Your Wonderful Project Topic

Author #1, Author #2

Abstract

For this project, I am positioning myself as a scouting agency that uses analytics to, among other things, enhance the discovery of talents and help soccer clubs better understand the dynamics (features) that come into play when determining the value, overall and future potential of a player.

Although the *overall* rating is the most important value either in the real football market or FIFA games, the official rating criteria is still not published. Our model is to predict the *overall* rating of players given their several capability values, which use the data classified by players' position and the whole position data. Besides, our project only uses some of these attributes concerning only player capability according to official guideline.

To demonstrate the benefits of our model, we built various charts and evaluated it. In our experiments, neural networks could ... compared to liner regression.

Introduction

The Fédération Internationale de Football Association (FIFA) will publish the overall rating and other metrics for soccer players every year. Although the metrics can affect the overall rating of a player, the exact rating criteria is still not published. Our project is to predict the overall rating of players given their capability values. Several data analyses projects have been designed for this dataset. However, these projects only focus on evaluating the metrics of players but failed to predict the overall rating of players. Some projects tried to predict the overall rating of players, but these projects still lack the comparison among different models to choose the best one or give any persuasive insights or conclusions. Our project tries different models for rating prediction and determine their performances to find the best model and give insightful conclusions and opinions.

We implement regression methods in machine learning to this problem. We have trained our model based on a set of data randomly selected from our dataset, and then test our model on a stand-out test set also selected from our dataset. In this way, we create and validate a model predicting the overall rating of a player.

Our project dataset is FIFA 2019 players attributes dataset, which is collected from Kaggle. This dataset includes 89 attributes of a soccer player, and our project only uses some of these attributes concerning only player capability. Besides, considering that goalkeepers have a totally different set of capability metrics compared with other players.

We implement linear regression and neural networks in our project. For each model, we use cross validation to find the best hyperparameter for each model. We compare their performance based on their $r2score$ and testing $RMSE$. Finally choose the most accurate model, neural networks. In order to finish this model fitting and predicting process, we plan to randomly select data from our dataset, and then split the selected data into training and testing sets to implement cross validation.

Dataset

0.1 Dateset overall

Our project dataset is FIFA 2019 players attributes dataset, which is collected from Kaggle. This dataset includes 89 attributes of a soccer player, and our project only uses some of these attributes concerning only

Category	*Attributes*
Basic Information	*ID, Name, Age, Photo, Nationality, Real Face, Height, Weight, Body Type, Special, Flag, Position, Club, Club Logo, Work Rate, Jersey Number, Weak Foot, Preferred Foot, Skill Moves, International Reputation*
Ratings	*Overall, Potential, LS, ST, RS, LW, LF, CF, RF, RW, LAM, CAM, RAM, LM, LCM, CM, RCM, RM, LWB, LDM, CDM, RDM, RWB, LB, LCB, CB, RCB, RB*
Market Value Related	*Value, Wage, Joined, Loaned From, Contract Valid Until, Release Clause*
Abilities	*Crossing, Finishing, HeadingAccuracy, ShortPassing, Volleys, Dribbling, Curve,FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties,Composure, Marking, StandingTackle, SlidingTackle*
Goalkeeper Abilities	*GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes*

Table 1: Columns (attributes) in the original dataset.

Category	*Attributes*
Basic Information	*Age, Position*
Ratings	*Overall* (as target value)
Abilities	*Crossing, Finishing, HeadingAccuracy, ShortPassing, Volleys, Dribbling, Curve,FKAccuracy, LongPassing, BallControl, Acceleration, SprintSpeed, Agility, Reactions, Balance, ShotPower, Jumping, Stamina, Strength, LongShots, Aggression, Interceptions, Positioning, Vision, Penalties,Composure, Marking, StandingTackle, SlidingTackle*
Goalkeeper Abilities	*GKDiving, GKHandling, GKKicking, GKPositioning, GKReflexes*

Table 2: Columns (attributes) in the dataset for model prediction.

player capability. Besides, considering that goalkeepers have a totally different set of capability metrics compared with other players, we will only consider data records of nongoalkeepers in our model.

0.2 Data processing

Now we beginning the data processing. There are 18206 players for which 89 features each are provided. Firstly, we should drop the duplicate entries. Then we check the columns (attributes). These columns can be grouped into 6 categories as follows:

Our model only focuses on predicting the overall rating of a player. Thus, we only need to keep those attributes related to the ability of players and drop the other irrelavent attributes. Besides, our model does not consider ratings on different positions for the same player. Thus, we only keep attributes as follows:

After deleting all the irrelavent attributes, we still have to delete data records with missing values. These records are resulted from , and they could introduce errors into our prediction. The null values in each column are counted and sorted as follows:

After deleting these missing values, we have 18147 rows left. After that, we still should inspect the data types of each column, and we should ensure the input attributes of our model are all of type *int* to guarantee that our model works. The only attribute violating this in our model prediction dataset is *Position*. This attribute only indicates the position of a soccer player. Thus, we can divide all records into

Index	*column name*	*Total missing*	*Percent missing*
0	*Position*	*60*	*0.003295*
1	*GKReflexes*	*48*	*0.002636*
2	*Curve*	*48*	*0.002636*
3	*Agility*	*48*	*0.002636*
4	*SprintSpeed*	*48*	*0.002636*
5	*Acceleration*	*48*	*0.002636*
6	*BallControl*	*48*	*0.002636*
...	...	...	...
35	*Stamina*	*48*	*0.002636*
36	*Overall*	*0*	*0.000000*
37	*Age*	*0*	*0.000000*

Table 3: Missing value percent.

different categories according to their positions.

We mainly divide players into 4 positions: **Forward**, **Midfielder**, **Back**, and **Goalkeeper**, and 4 CSV files are built for each of these categories to store them separately. Specially, we also built a CSV file to store all data together after our processing above.

Solution

The solution section covers all of your model design, algorithms, formulas, findings etc. It explains in detail each contribution, if possible with figures/schematics.

Results and Discussion

The results section details your metrics and experiments for the assessment of your solution. It allows you to compare your idea with other approaches you've tested.