DSCI 562 Lab 1

Rohit Rawat

Contents

Repo link: https://github.ubc.ca/mds-2021-22/DSCI_542_lab2_rrrohit1

Audience Persona

Joe Devine[1] is an Executive Producer at The Athletic, a subscription-based sports website that covers Football and other sports. He is also the Creative Director at Tifo Football, a Youtube channel dedicated to Football analysis. `Sensible Transfers`[2] is a football club-specific video series on the channel on which Joe makes analytics-based reporting on football transfers, staying away from rumors. This analysis is supposed to act as a brief for Joe to consider for making the Sensible Transfers edition for the club Manchester United. He is well-versed in football terminology and knowledge of advanced machine learning topics is not assumed. The focus of the report is on interpretability and giving suitable suggestions while marking the key issues with the club. He has more than 10 years of experience in analytics-based football journalism.

Abstract

`FIFA 21` is a football simulation video game developed by EA. It is popular among the younger generations and even among actual football players[3]. Manchester United is struggling in the Premier League, the top-tier football league in England. A series of bad player investments have to lead to their downfall. In this study, a dataset of 806 football players listed in `FIFA 21` and currently active in the Premier League in 2021 was scraped from the website `sofifa.com`. This publicly available data on football players were used to discern the differences in player attributes between Manchester United and the other clubs in the league. The major issue in the team was a low-growth squad with an aging midfield section. It was inferred that age, international popularity, and social media following were the most important features associated with player growth. To fill the gap in the club's squad, a Ridge regression was the best model to predict and interpret players' growth. The model had an R-2 score of 0.84 on the test set from where Rico Richards was chosen as an appropriate midfield choice for transfer due to his low weekly salary and high growth potential.

Introduction

`FIFA 21` is a popular association football simulation-based video game that has a useful database of actual football players. In the past, another football game `Football Manager` has been extensively used for scouting players around the world. Popular football teams have used data analysis to scout players and make strategic decisions[4].

Manchester United were the Premier League champions in the 2013-14 season but have seen a dip in their performances for the last eight years. Bad decision-making and poor recruitment choices have led to this downfall. Three key issues which hamper the growth of the club currently are as follows:

1. The highest player wage bill in the league of £ 226 million
2. A aging squad with an average of 27 years, 10^{th} in the league, especially the midfield players

To turn their fortune around and ensure the Top 4 finish is essential to the club's functioning as it ensures additional revenue by booking a place in the UEFA Champions League. Currently, the team is 5^{th} in the league and has been witnessing poor league performances. To overcome this, the team needs to recruit young new players who are not on an expensive weekly wage bill.

In this study, the dataset of Premier League football players is taken, with a focus on recruiting from the bottom teams in the league whose players are likely to transfer teams to avoid relegation. Their current playing attributes along with their potential growth are assessed to decide which players have high growth potential. The Linear regression model with its extensions of Ridge and Polynomial regression is used to predict the growth of bottom league team players and present some possible choices based on the model.

Methods

Study Design and study population

The project was designed for Manchester United to scout premier league football players for their own team. The total number of 18,541 players available in major league teams around the world coming from 170 countries around the world.

For the analysis, the English Premier League football players are selected since it would be easier to assess their performance and scout them.

Data collection

The original data is updated by the game maker `FIFA` on a yearly basis during the game release in October every year. They have not made their API available for public use. The data can be extracted by buying the PC game and scraping player profiles in the game files.

This job has been done by the website `sofifa.com`[5] which updates the data on a yearly basis for public use and publishes it on their website. I have used web scraping to extract the dataset in a tabular form[6].

For each player, the player's demographic details, current team, nationality, position on the field, and the different playing attributes measured on a scale of 0 to 100 are populated.

Data preprocessing

The original scraped had details of 18,541 players available all over the world. The focus of the project was Manchester United and it plays in the Premier League in England. So, the dataset has been filtered to get the details of only Premier League teams. The selected dataset has 806 players and 91 attributes.

The details of dropped 43 features are as follows:

- Fields related to the image file name of the player's picture, the club logo, and the national flag
- The possible player rating if he was played out of position. This can be avoided as the focus is on the growth of the player in his original position
- The jersey number and the unique identifier of the player

The `Position` feature had 28 categories. I reduced the categories to 4 major segments:

1. GK: Goalkeeper
2. DF: Defender
3. MF: Midfielder
4. ST: Striker

There are two player quality columns in the dataset: `Overall` and `Potential`. In general, the growth of the individual is important for scouting young talent in the league. So, a new column `Growth` is created using the difference between `Overall` and `Potential` of the player.

The `Club` column has 20 different football teams. Since Manchester United is a top-tier club, it is fair to evaluate it with the top 3 teams in the league table. So, the column has been divided into four categories:

1. ManUtd: Manchester United
2. Top3: Manchester City, Liverpool, Chelsea
3. Bottom3: Bournemouth, West Bromwich Albion, Watford
4. Others: The remaining 13 teams

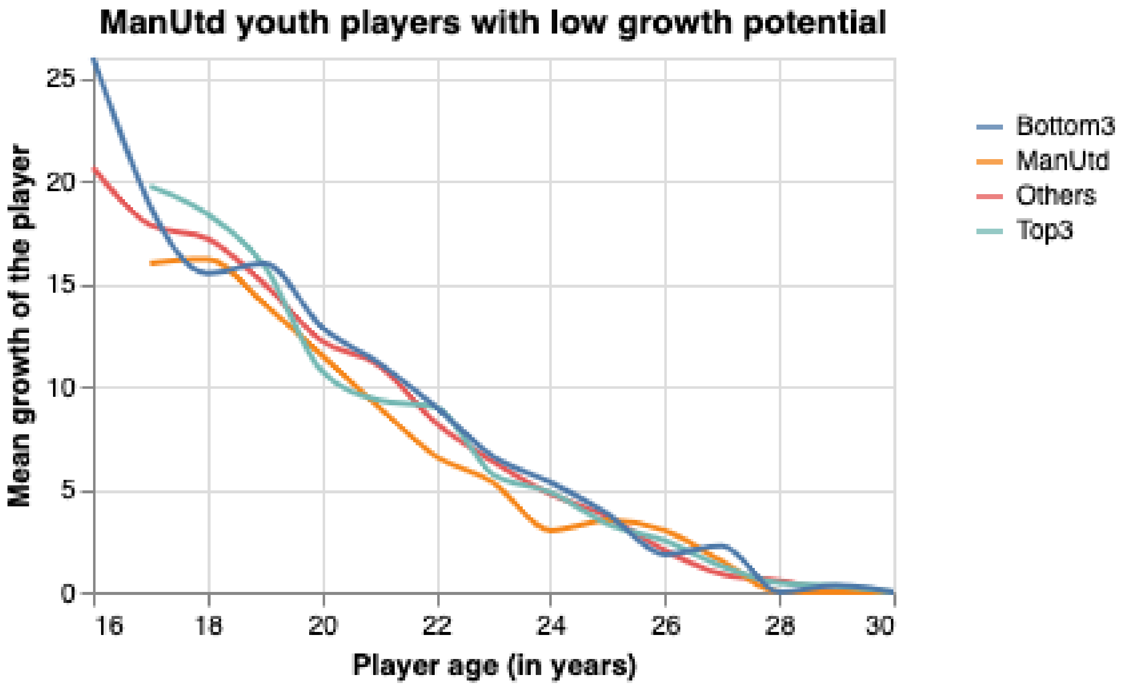
This categorization would be used in the visualization in the report.

Three columns `Weak Foot`, `International Reputation` and `Skill Moves` have ordinal data with a range of 1 to 5, suffixed with a special star character. So, the columns are truncated to include only the numbers.

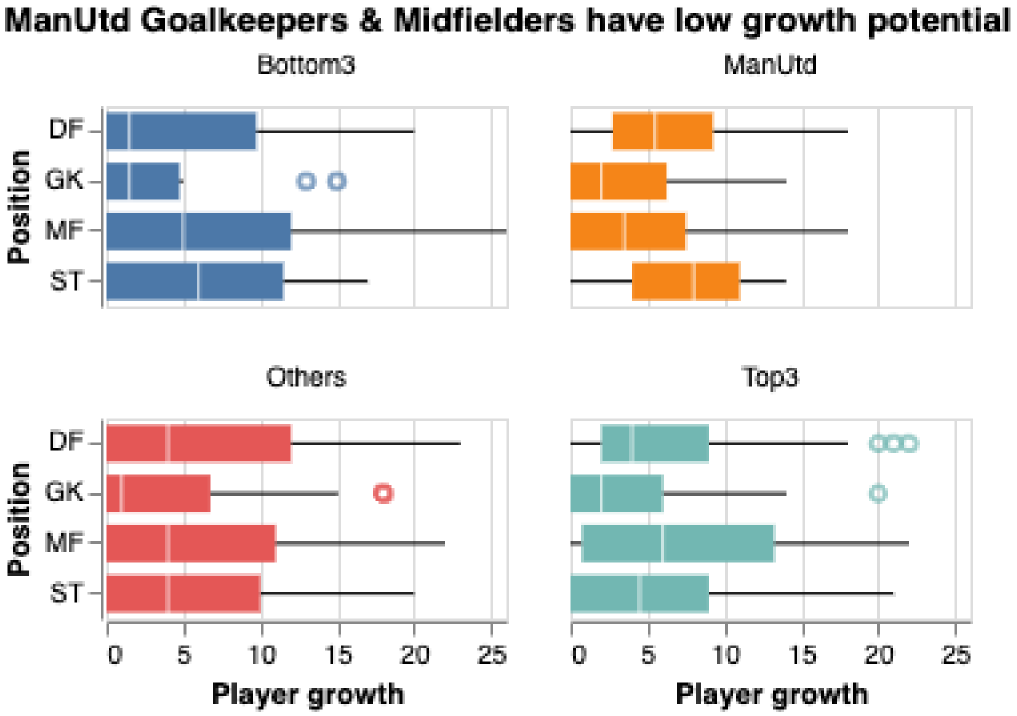
The dataset has missing values in some player attributes but those can be replaced with zero. The reason is that the missing values are concerned with out-of-position attributes of the player such as the goalkeeping ability of the striker is missing which makes sense as it is not his position. So, these NaN values are imputed with 0. The test data split comprises the Bottom3 teams from where the team is looking to scout players.

Assessing key team issues

There have been a series of bad signings made by the team and the quality of the youth at the club is also questionable. From the visualization below, we can observe that the growth potential of the players under the age of 25 at the club is comparatively worse than the other teams.



The feature `Age` and the response `Growth` are negatively related with a correlation of -0.86. Secondly, the team has been leaking a lot of easy goals due to poor defense and losing the ball in the midfield. These two could be the areas that need fresher legs. The graph below illustrates that the growth potential of the midfielders and the goalkeeper is the most questionable.



Statistical Analysis

All statistical analysis was conducted by using Python version 3.9.7. The goal of the analysis is two-fold. Firstly, the study wanted to find out the factors which impact the growth of players. Secondly, it predicted the possible growth of Premier League players of other teams. Due to this duality, where the prediction and the interpretation are important, we are using Linear Regression and its extensions such as Ridge regression and using polynomial features for modeling the problem. The coefficients of these regression equations were used to interpret our results. Our model performance is assessed by the R2 score. R2(R-squared) is a goodness of fit measure for regression models, which explains the variation in the response variable around its mean.

The analysis is started by highlighting the key issues affecting the club and its comparison with the other teams. Then, a column transformer is made to perform one-hot encoding on the categorical features and scale the numeric ones. A pipeline was created to pass the preprocessed data to the models which were fitted. A Dummy regressor was used as the baseline and the other models mentioned above were fitted one by one on the training data.

In this study, Tree-based methods such as RandomForest, CatBoost and XGBoost are avoided due to interpretability issues, making them difficult to communicate to a larger audience.

Results

Model selection

Keeping these two things in mind, the dataset on modeled on the baseline dummy regressor which gave bad training and validation scores of 0 and -0.08 respectively. After scaling the numerical features and one-hot encoding the categorical features, we passed the processed data into Ridge Regression with default parameters. The `R-squared` score for the model was 0.88 for the training set and 0.82 for the validation set. The model was able to explain 82% variation in the validation set by the model predictions.

The ordinary Linear regression gave comparable results. Polynomial features overfitted the training data set and gave a poor validation score of 0.053. So, Ridge regression is the best choice here since it would generalize better on the test dataset due to regularization. The complete list of train and validation scores is as follows:

	Dummy regressor	Linear regression	Ridge regression	Polynomial regression
fit time	0.01	0.01	0.01	0.04
score time	0.00	0.00	0.00	0.00
validation score	-0.08	0.82	0.82	0.53
train score	0.00	0.88	0.88	1.00

Interpretation

Using the model pipeline, the `Age` of the player had the regressor coefficient with the highest magnitude with a negative sign. Given the level of significance $\alpha = 0.05$, the coefficient had a p-value less than 0.05, making it statistically significant for the study. This aligns with our earlier observation that `Age` and `Growth` are negatively correlated. Other features with high magnitude are as follows:

1. International Reputation
2. Social profile of the player: Following, Likes
3. Wage of the player
4. Club of the player

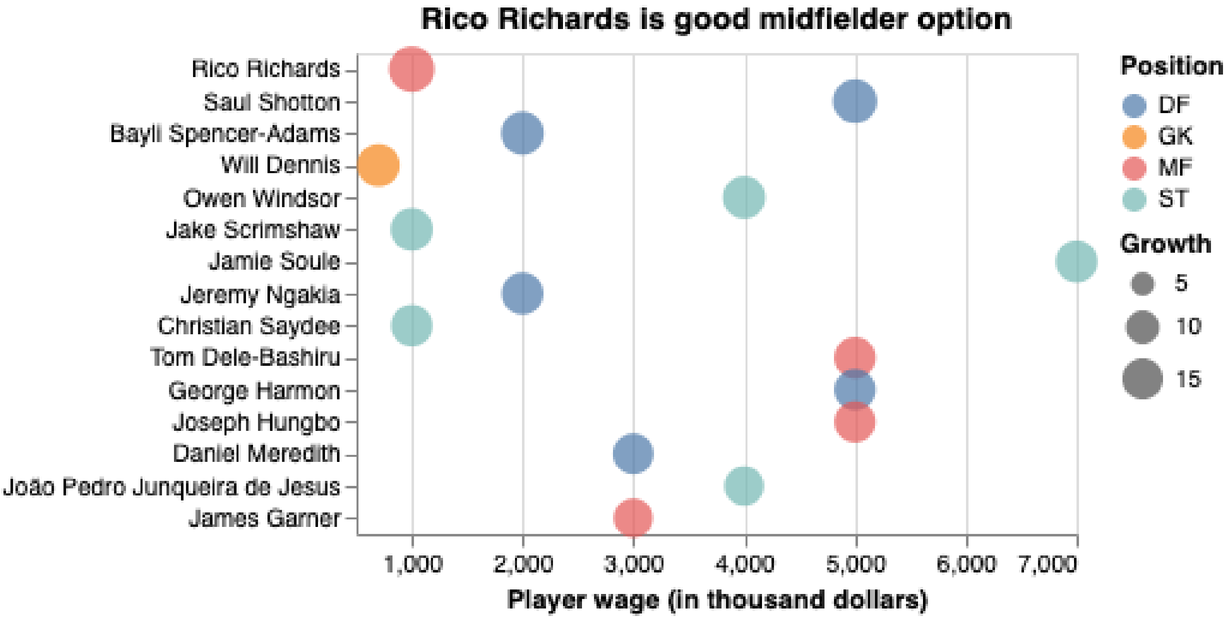
These regression coefficients are statistically significant and ensure that the model performs better than the Null model.

Given below is the table with the top 5 regression coefficients of the highest magnitude.

Coefficient	Magnitude
Age	-4.88
International Reputation_4	2.57
International Reputation_1	-1.88
Following	-1.12
Likes	1.12

Prediction on the test set

The predictions on the test set had an R-2 score of 0.84 which is comparable to the validation score of 0.82. This reconfirms that the model does a good job of generalizing on the unseen data. Since the priority of the club is signing a low-wage player with high growth potential, the below shows the list of high growth potential young players under the age of 25. Rico Richards is the inexpensive youth player in the bottom 3 teams which solves the midfield issues of Manchester United. Along with this, he also has the greatest growth potential in the relegation zone teams.



Conclusion

The problem statement was to point out the key issues in the current team players of Manchester United, commenting on the significant factors for player growth and making a suitable prediction based on that for scouting purposes.

Two issues in the club were the presence of low-growth players and an aging midfield player pool. After assessing the 806 players in the premier league, Rico Richards was a good young midfield player which could be scouted from the relegation-prone teams. The low current wage of the player ensures that the transfer is not a burden on the club's finances. At the same time, it provides solidity and future-proofing the midfield area.

Age was highly correlated to the growth of the football player which was statistically significant with the Ridge regression. In addition to it, the player's international popularity and social media following were also associated with the player's growth. The Ridge regression model had an R-2 validation score of 0.82

and a test score of 0.84. It was selected as the best model out of ordinary linear regression and polynomial regression due to a better validation score. The comparable validation and test score ensured that the ridge model is able to generalize well on unseen data.

Further Improvements:

- In its current version, the model looked at only the Premier League players. Players from across the world can be incorporated.
- The out of position player attributes were not used in the model which could be incorporated to improve the overall assessment of the player

References

[1] https://www.linkedin.com/in/joe-devine-21a95016a/?originalSubdomain=uk

[2] https://www.youtube.com/watch?v=5rMx3UE8XeU

[3] https://en.wikipedia.org/wiki/FIFA_21

[4]https://www.sportperformanceanalysis.com/article/2018/6/8/the-history-of-brentford-football-analytics

[5] https://sofifa.com/

[6] https://github.com/4m4n5/fifa18-all-player-statistics/tree/master/2021