CS584 Final Project Report

Ting Jiang A20386491,Chen Gong A20379606, Yizhi Hong A20386348

1. Introduction

This project will analyze the famous video game: FIFA 18. There are over 20k soccer players to be analyzed. The project targets to analyze the relationship between the player's overall and the player's attribute, the relationship between the player's position and the player's attribute. We are quite interesting in the rating system of the player. And how the abilities(attributes) needed in certain position. This will help the game player to determine the selection of the player, and even helps the real soccer to understand the attributes that they need to be improved. FIFA 18 provide a dataset (fifa-18-demo-player-dataset) for this research.

Main Problem:

1. The overall statistics of FIFA 18.
2. How the attributes influencing the rating of the player.
3. Predict the type of the player. (Classification)
4. Predict the position of the player. (Random Forest)
5. Predict the overall rating of the player. (Regression: Linear / Polynomial)

2. Related work

[1] Exploring FIFA 2018 dataset Hitesh palamada
The author shows a lots of statistics from FIFA 18. Such as top player, the attribute in the dataset.
He implemented in python and take the dataset as a data frame to do the operation. He look up the whole dataset and get something we might interesting in, e.g. most valued player, age density, best club and so on.

[2] FIFA 18 - Predict player's positions DLao
The author shows the relationship between the position and the attribute. He use classifier to group them together. Also tune the attributes to train the model. He gives us the inspiration of our work.

3. Approach

Generally, the methods of regression and classification would be used in this research, also we will apply cross validation and regularization in order to get the better result for our analyzing. The data will are using is the real dataset in the FIFA 18 video game.
We are following the rule that we learned from class:

a. Pre-process the data

Since the dataset is so huge, It includes a lot of informations that might redundant in our work. (e.g. club, nationality, age, name, foot). At first, we transfer the dataset to the data frame then taking the columns that we need. Also, the goalkeeper has totally different rating system in the game and it might not in our consider, so we should remove the goalkeeper in our data frame. At second, we take a look the data, and try to deal with the null/signed value. Some attribute might have 80 - 5, we just do the calculation as the system do, Also we normalize our data in case the outlier occur . At last, we set the most preferred position for each player (handle multiple positions) then categorizing the position for prediction. Since we have 3 different type of trainings and predictions. We will perform the pre-process in the data analyze.

b. Data analyze

For common sense about soccer player, we take some attributes might relevant to the overall rating and the positional decision. Then we separated 3 parts of the attributes: Attack Attributes, Mixed Attributes, Defend Attributes which will consequently change the main types of the player: ST(Attacker) , CM(Mixed player), CB(Defender). We do the hypothesis below then randomly pick the players to plot the attribute and the rating with 3 lines of position: ST, CM, CB. We notice that it did prove our hypothesis a little bit when we eyeball the plot. So now we try to perform the model to do our prediction.

c. Fit model and model regularization

1. **Logistic Regression - predict the attacker or the defender**
    - Categorize the position in 1 or 0, separated the 14 positions in attacker group and defender group.
    - Set the ST/RW/LW/RM/CM/LM/CAM/CF as an Attacker group as 1,Set the CDM/CB/LB/RB/RWB/LWB as an Defender group as 0.
    - Perform 5 cross validation for Logistic regression.
    - Measuring by accuracy, since either 1 or 0.

- Tune the model by lasso, since we need to get rid of some attributes which will not contribute the defender or attacker position. E.g. Shot power might not important for the defender. But important for Attacker.
- Tune the alpha to get the coef for each attribute, sort the coef by descending order. Then try through the features(attributes) to yield the best accuracy.

2. **RandomForest - predict all the positions**
    - Categorize the position in 0 to 13, factorized the 14 positions.
    - Perform 5 cross validation for Random Forest.
    - Tune the model by ridge, since all the attributes which will contribute to different position.
    - Measuring by accuracy, since we know the true position for the player.
    - Tune the alpha to get the coef for each attribute, sort the coef by descending order. Then try through the features(attributes) to yield the best accuracy.
    - Show the features are significant.

3. **Linear Regression - predict the overall rating**
    - Measuring by MSE in order to get accuracy
    - Perform 5 fold cross validation for Linear Regression
    - Tune the model by ridge, we need to consider all the attributes to evaluate a player.
    - Measuring by MSE to determine the performance. Since we get the numerical result and numerical rating.
    - Tune the the Linear Regression in Linear model and polynomial model.
    - Show the MSE and best performance model.

4. Result

As the result. We find out lots of interesting information in our work. As each approach, we set up the baseline. We find the top 5 important features that to determine your position are 'Finishing', 'Crossing', 'Marking', 'Ball control', 'Sliding tackle' in this game.
In predicting the attacker or defender, we set 0.5 as baseline by common sense. We perform 0.833513325558 accuracy through lasso which is outperform our baseline.
In predicting all the position, we set 0.71 as baseline by 1/ 14 positions, We perform 0.395765861155 accuracy through ridge which is outperform our baseline.
In predicting the overall rating, we set average as our baseline: total rating over attribute.
We find that very low MSE in Polynomial model regularized by ridge which is 1.8e-7. We random test the player, we get a similar result rating. I think we found the formula in FIFA system which is impressive.

5. Conclusion

In conclusion, We find that really interesting since we are soccer fans. We use the techniques we learned from class, classification and regression, solving an interesting issue in real world. Most of our work performs well with decent accuracy.

In addition, datasets from real world are always huge and calculation intensive. Sometimes, the calculation work will take long time within one calculation node. So, the data analysis in real world will requires calculation clusters with distributed calculation techniques such as mapreduce.

6. Reference

[1] FIFA18 Dataset. https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset/data

[2] Taking a look at the FIFA 18 Player Dataset.https://www.kaggle.com/donyoe/taking-a-look-at-the-fifa-18-player-dataset

[3] FIFA 18 - Predict player's positions DLao

[4] Exploring FIFA 2018 dataset Hitesh palamada