SOCCER PROJECT

Report By: ROHIT AKASH R

This project has been submitted as the QSTP final project held during the summer of 2021

Problem Statement:

You are appointed as the Sporting Director of a newly established football club that aims to be competitive among the top European clubs of the world. You have been assigned to make signings for the main lineup squad of 11 players that must match up to those standards but you have been allotted only with a limited budget of 150 million Euros for the same. You are also required to predict the annual wage of the squad that you have signed from the same dataset.

Dataset Used:

This project uses FIFA 2020 dataset available on Kaggle by Stefano Leone.

Link: https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

Methodology:

This project can be done using the methodology of 'Moneyball' introduced by the book, Moneyball: The Art of Winning an Unfair Game by Michael Lewis. The book argues that there are statistics of players that are very important for their performance but are generally overlooked. This method uses those overlooked attributes to sign players who are undervalued.

For feature selection, this project uses Mutual info regression to calculate the dependency of the players with their overall rating to detect the most important features

Then, a normalized score is generated for the players according to their performance in those features which is used to shortlist the players and a value-for-money squad is chosen by considering other external factors like age, position, etc.,

As final part of the problem, the dataset is used to predict the wages of the players using Regressor with pipeline processing

Procedure:

1. Data Preparation:

First the data is categorized into basic info, skill attributes, positional attributes. Another categorization is done based on the player's team_position, national_position, player_positions based on which position they play. These positions are:

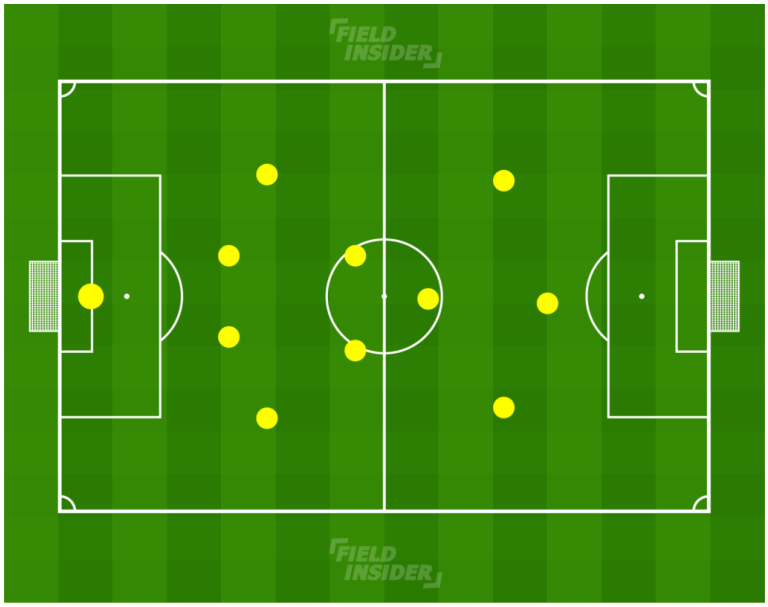
- Goalkeeper
- Defenders
- Fullbacks
- Defensive Midfielders
- Attacking Midfielders
- Wingers
- Striker

While there are many more positions available in a football line-up, these are the basic player categories. Though attacking and defensive midfielder fall under the same category, but they are separated here for convenience and positional similarity with the chosen model European team

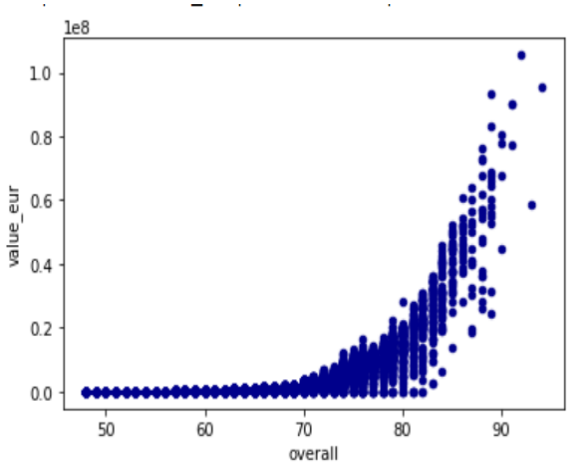
2. Formation Planning:

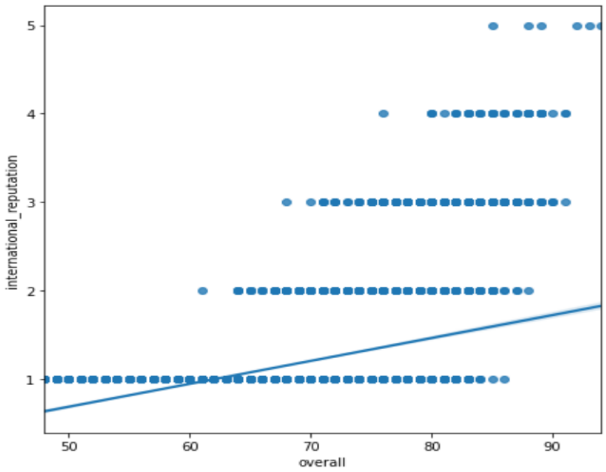
The Model European team chosen for the line-up and formation is Manchester City. Manchester City competes in the English league and is one of the most consistent and successful teams in the world. Even though, its budget limit surpasses our team's budget by a huge margin, Manchester City has recently made many good value-for-money signings with huge potential. Manchester City's Manager, Pep Guardiola is one of the most successful managers in football history and he uses a stable formation with the right balance in attack and defence. The following image shows his formation this year.

The squad that has been planned for the project is quite similar, but with 2 Defensive Midfielders (DMs) and one attacking midfielder instead of the to Central Midfielders and 1 Defensive Midfielders in the shown line-up to maintain the balance. But flexibility is also provided with an optional signing of an Attacking Midfielder in case the Team Management wishes to follow the above model. This is a typical 4-3-3 formation.



Then, 'overall' property, which is given as the overall rating of a player by FIFA is checked to see if the property depends only on skill level of the player or other factors too. But it is observed that the overall rating increases with increase in the player's market value and international reputation.





This means that overall rating of a player also depends on their popularity in the football world. So there needs to be another property that calculates the player's rating only based on their skill level. To do this Feature Selection is done.

3. Feature Selection:

Feature Selection is done to identify the most important skill attributes of a player with respect to the overall ratings of players playing in that position. This is done to find any specific skill attributes that players with high overall ratings tend to be particularly strong at. These attributes are identified for each position stated above using Mutual Information Regression. Mutual information (MI). between two random variables is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. In this case, the overall rating property is compared with each of the skill attribute and relevant skill attributes are filtered by selecting all the properties with MI Score greater than or equal to 0.6. While Correlation constant was initially planned to be used, it was later decided that Mutual Information Regression was to be used because it can detect any kind of relationship, while correlation only detects linear relationships. Further advantages to Mutual Information Regression are:

- It is easy to use and interpret,
- It is computationally efficient,
- It is theoretically well-founded,
- It is resistant to overfitting, and,
- able to detect any kind of relationship

Finally, the following attributes were selected for each of the player category:

GoalKeeper:

1. 'gk_positioning',
2. 'goalkeeping_positioning',
3. 'goalkeeping_diving',

4. 'gk_diving',
5. 'goalkeeping_reflexes',
6. 'gk_reflexes',
7. 'goalkeeping_handling',
8. 'gk_handling',
9. 'movement_reactions'

Defender:
1. 'defending',
2. 'defending_standing_tackle',
3. 'mentality_interceptions',
4. 'defending_marking',
5. 'defending_sliding_tackle',
6. 'movement_reactions',
7. 'attacking_heading_accuracy'

FullBack:
1. 'defending',
2. 'movement_reactions',
3. 'mentality_interceptions',
4. 'defending_standing_tackle',
5. 'defending_sliding_tackle',
6. 'skill_ball_control',
7. 'attacking_short_passing',
8. 'defending_marking'

Defensive Midfielder:
1. 'movement_reactions',
2. 'defending',
3. 'attacking_short_passing',
4. 'mentality_interceptions',
5. 'skill_ball_control',
6. 'passing',
7. 'defending_standing_tackle',
8. 'skill_long_passing'

Attacking Midfielder:
1. 'skill_ball_control',
2. 'movement_reactions',
3. 'passing',
4. 'dribbling',
5. 'attacking_short_passing'

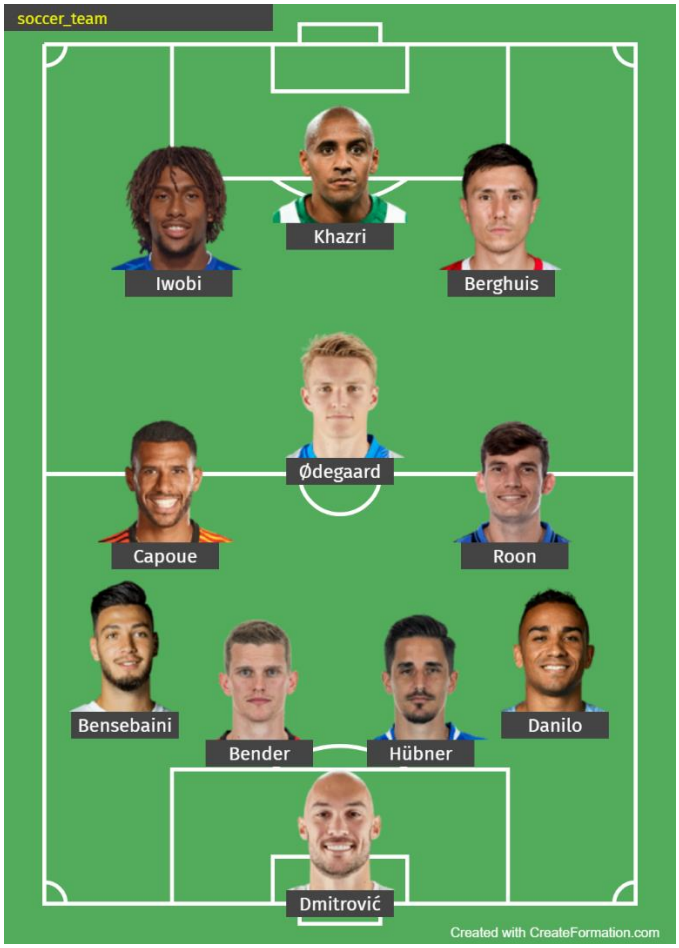
Winger:
1. 'skill_ball_control',
2. 'movement_reactions',
3. 'dribbling',
4. 'attacking_short_passing',
5. 'passing',
6. 'skill_dribbling',
7. 'mentality_composure'

Striker:
1. 'shooting',
2. 'mentality_positioning',
3. 'skill_ball_control',

```
4. 'movement_reactions',
5. 'attacking_finishing',
6. 'power_shot_power',
7. 'dribbling',
8. 'mentality_composure'
```

4. Final Squad Selection:

A normalized score is generated for each player based on their skill levels on these important attributes and that score is used to shortlist top ten players for each position. Further filtering is done using the budget limit and age limit. Other factors such as positioning strength and work rate, etc., are used to finally select the squad from the compiled list.
The main squad signed is provided after generating the line-up in createformation.com as image:



The final budget of the squad came up to 136 million Euros against the set budget limit of 150 million Euros. If the Optional signing (Index Number 862, Player Name: Denis Suarez), was also signed, the total budget comes up to 149 million Euros, which is still under the budget Limit.

5. Wage Prediction:

As a final part of the project, the Dataset minus all the signed players was taken as the training data and the main squad was taken as the test data. Numerical Variables are imputed using Simple Imputer and Categorical Variables are imputed using One Hot Encoding. Finally, prediction is done using XGBoost (Gradient Boosting). Gradient boosting is a method that goes through cycles to iteratively add models into an ensemble. This ensemble method is one of the most accurate prediction methods by combining the predictions of several models. The final wage sheet is prepared with the wage of all the players and final Wage of the squad comes up to 492657.537 Euros. (Note that this value might change each time the prediction model is run).

Note:

- There may be discrepancies in the process of Squad selection and wage prediction of noise within the prediction models, outdated dataset (this dataset gives the data for the 18/19 season), inaccurate attributes allotted by FIFA 20 Game, etc.,
- Further many other properties of the players have been ignored due to lack of available data and other complications like injury proneness, being loaned to other teams, financial regulations, etc.,
- These factors are overlooked in this project because this project has been done for educational purposes only and is not intended to be used for real life football prediction situations.

Result:

A competitive squad has been signed with limited budget and final wage of the squad has been predicted and proposed. The methodology used in this project has several use cases in sports analysis like scouting, squad selection, player analysis, team analysis, football score prediction, etc., Its use case can also be extended to other domains like Management, Biotechnology, Pharmaceuticals, etc.,