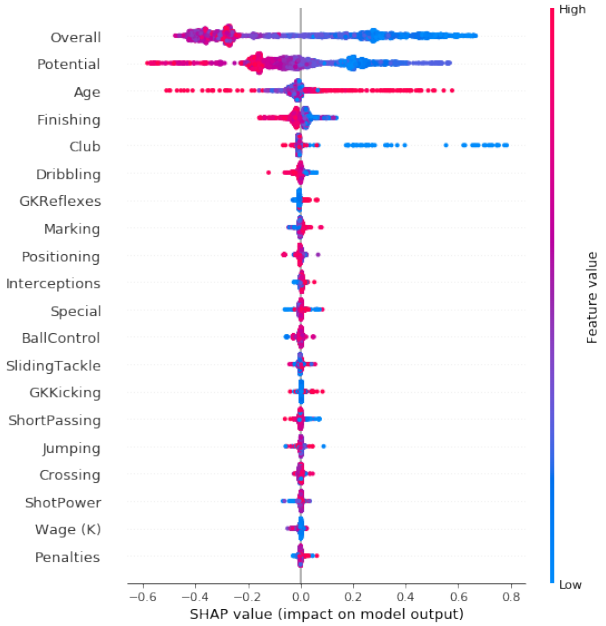

Figure 6: SHAP Summary Plot for Test Variables

Figure 5 shows the SHAP summary plot for training variables, which gave us that age is the most important feature. Figure 6 shows the SHAP summary plot for testing variables, which also gave us that age is the most important feature. From these two figures, we can easily see that SHAP summary plot gave us almost the same result for training and testing variables.

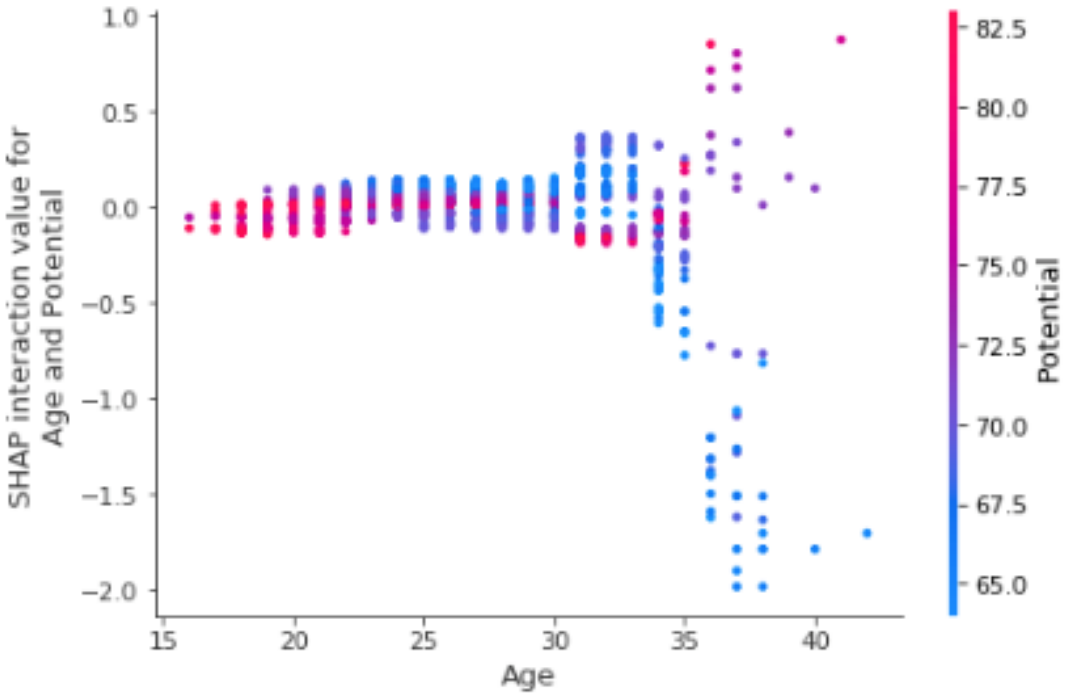

Figure 7: SHAP Interaction Plot

Figure 7 shows SHAP interaction plot, which we choose the most important two features from SHAP summary plot we tried previously. From Figure 7, age was interacted with potential, we found out that as the age goes higher the potential goes lower. Most of the players will have high potential in their early age.

6 Conclusion

We thought the reason all models appear very high AUC 0.99 is the variables are highly correlated with each other, which makes the model very good. We have selected all exercises to be our variables, and most players have very high grades for these exercises, it makes our model too good to predict the value players worth.

Since our results were unexpected, we worked on to see if we get different results. We re-selected the variables, to filter similar variables to make the model cogent, and run through our code again. The results were similar to our old models with AUC of 0.99 so we are very confident on our results. Since our model is good, we can use it to predict the value the player worth by given features.

From SHAP summary plot for our best model, we know the most important variables would influence the value the players worth, which includes overall scores, their potential, their age, their position, and specific exercise includes finishing, dribbling, GKReflexes, ball control, etc.

Football is a popular sport in the whole world and often times they are traded and bought. It is interesting to find if the player value is more than 100 million euros just by looking at their data so as to find the top player of the world. This will help not only the clubs but also the fans who are curious about the future of the player.

References

[1] FIFA.com, “About FIFA: Organisation,” FIFA.com. [Online]. Available: https://www.fifa.com/about-fifa/index.html. [Accessed: 10-Dec-2019].