PREDICTION OF FIFA PLAYER VALUE

Group member: Tshering Sherpa, Yanru Li, Yuyan Li

Data Description:

FIFA stands for the Fédération Internationale de Football Association which is a non-profit organization which describes itself as an international governing body of association football. In the USA it is called soccer. 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/.

Problem:

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.

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.

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 just by looking at their data. This will help not only the clubs but also the fans who are curious about the future of the player.

Techniques:

From this data set we chose the features that we thought were important for our project. We split the data in train and test set and use XGboost, lightgbm, Catboost, and logistic regression models to get our auc score. To get the better results we did grid search and hyperopt. We then did shap and interaction plot to find out the most important features from our data.

Findings and Conclusions :

After running through our code multiple times we got the same results with 0.99 auc score for XGBoost, lightgbm, and Catboost models. For the logistic regression model, we got the auc score for training set is 0.9 4 and for test set is 0.93. Figure 1 shows the result for XGBoost model and logistic regression model which auc for training and testing data.

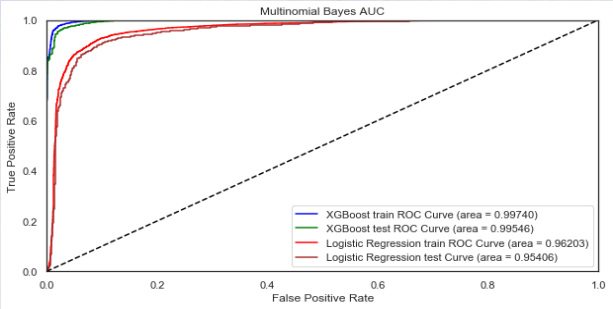


Figure 1

Figure 1 shows the result of Area under the ROC (auc) curve for all four models we tried. We got XGBoost model with highest train and test score.

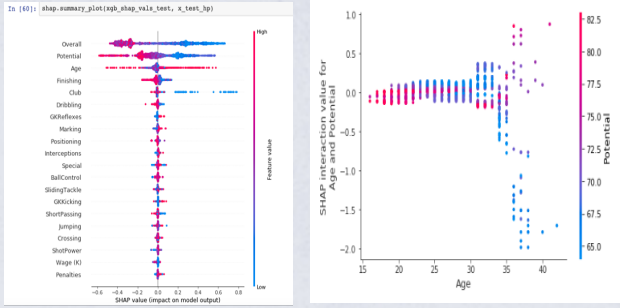


Figure 2

Figure 2 shows the shap summary plot, which gave us that age is the most important feature from our data and when interacted with potential we found out that the higher the age then there is more chance to have low potential.

Interesting Finds:

We did the heatmap for the variables we have used to classify. From Figure 3, we can easily see the correlation between variables in which black means variables have highly negative relationship, and white means variables have highly 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. We also did the histogram (Figure 4), it shows most attributes of exercise are left-skewed. It also means most players earn the good scores in each exercise.

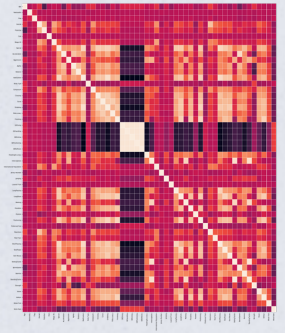


Figure 3

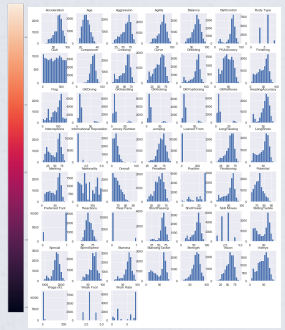


Figure 4

Step Forward:

We thought the reason all models appear very good with 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.

Confidence:

Since our results were unexpected, we worked on to get well fit models. We reselected the variables, to filter similar variables to make the model cogent, and run through our code again to see if we will get different or better results but it was similar to our old models, so we are very confident on our results.

University of Pittsburgh

FIFA For the Game. For the World.

PREDICTION OF FIFA PLAYER VALUE

Data Mining Project by:

Tshering Sherpa

University of Pittsburgh

FIFA
For the Game. For the World.

Data Description:

1. FIFA stands for the Fédération Internationale de Football Association.
2. It is a non-profit organization which describes itself as an international governing body of association football.
3. In the USA, football is called soccer.
4. We got our data from kaggle datasets.
5. It is named “FIFA 19 complete player”
6. Our dataset is the latest edition of FIFA, and the data is scraped from https://sofifa.com/.

Problem:

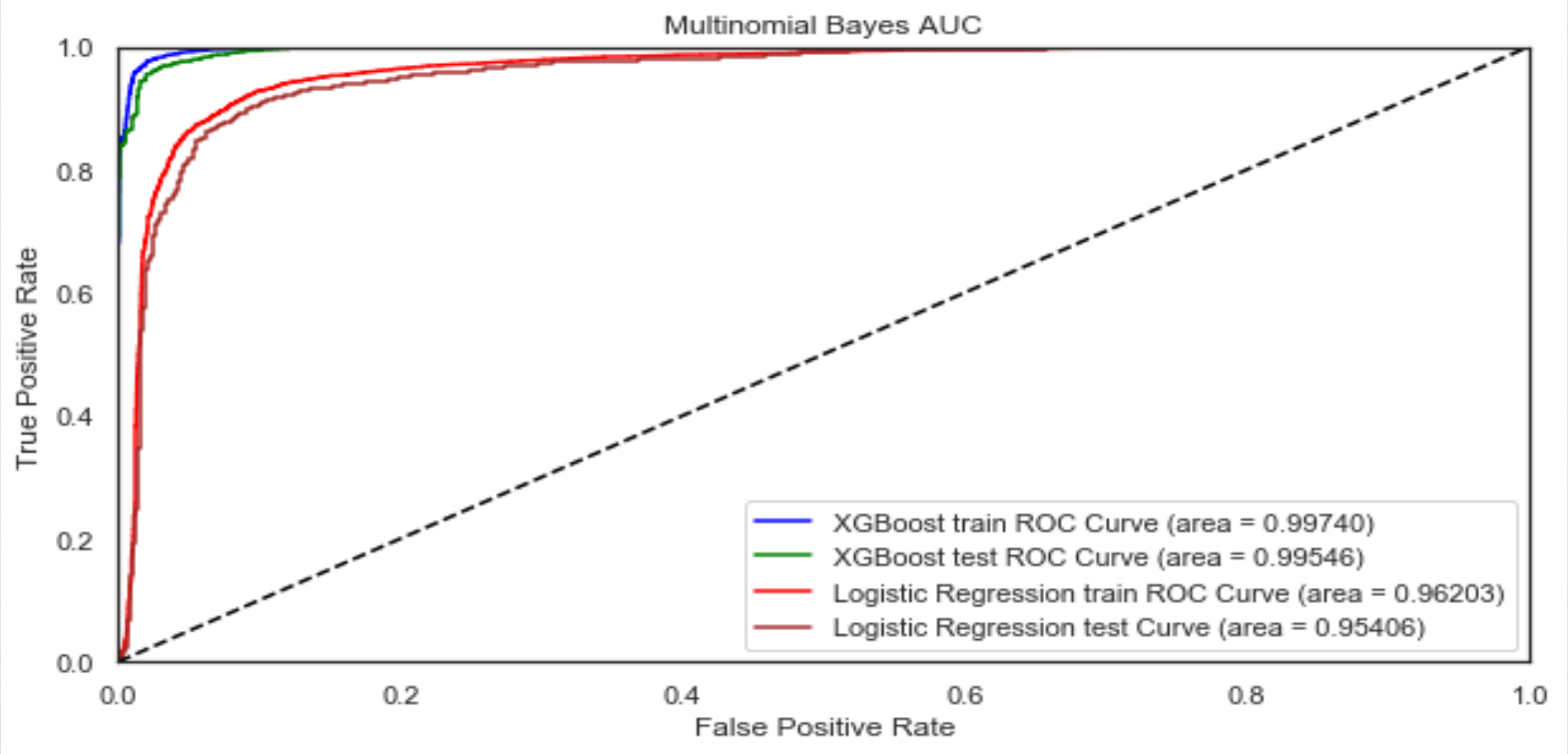
1. To find the top valued fifa players.
2. In our research, we decided to take player whose values are more than 100 million euros as top valued fifa players.
3. To predict the value of player.
4. To make the value of the player as a target.
5. To let the club owner and fans know the worth of the fifa player.

Techniques:

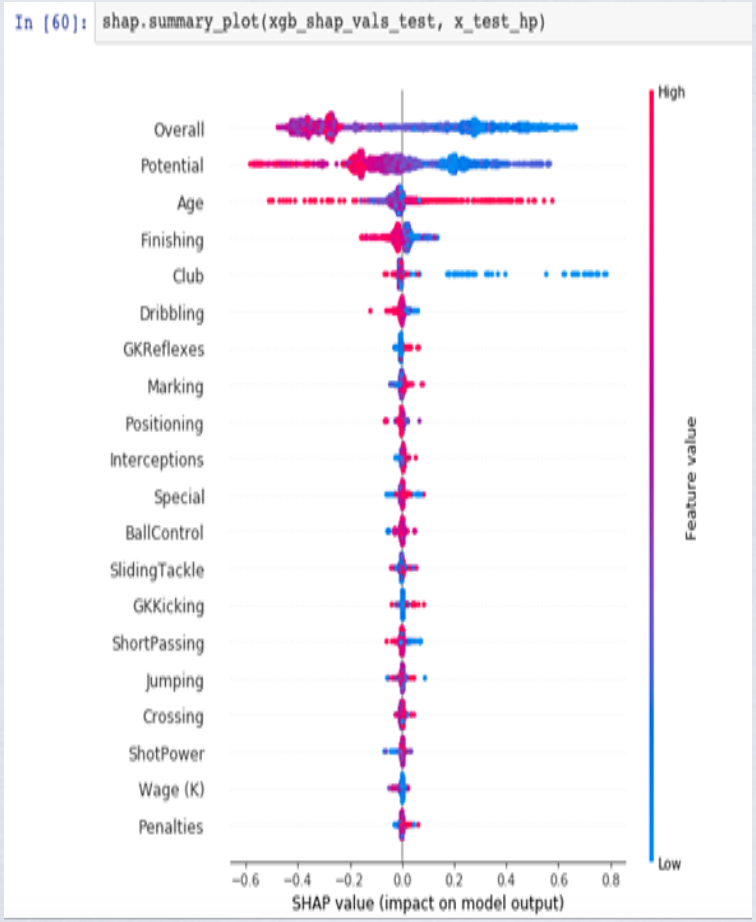
1. Choose the best feature/variables from the data set that are important which will give us the best model.
2. Binary classification of the value of the player.
3. Split the data in train and test set with 80% and 20% respectively.
4. Use machine learning algorithm like XGBoost, LightGBM, CatBoost packages to get our auc score.
5. Use SHAP package and SHAP interaction plot to find out the most important features from the data.

Findings and Conclusions:

1. We run our code multiple times with different variable and with different packages.
2. The best results we got was 0.99 auc score.
3. Even though we used XGBoost, LightGBM, and CatBoost models for the prediction and got almost similar result.
4. For logistic regression model, we got the auc score for training set 0.96 and for test set 0.95.
5. Out of all of the models we got, XGBoost performed the best with 0.99 auc score of train and test set.

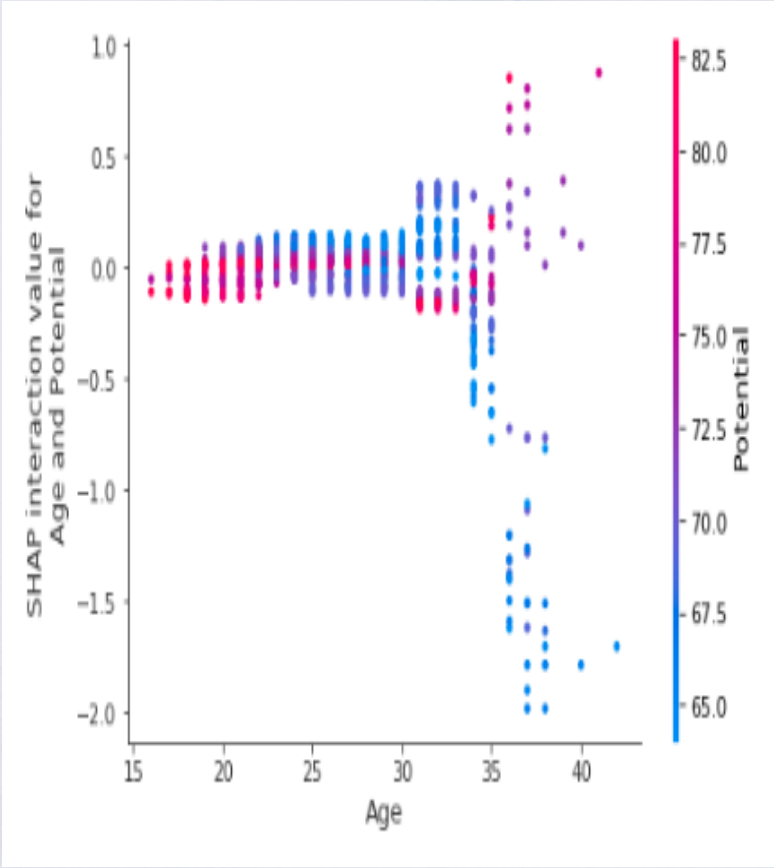


The above figure shows the result of Area under the ROC (auc) curve for all four models we tried. We got XGBoost model with highest train and test score.



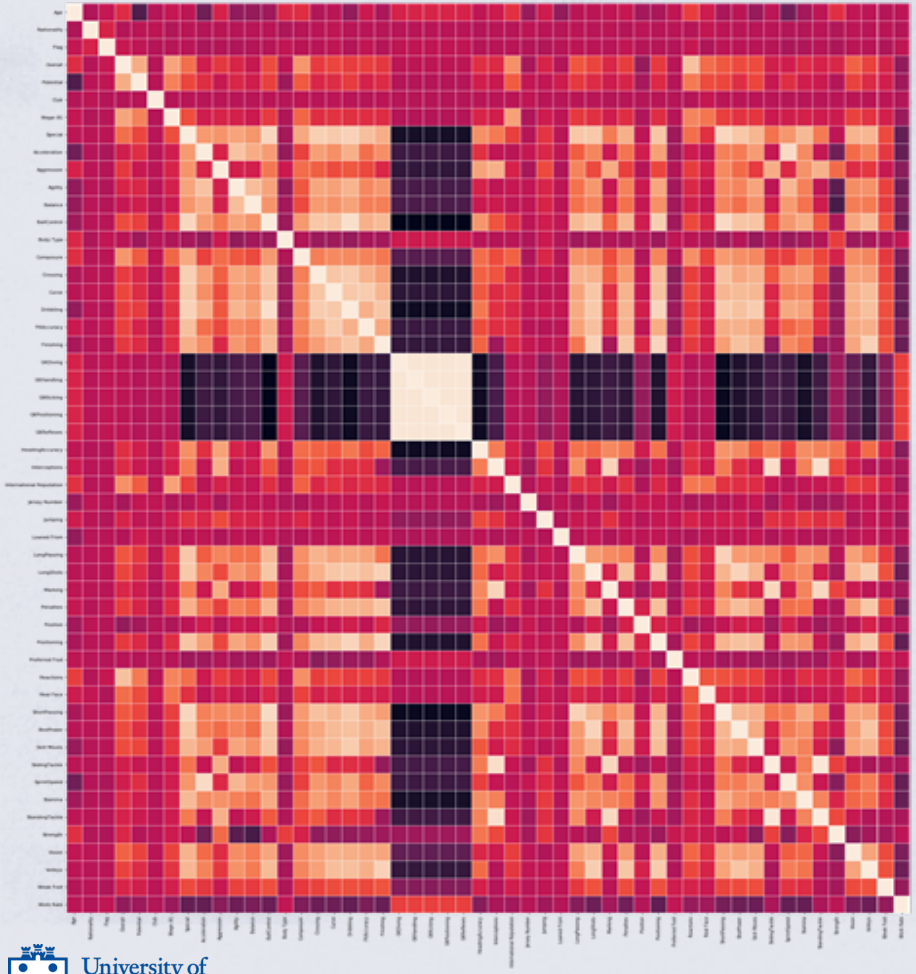
1. This figure shows the SHAP summary plot of our data.
2. Here the red color denotes higher feature value and blue denotes lower value.
3. Since age has the most red data points in the figure, we can say that age is the most important feature from our data.

University of Pittsburgh



1. The figure is the SHAP interaction plot of age and potential.
2. We can see that sign of high potential player are shown in early age from 15 to 25.
3. We found out that the higher the age, there is more chance to have low potential.

University of Pittsburgh



1. From this figure, we can see the correlation between variables.
2. Darker color means variables have highly negative relationship, and lighter color means variables highly positive relationship.
3. It makes sense since the dataset is about the scores the players get in each exercise and the players would get higher score if they are great.
4. Heatmap is one of the feature that tell us the correlation of each variable in the dataset just by looking at the figure.

1. This figure is the histogram of the variable of the data.
2. It shows most attributes of exercise are left-skewed.
3. It means most players earn the good scores in each exercise.
4. It also means that most of the players from our dataset are highly skilled.
5. Histograms plots are very useful to see where the data are skewed which tell us where the data are mostly dense.

University of Pittsburgh

Step Forward:

1. We thought the reason all models appear very good with auc 0.99 is because the variable are highly correlated to each other.
2. We have selected all exercises to be our variables, and most players have very high scores for the exercises, it makes our model great to predict the value players worth.
3. We also thought that our dataset were large enough to give us the insights of the players.

Confidence:

1. Since our results were unexpected, we worked on to get other models.
2. We reselected the variables, to filter similar variables to make the model cogent, and run through our code again.
3. We got our results to be similar to our old models, so we are very confident on our results.

Summary:

1. We have a very good model that can predict the value of the fifa player.
2. Our XGBoost model performed the best with 0.99 auc score.
3. We found out age is the most important feature from our data that will give us the best models.
4. Using our model owner of clubs can buy or trade their players or use it to negotiate with the players according to their needs.
5. Using our model even fans can tell who is worth more than 100 million euros.