Analyzing and Predicting FIFA Player Data

Alaa Ali
ala527@g.harvard.edu

Masaoud Haidar
masaoudhaidar@college.harvard.edu

Vinit Patankar
pvinit@hotmail.com

Robert Roessler
robertroessler@g.harvard.edu

TF: Jovin Leong
jovinwleong@gmail.com

INTRODUCTION

The past decade might very well be described as the decade of data science. All academic fields as well as professional domains have tried to include or a least experiment with new approaches that are commonly branded with buzz words such as 'big data', 'machine learning', or 'artificial intelligence'. While in this context our first association might be new data driven breakthroughs in medicine or the internet of things, this also hold true for professional sports, particularly the ones that are being watched by billions and that generate billions of revenue each year such as soccer. In an interview, FC Sevilla's Sporting Director Monchi, who is regarded as a soccer mastermind and brilliant strategist in the soccer community, stated that "Big Data is the future of football" [2]. The writing is already on the wall: In the soccer world, more data than ever is recorded, far beyond score lines or basic match statistics. Scouts have started to gather data on player movements on the pitch, passing accuracy for various types of passes, dribbling success rates, and a lot more. Scores and rankings have been developed to classify players, which are starting to strongly influence manager decisions on whether to buy a player or not.

In this project, we will be using player data from the soccer simulation video game FIFA to develop models that are able to answer relevant domain questions and make insightful predictions. After describing the data and laying out the general impact that our project could have, we will present our models and findings for our four main project tasks:

1. Predicting the overall rating (OVR) of a player
2. Predicting the player positions
3. Predicting the best staff of all teams in Division 1 European League
4. Predicting future skill changes of all the players

GENERAL IMPACT

As mentioned briefly in the *introduction* and as will be laid out in greater detail in the *data* paragraph, we will be using data from EA Sports' FIFA video game in this project. This might lead to the hasty deduction that our project has barely any applicability in the real world. This would be a misconception for two reasons: First of all, the data in the simulation is not just data that was generated solely for a virtual universe detached from the real world. Rather, the data is generated by EA Sports hired FIFA Talent Scouts[1] that evaluate players and their skills based on actual game performances. Furthermore, the tasks we're working on in the project are fundamental in the professional lives of soccer coaches, managers, and scouts. Our models could thus be adapted by clubs and fed with actual real world data that a club has gathered though their scouting network to make meaningful predictions that could have a positive impact on a soccer club's day-to-day operations. Our models can predict overall ratings and how players are expected to improve, which could help in making decisions of whether to buy a player or not. Our models could also make suggestions on where to place a player and what other positions might suite their skill set. Finally, we offer a way to evaluate club staffs, which helps clubs with hiring and helps young players in choosing their destinations.

DATA

As pointed out earlier, the data set that we're using for this project comes from the player database of the FIFA video game series. It has been generated by a network of FIFA Talent Scouts that have been hired to evaluate more than 15.000 soccer player for each year between 2015-2020. For each player, there are dozens of features ranging from meta data such as name, age, contract, etc. to detailed skill related attributes such as shooting, dribbling, or passing. It is worth mentioning that there are actually two groups of player attributes:

- a set of 29 very specific skills (acceleration, sprint speed, heading, jumping, vision, short pass, long pass, etc.) used for the core FIFA game
- a set of 6 more general but comprehensive skills (pace, shooting, passing, dribbling, defending, physical) used for the so called 'FIFA Ultimate Team' mode (FUT), which are the results of an aggregating algorithm the FIFA game uses on the first set of the 29 skill attributes.

In addition to that, there are various other attributes:

- player-related meta-data such as name, age, height, weight, as well as club-related meta-data such as club name, contract length, and salary.
- data regarding player positions showing the role in the club and in the national team and position specific ratings.
- overall rating and potential of players.

y

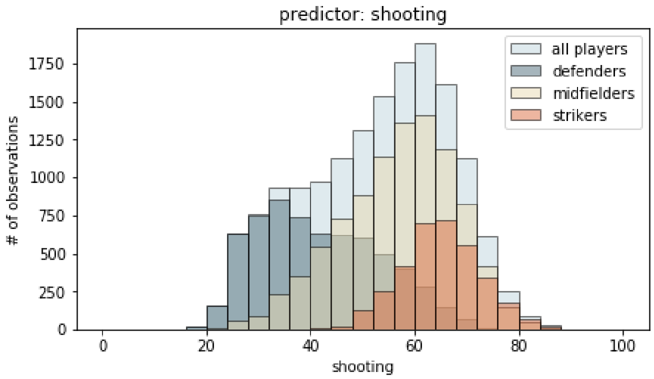


Figure 1. ***Distribution of the skill “shooting” across players. The sky blue distribution across all players is not very informative. Only after encoding position visually, we can see that shooting can be regarded a strong predictor***

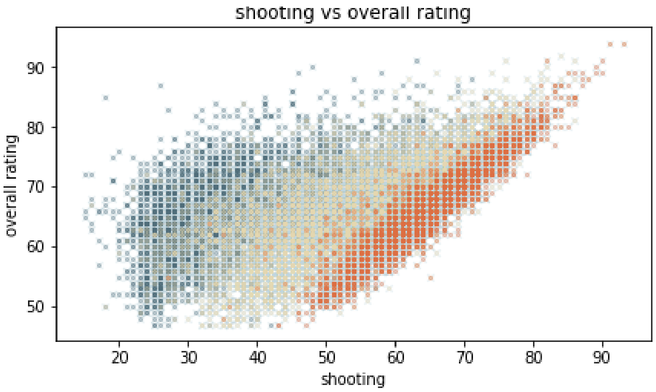


Figure 2. ***The scatter plot shows that there are linear correlations between the skill “shooting” and overall rating if you factor the player position in***

EXPLORATORY DATA ANALYSIS

The following section is divided into four different aspects, namely *Distribution of Skill Attributes*, *Player Positions*, *Best Staff*, and *Skill Changes* that will bring to light some helpful insights in preparation of the four main tasks.

Distribution of Skill Attributes

For our initial EDA, we have used the set of the 6 FUT skill attributes as they are distinct and meaningful skill categories and, thus, potentially strong predictors. In contrast, we know from our own robust domain knowledge that many of the 29 specific skill attributes would be collinear predictors (e.g. *acceleration* and *sprint speed*, which are aggregated to *pace* in the FUT set, to name just one out of many possible examples). For each of the 6 skills, we then plotted both a histogram (cf. Figure 1) as well as a scatter plot (cf. Figure 2) to analyze not only the distribution of a skill across players (i.e. a histogram for shooting, passing, etc.) but also to see how a skill is related to our response variable *overall rating* (i.e. a scatter plot with the skill on the x axis and the overall rating on the y axis). While we can show only a subset of these visualizations in this paper (with shooting being our exemplary predictor), all visualizations can be found in the accompanying notebook.

Initially, these two visualization types didn't allow us to draw any meaningful conclusions in terms of predictive strength of a skill - each skill seemed normally distributed across players. We then visually encoded also the player positions such as defender, midfielder, or striker to identify patterns. This was a key step as the visualizations now showed a clear linear trend for each position between overall rating and the four skills *shooting*, *dribbling*, *passing*, and *defending*. This observation had strong implications for our first task, which was to build a model that would predict the overall rating of a player. Not only were we able to find out the best predictor but also that including dummy variables for the various positions and using interaction terms would most likely help to improve our model.

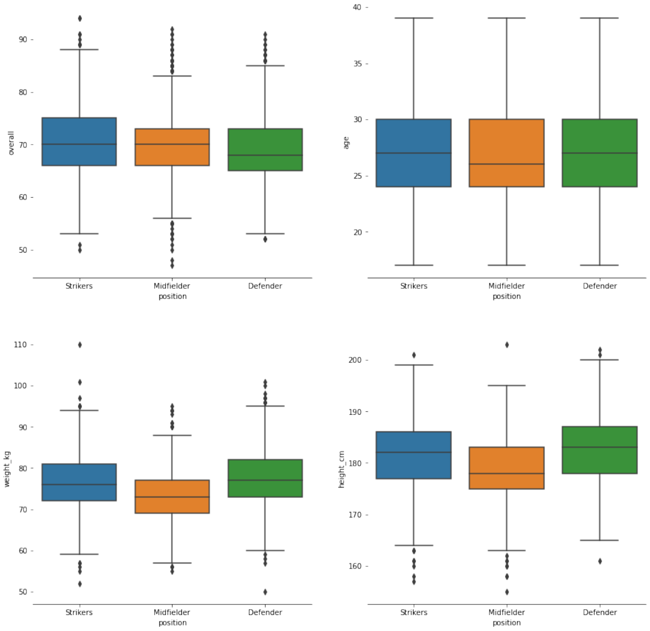


Figure 3. ***The four box plots show the the relationship between the three team positions vs four of our data attributes (age, ’height cm’, ’weight kg’, overall)***

Player Positions

To find the player position, we focus on three groups which are ”Strikers”, ”Midfielder”, and ”Defender”. For example: all players with the Central Attacking Midfielder (CAM) and Right Defending Midfielder (RDM) attributes are all categorized to be ”Midfielder”.

We started with a countplot figure to plot the count for each category under the different team position (Strikers, Midfielders, and Defenders) We noticed that the total count of Midfielders is larger than that of Strikers and Defenders.

In order to plot some physical appearances such as age, height and weight and see effect on the players positions, We provide four boxplots to show the relationship between the different team positions (Strikers, Midfielders, and Defenders) vs four of our data attributes (age, ’height cm’, ’weight kg’, overall) as shown in the figure 3.

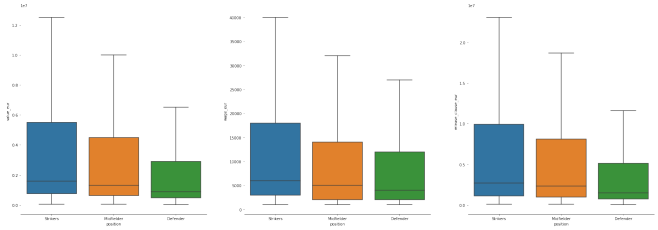


Figure 4. *The three box plots show the the relationship between the team positions vs the value attributes (value eur, wage eur, release clause eur)*

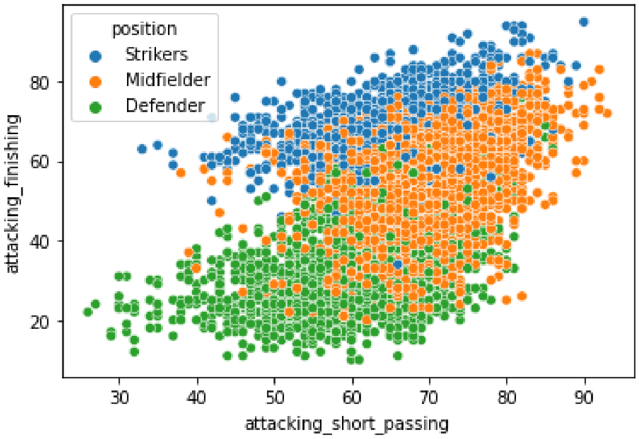


Figure 5. *Scatter plot attacking finishing by attacking short passing classified by position*

- There is no significant difference in Overall Rating among all three positions.

- Midfielder tends to be shorter and lighter since they need to be more flexible.

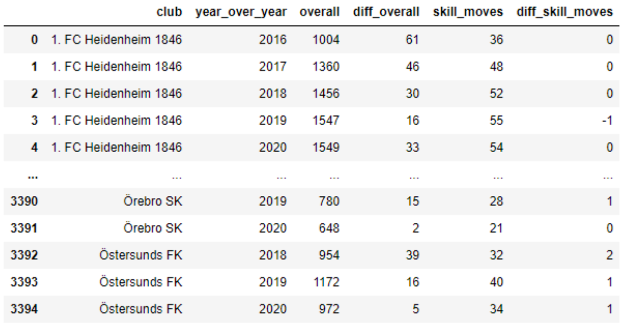
A countplot figure to plot the count for the preferred foot attributes (either left, or right) for each team position (Strikers, Midfielders, and Defenders)

- All three positions, Strikers, Midfielder, and Defender are heavily right feet used.

Similarly, we provide three box-plots to represent the relationship between the different team position (Strikers, Midfielders, and Defenders) vs three of our data attributes which are relevant to each player value (value eur, wage eur, release clause eur) as shown in fig. 4. Strikers tend to have higher value, wage, and Release Clause compared to other positions. Defender position has the lowest value, wage, and Release Clause.

In order to provide Scatter plot for the skills and positions, we plot the attacking finishing by attacking short passing as an example to be classified by position as shown in Fig. 5.

We noticed that the Defender player position has a lower skill move score compared to Midfielder and Strikers player positions. This explains why the Defender class has values less than the other positions due to not performing good eye-catching moves due to their low skills move scores.



	club	year_over_year	overall	diff_overall	skill_moves	diff_skill_moves
0	1. FC Heidenheim 1846	2016	1004	61	36	0
1	1. FC Heidenheim 1846	2017	1360	46	48	0
2	1. FC Heidenheim 1846	2018	1456	30	52	0
3	1. FC Heidenheim 1846	2019	1547	16	55	-1
4	1. FC Heidenheim 1846	2020	1549	33	54	0
...	...	...	...	...	...	...
3390	Örebro SK	2019	780	15	28	1
3391	Örebro SK	2020	648	2	21	0
3392	Östersunds FK	2018	954	39	32	2
3393	Östersunds FK	2019	1172	16	40	1
3394	Östersunds FK	2020	972	5	34	1

Figure 6. *Data Frame representation for the prepared data frame that represents the aggregated overall difference at a club level*

Best Staff

The available data set comprises several attributes for individual players. The provided features for the players comprise the skills the players are measured upon. We decided to leverage the numerical attributes that are associated with the players to ensure that we could quantitatively evaluate the staff. In addition a players improvement is reflected by the change in the score year over year. Our data analysis hence focused on computing the change in the values for the different parameters.

The team initially considered only players who stayed at the club year over year to evaluate the impact of the coaching staff year over year. However, given not many players stay at the club, we decided not to enforce the stay at club metric. The credit for the increase is assigned to the club in the earlier year. This approach ensured that we had a comprehensive data set of all the players and not only the ones that played for the same club.

Our initial EDA earlier revealed that age of the player is an important metric to determine the effectiveness of the club. However on investigation we realized that certain clubs with younger players would get significant score increases over other clubs which had a mature player population. The grouping of data based on age was discontinued.

The prepared data frame shown in (cf. Figure 6), shows changed values which are prefixed with diff, and actual values of the metrics from the year. The values are also aggregated by using sum every year for the club. The reason the sum was selected as the aggregating operation is because that helped in identifying the clubs that have improved their players over multiple years vs those who had data only for a few years.

Skill Changes

To predict how things will be in 2021, we decided to build models that predict changes from a year to the next one. This means that to predict the data for 2021, we would use the same data as 2020 and add the predicted change to each variable. This is because most of the data don't change over one year. See figure 7 for an example.

To predict the changes in a variable for a year, we use all the data we have for the previous year. This approach assumes that knowing the data for a year makes the data for previous

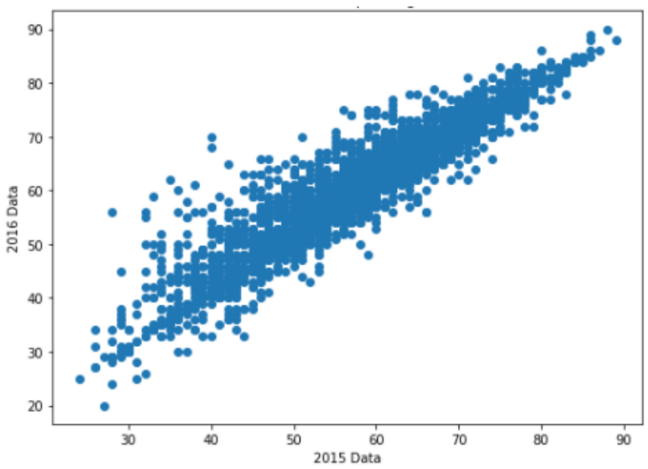


Figure 7. ***Scatter Plot of the Passing attributes of all players in 2016 vs 2015. We can see that most data points are on the line y=x, which means that most players don't have a change in passing from 2015 to 2016.***

years irrelevant, an assumption usually common when studying Markov Chains. Although this assumption might not be ideal, as players might have trends of development or decline over multiple years, we think it should still work in general. In addition, it simplifies our work and allows us to have a lot of data, about 60k observations of transitions from year to year.

Looking into the data we have, We can divide our variables into three groups:

1. Variables that aren't useful for predicting changes and aren't interesting to predict their changes, such as names, urls, jersey numbers, etc.
2. Variables that are useful for predicting changes but aren't interesting to study their changes, such as age and team position.
3. Variables that are interesting to predict their changes and are useful as predictors. This includes the rest of the numerical variables.

For this task, we have 73 different response variables. This can make the task really hard. We don't want to do individual modeling for every single response variable. Luckily, we can group some of our response variables together. We believe that models will perform similarly on variables from the same group. This also allows us to take one example from each group to do visualizations and interpretation. We will also use the same hyper parameters for variables within one group. These groups are:

1. Base skills: The 6 general skills,
2. Specific skills: The 29 detailed skills.
3. Position ratings: Position-specific ratings.
4. Miscellaneous: Only includes the overall rating

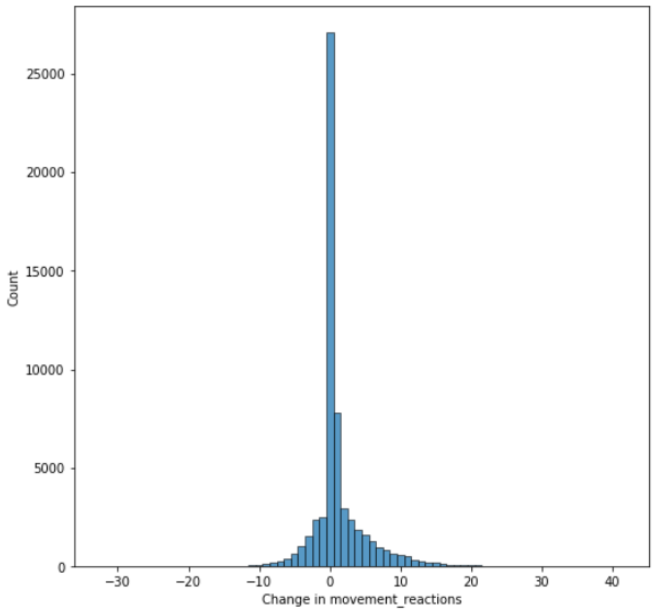


Figure 8. ***Histogram of changes in the movement reactions skill for all player transitions over all five years. We can see that the most common value of change is 0 and the changes are 0-centered with a slight positive skewness.***

After we prepared the data, we visualized histograms of changes. Take Figure 8 as an example. For all histograms we visualized, the data is 0-centered with a slight positive skewness. We believe that this skewness can indicate a yearly inflation in player ratings. Also, the values of 0 were so common, with about third of the data points being unchanged from year to year.

We also visualized changes vs age. Take Figure 9 as an example. All visualizations showed that the changes have a negative correlation with the age. The older a player is, the less their change would be, with possibly a negative change. This is consistent with our general knowledge about athletes, that they develop a lot when they are young until they reach their peak, and then they start declining.

TASK 1: PREDICTING OVERALL RATING

Approach

For the task of predicting overall rating, we decided to build a basic model initially that would help us to compare and evaluate other, more advanced models. These more advanced models would be based on our findings in the initial exploratory data analysis, namely the use of dummy variables and interaction terms to account for different player profiles that could result in high overall rating despite strong differences in individual attributes. We would then develop an optimized model using lasso regression before building an ANN to see whether these techniques could help us improve the accuracy even more.

Modelling

Our base model is a linear regression model using the 6 FUT attributes (c.f. section 'Data') as predictors.
We then continued to use include dummy variables and interaction terms for player positions to train a second, more advanced linear regression model. In a next step, we used lasso regression to optimize our model even further. Lastly,

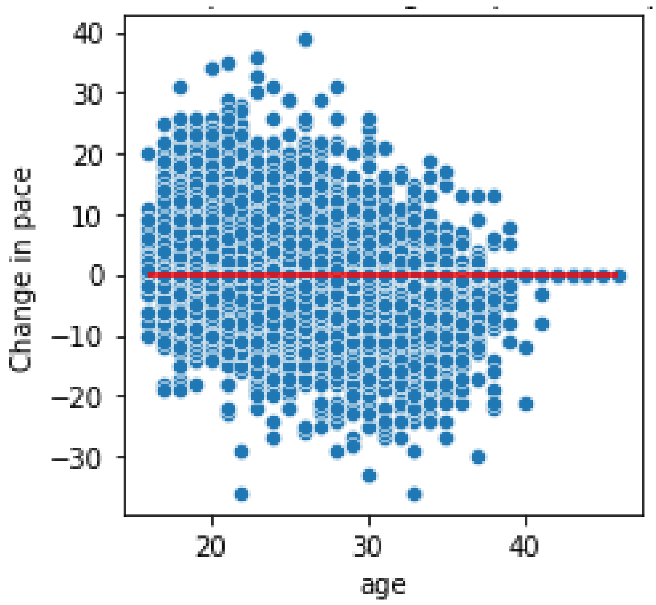
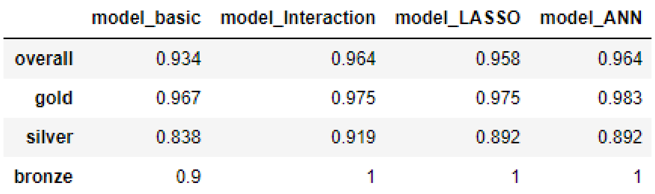


Figure 9. *Scatter plot of changes in Pace vs age. The line y=0 is marked with red. We can see that the younger a player is, the more likely that they will have a high positive change in pace, and vice versa. This could be because of the natural development and decline of players over their career.*



	model_basic	model_Interaction	model_LASSO	model_ANN
overall	0.934	0.964	0.958	0.964
gold	0.967	0.975	0.975	0.983
silver	0.838	0.919	0.892	0.892
bronze	0.9	1	1	1

Figure 10. *Basic model performance evaluation and scoring metric - prediction accuracy for ranking players in gold, silver, or bronze tier*

we trained an ANN on the data (including the dummy variables) to examine whether this could help us improve our predictions even more.

Evaluation

Initially, we were using MSE to compare the performances of our models. But while the MSE might be a useful initial indicator that could help us with model optimization, it is not necessarily a meaningful measure for a reader of this paper let alone a soccer coach, or talent scout. Instead, we started looking into classification and ranking metrics to ultimately provide our target audience with some sort of accuracy score that would give them a clear understanding of the precision of our predictions. Conveniently, the FIFA game itself has already such a system in place that we were able to use as our initial metric: They have gold, silver and bronze tier players. With a maximum score of 99, gold players have an OVR of 75+, silver players are between 74 and 65, and bronze players are 64 and lower. We would use this tier-based ranking system as a first indicator to measure how accurately we could predict OVR (cf. Figure 10).

However, this ranking system had quite a few disadvantages.

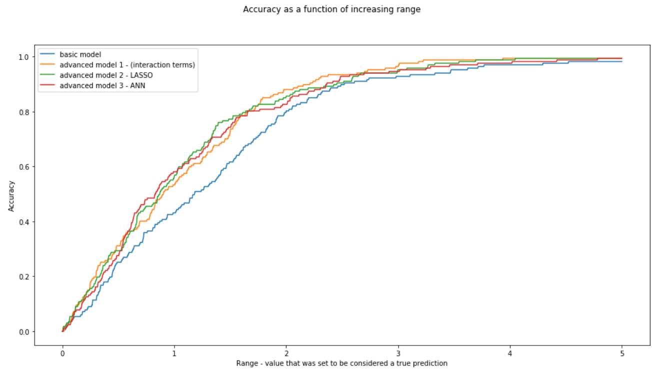
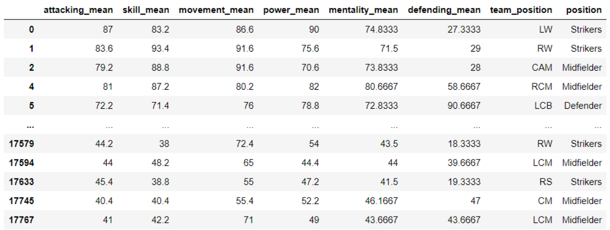


Figure 11. *Range-based performance evaluation and scoring metric - calculating the accuracy for each model based on the current range in which a prediction is considered to be accurate*

Not only seem the three threshold values that would determine the tier rather arbitrary particularly for real-life applications, but it would also allow for predictions that are off by a dozen of OVR points to be counted as accurate as long it is in the right tier (i.e. a player with a predicted OVR of 75 would be considered a gold tier player even if his actual rating is 85). Thus, we developed a metric that would allow domain experts to define and set their own accuracy range, i.e. a range in which they would consider a prediction to be accurate. To provide an example for a range with value 2: If a player has an actual OVR of 85, then a prediction is only to be considered accurate if the absolute value of the discrepancy between prediction and actual OVR is smaller than 2, i.e. if the predicted OVR is larger than 83 but smaller than 87. To prepare a visual representation of our metric, we looped over a linspace of ranges from 0 to 10 and calculated the accuracy for each range based on the discrepancy between predicted and actual OVR. Figure 11) shows the results of his metric.

In contrast to our initial gold-silver-bronze scoring method, this new metric allowed for a much more nuanced evaluation. Figure 11) shows that there is not one best model overall, but different models with alternating performances in different ranges. In other words, the different models are suited for different and slightly tweaked domain tasks. E.g. let's assume that a sporting director asks a scout to evaluate a player but allows only for a margin of error of one OVR point. In that case, the best model for the scout would be the ANN model and the chances that his prediction is within the demanded margin of error is 55%. However, if the sporting director is not happy with 55%, but in return allows the scout to be off by 1.5 OVR points, then the best model would be the Lasso model, which will predict with an accuracy of 80% for that range. Lastly, if the sporting director is still not happy and decides that he want's an accuracy of at least 90% and he states the margin of error should be as low as possible but as high as necessary, then suddenly, the best model would be the linear regression model with dummy variables and interaction terms, which did worse for small ranges than the other models, but best if the range is greater than 1.8. This goes to show that this new scoring method allows us to evaluate our models very precisely and can enables task-based model selection.



	attacking_mean	skill_mean	movement_mean	power_mean	mentality_mean	defending_mean	team_position	position
0	87	83.2	86.6	90	74.8333	27.3333	LW	Strikers
1	83.6	93.4	91.6	75.6	71.5	29	RW	Strikers
2	79.2	88.8	91.6	70.6	73.8333	28	CAM	Midfielder
4	81	87.2	80.2	82	80.6667	58.6667	RCM	Midfielder
5	72.2	71.4	76	78.8	72.8333	90.6667	LCB	Defender
...	...	...	...	...	...	...	...	...
17579	44.2	38	72.4	54	43.5	18.3333	RW	Strikers
17594	44	48.2	65	44.4	44	39.6667	LCM	Midfielder
17633	45.4	38.8	55	47.2	41.5	19.3333	RS	Strikers
17745	40.4	40.4	55.4	52.2	46.1667	47	CM	Midfielder
17767	41	42.2	71	49	43.6667	43.6667	LCM	Midfielder

Figure 12. ***Data Frame representation for the prepared data frame that represents the aggregated mean values for skills stats***

TASK 2: CLASSIFY PLAYER POSITION

Approach

For this task, we had two step plan. Initially, we build a basic model which leverage the dataset of all players and train a logistic regression model to predict the position of players using their other statistics.

There are many attributes that could be used to predict the team player positions (Strikers, Midfielders, or Defenders). We selected the following attributes to creates the features used for training:

- Five attributes describing the attacking statistics ”attacking crossing”, ”attacking finishing”, ”attacking heading accuracy”, ”attacking short passing”, ”attacking volleys”.
- Five attributes describing the skill statistics: ”skill dribbling”, ”skill curve”, ”skill long passing”, etc.
- Five attributes describing the movement statistics: ”movement acceleration”, ”movement sprint speed”, ”movement reactions”, etc.
- Five attributes describing the power statistics: ”power shot power”, ”power jumping”, ”power stamina”, ”power strength”, etc.
- Six attributes describing the mentality statistics: ”mentality aggression”, ”mentality interceptions”, ”mentality positioning”, ”mentality penalties”, etc.
- Three attributes describing the defending statistics: ”defending marking”,”defending standing tackle”,”defending sliding tackle”.

We decided to find the mean of each group of these attributes mentioned above and use the list of mean values as six main features for training on FIFA 19 data as shown in Fig. 12 to predict the team position to build our baseline model.

Our second approach is to train a neural network model on the FIFA 19 data for training in order to classify the three the team player positions (Strikers, Midfielders, or Defenders).

In addition to classifying player positions, we plan to also recommend secondary positions for each player. This could be important not only in-game (in case you need a player to fill in a position due to injury, etc.) but even in real life, if you want to provide recommendations for coaches or scouts.

Figure 13 shows the heat map of six features correlation by plotting all of the mean numerical features for correlation. Some of the features are correlated which could be useful for doing the features engineering.

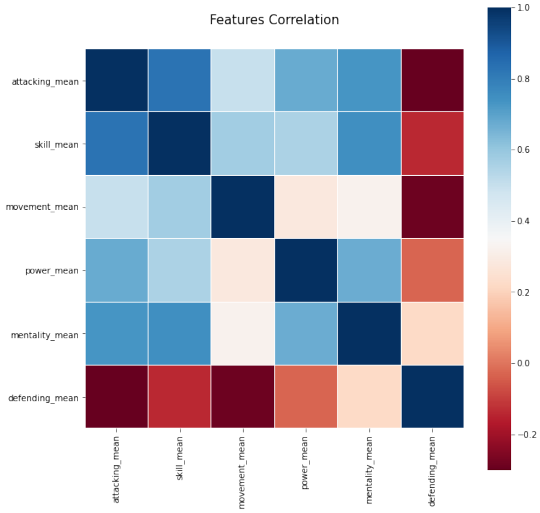
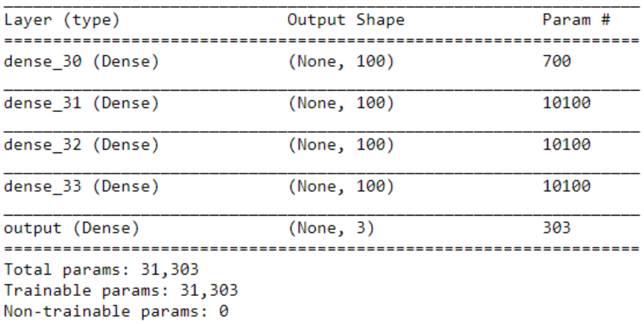


Figure 13. ***heat map of features correlation***



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
dense_30 (Dense)             (None, 100)               700       
_________________________________________________________________
dense_31 (Dense)             (None, 100)               10100     
_________________________________________________________________
dense_32 (Dense)             (None, 100)               10100     
_________________________________________________________________
dense_33 (Dense)             (None, 100)               10100     
_________________________________________________________________
output (Dense)               (None, 3)                 303       
=================================================================
Total params: 31,303
Trainable params: 31,303
Non-trainable params: 0
```

Figure 14. ***Neural Network model architecture for predicting the team player position***

Modelling

We build two multi-class logistic regression model using the six mean features mentioned above in in Fig. 12 from FIFA 19 data to train a classifier for the team position. One model uses Multi-nomial logistic regression which uses the cross-entropy loss and the second model uses the one-vs-rest (OvR) scheme. We added a normalization step to the six mean of skills features using sklearn.StandardScaler processing.

Then we trained a neural network model on the FIFA 19 data for training in order to classify the three main team player positions (Strikers, Midfielders, or Defenders) as shown in Fig. 14.

Evaluation

Once the model was selected to predict the possible team player positions, we used the accuracy classification score to evaluate each of the logistic regression models to determine the validity of the approach.

The initial evaluation of the regression models has revealed that there is an opportunity to further improve the performance of the model in terms of classifying the team player. The evaluation results are shown in Table below.

Model	multi-class	Train Acc	Test Acc
Logistic Regression	OvR	77.21%	43.66%
Logistic Regression	Multinomial	78.51%	41.14%

Additionally, we run performance evaluation for the neural networks models which have been trained on the same six mean value features representing the skill set.

Model	# response	Train Acc	Test Acc
Neural Network	3	79.65%	42.13%
Neural Network	26	33.36%	12.66%

TASK 3: BEST STAFF

Approach

As the clubs staff has to be ranked, our approach is to score every club based upon the players ratings that played for the club that year. The data for all the players in the club was aggregated year over year. Our scoring approach is a hybrid approach that involves using the data available about the improvements year over year. In addition we also identified that just by selecting the overall score alone, we were not able to capture the improvements by the staff. Hence the overall scores differences for the players are a more accurate reflection of the staff's contribution to the improvement of the players. In order to rank the teams based upon the improvement in overall score for the players, we would like to compare teams with identical improvements and come up with a predicted score.. This predicted score would indicate what the club staff should be able to improve in the player and if the actual difference in overall score is better then the club staff has definitely done better. For the sake of prediction, we create a Model with the overall score difference (diff-overall) as the outcome and the individual difference parameters as well as the actual parameters as the predictors. This approach would ensure that 2 teams with similar change in parameters values year over year, would get the same target values in this case the change in overall rating.

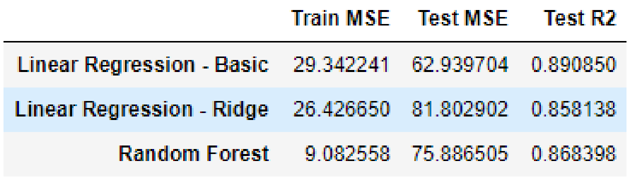
Modelling

As our approach involved predicting the overall score difference based upon the difference metrics captured for the individual feature, we needed to develop a regression model. In order to ensure that we come up with an effective model that fits the train data appropriately, we fitted more than one model on the train set. The following models were fit on the data and the score was calculated for the train and test set. In all the models we have excluded the non numeric parameters in a prepared data set. The diff-overall is the target parameter.

- Linear Regression
- Linear Regression with Ridge Regularization
- Random Forest Regressor with 50 percent max features.

Models were evaluated for MSEs and R2 Scores and below is summary: The Random Forest Model was selected given its low train MSE and better R2 value than the regularized model.

In order to compute the score we implemented 4 scoring formulas:



	Train MSE	Test MSE	Test R2
Linear Regression - Basic	29.342241	62.939704	0.890850
Linear Regression - Ridge	26.426650	81.802902	0.858138
Random Forest	9.082558	75.886505	0.868398

Figure 15. ***Model Metrics for predicting the overall difference for clubs aggregated for individual players***

- Diff-overall - diff-overall-predicted: The rationale behind this approach is that a team that does better than predicted indicating that it did better than other teams and hence definitely contributed to the overall improvement of the players.
- diff-overall * (diff-overall - diff-overall-predicted): This approach gave the diff-overall which is the actual value of the prediction more importance than the approach earlier.
- Overall + (diff-overall - diff-overall-predicted): Using the Overall score in addition to the difference in the overall rating and predicted difference in overall.
- diff-overall + (diff-overall - diff-overall-predicted): In this approach the diff-overall scores have been added to the difference in the actual and predicted values and thus both factors contribute evenly.

Since the scores were generated using the scoring approach outlined are not on a scale, we went further to normalize the scores on a scale of 100. The normalization was achieved using sklearn.MinMaxScaler processing and then multiplied by 100.

Evaluation

Once the model was selected to predict the overall score, we evaluated the individual scoring approaches to determine the validity of the approach. The process in addition to reviewing the generated normalized scores also include manual validations and Subject matter expertise to determine the results. Here is an evaluation of the every scoring approach:

- Diff-overall - diff-overall-predicted: This approach was rejected because the accuracy of the score heavily relied on the accuracy of the prediction. Clubs where the prediction was significantly off were being rewarded and punished significantly in this case.
- diff-overall * (diff-overall - diff-overall-predicted): This approach gave the diff-overall metric, which represents numerical value of player improvements by the club staff over multiple years, appropriate significance. However as the diff-overall is multiplied by the difference, the impact of the incorrect prediction seemed to be identical to the first approach and hence was rejected.
- Overall + (diff-overall - diff-overall-predicted): This approach clearly resulted in some very well known clubs with some good players being selected as the best staff clubs. Clubs with good players are bound to have higher overall scores but do not reflect the quality of the club staff as they

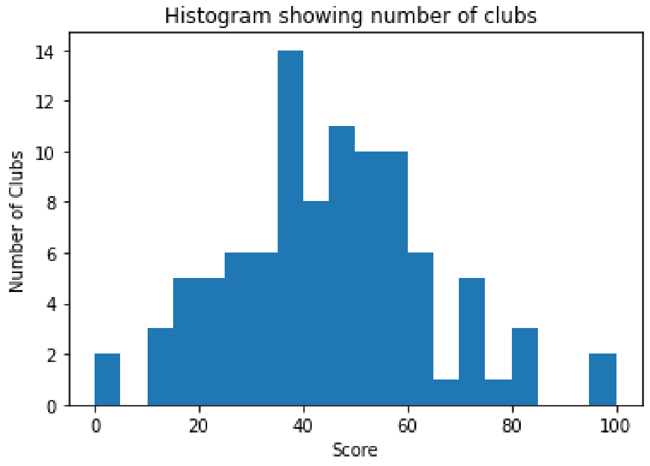
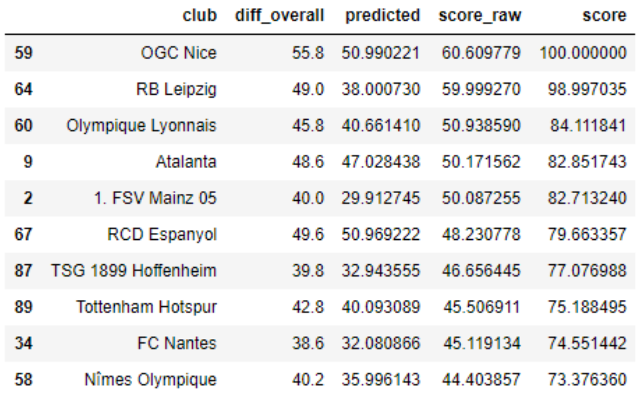


Figure 16. *Histogram of Club Scores - Scoring Approach that combines overall difference in addition to the difference between the overall and the predicted overall*



	club	diff_overall	predicted	score_raw	score
59	OGC Nice	55.8	50.990221	60.609779	100.000000
64	RB Leipzig	49.0	38.000730	59.999270	98.997035
60	Olympique Lyonnais	45.8	40.661410	50.938590	84.111841
9	Atalanta	48.6	47.028438	50.171562	82.851743
2	1. FSV Mainz 05	40.0	29.912745	50.087255	82.713240
67	RCD Espanyol	49.6	50.969222	48.230778	79.663357
87	TSG 1899 Hoffenheim	39.8	32.943555	46.656445	77.076988
89	Tottenham Hotspur	42.8	40.093089	45.506911	75.188495
34	FC Nantes	38.6	32.080866	45.119134	74.551442
58	Nîmes Olympique	40.2	35.996143	44.403857	73.376360

Figure 17. *Top 10 Clubs from the Test Set*

might not be contributing to the improvement of the players and hence diminished the importance of the change in the overall rating significantly and hence was rejected.

- diff-overall + (diff-overall - diff-overall-predicted): This approach actually resulted in a minimizing the reliance on the predicted score by the model while still giving credit if the model predicted a difference. The resulting clubs were also manually validated using other metrics like recent wins and Subject matter expertise and hence was the selected approach for scoring.

We also anticipated that the generated scores would be normally distributed with majority of the teams within a 30 to 60 range and few exceptional teams crossing 80. From the visualization in (cf. Figure 16) for selected approach, we observed that pattern most prominently and validated our scoring approach. Based upon the generated scores, we have provided the list of top 10 clubs rated by our scoring approach in (cf. Figure 17).

TASK 4: OUTLOOK FOR 2021

Approach

As said in the EDA, for this task, we will work with the data of the transitions of the players. To predict the data of 2021, we would use the same data from 2020 and then add the predicted change from 2020 to 2021. Our predictors are the data from each year, and our response variables are the changes in data from that year to the next one.

We will have two approaches to the problem. In the first, we will try to predict exact numerical changes for the players. In the second, we will predict categorical changes: Will this skill for the player increase, decrease, or remain the same in the following year? The first approach helps us predict how things will be in the following year. The second approach is helpful in simply predicting whether the player will improve or not.

Modelling

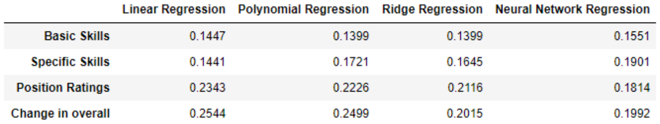
To predict the exact changes between years, we used linear regression. Our response variables are all numerical, so this model is appropriate for the task. In addition, we have 73 different response variables, and thus we will have to train 73 different models in every step, This makes Linear regression very good for this task, as it has a closed form solution and thus, it is less computationally intensive than other models, such as trees or neural networks. Our baseline model is a simple linear regression and our final model on all the variables is a polynomial model of degree 10 with ridge regularization. Here again, we used Ridge Regularization rather than Lasso because it has a closed form solution. In addition, we fitted Neural Network regressors on one example from each group.

To predict categorical changes, aka, whether a skill increases, decreases, or stays the same, we used logistic regression. Our baseline model is a simple logistic regression, and we used polynomial terms and ridge regularization to improve its score.

Evaluation

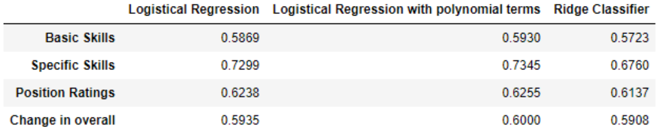
For predicting exact changes, we got the R^2 scores in Figure 18. We can see that the best models for each groups, in order, are neural network, neural network, simple linear regression, and simple linear regression, with linear regression always performing better than polynomial and ridge regression. This shows that the relationship between our inputs and outputs are mostly linear. The R^2 scores we got range from 14% to 25%. This means that our inputs don't explain a lot of the variations in our outputs. We think this is because the data we have is not enough to predict how a player will develop. From our domain knowledge in soccer, the most influential things about the development of a player in a season are the minutes he played in that season and how long he was injured (if he was injured). We don't have this data in our dataset, and so, this highly affects our ability to predict the changes.

For predicting whether a skill would increase, decrease, or remain the same for every player, we got the accuracies in Figure 19. The logistic regression with polynomial features performs the best on all groups. The accuracies of this model on the groups range from 59% to 73%. We think this is a very good accuracy considering that, as mentioned earlier, we don't have data about minutes played and injuries. This model can be used by clubs to predict whether a player would



	Linear Regression	Polynomial Regression	Ridge Regression	Neural Network Regression
Basic Skills	0.1447	0.1399	0.1399	0.1551
Specific Skills	0.1441	0.1721	0.1645	0.1901
Position Ratings	0.2343	0.2226	0.2116	0.1814
Change in overall	0.2544	0.2499	0.2015	0.1992

Figure 18. ***R2 validation scores of the linear regression and neural network models on the different groups of response variables. Note that the neural network scores are just for one example from each group while the rest is the mean score for all the variables of the group.***



	Logistical Regression	Logistical Regression with polynomial terms	Ridge Classifier
Basic Skills	0.5869	0.5930	0.5723
Specific Skills	0.7299	0.7345	0.6760
Position Ratings	0.6238	0.6255	0.6137
Change in overall	0.5935	0.6000	0.5908

Figure 19. ***Classification accuracies of the different logistic regression models in predicting the categorical changes of player skills. Note that these are the accuracies on the validation set.***

improve or not in the upcoming season. This can be helpful for hiring decisions. With these good accuracies, we believe the model can be very helpful for such a task.

CONCLUSION

In this paper, we were able to show that the models that we developed to solve the various domain tasks were able to produce meaningful results. Despite the fact that we were using data from a soccer video game simulation for training and testing, our models could well be fed with real-life data and utilized to assist soccer coaches, scouts, or sporting directors in the day-to-day operations. We can use the trained model not only to accurately predict the player position class but also to recommend secondary positions for each player which would be useful to soccer coaches in-game even in real life. Our trained models can also be used to predict the rating of a player and how he will improve in the upcoming season, which can be helpful for hiring purposes. Finally, our trained models can evaluate the performance of a club's training staff, which can be useful both for clubs to evaluate their staff as well as for players when deciding where to move and what club is the best for their next career step.

While the data set at our disposal provided extensive insights, the player data from the FIFA video game naturally didn't include information about the actual minutes a player played or the injuries a player incurred. In order to apply our models and help with real-life decision making in the soccer domain, adding such information would help to optimize our models even further. Particularly, this would allow us to predict player improvements more accurately. Furthermore, we could expand our staff evaluation also by how they are handling and preventing injuries. Such data could either be provided by a club who wants to use our models or one could use web scraping to generate that information.

REFERENCES

1. How is FIFA Data Collected? `https://www.fifagamenews.com/fifa-data-collected-interview/`.
2. Big Data is the Future of Football. `https://trainingground.guru/articles/monchi-big-data-is-the-future-of-football`.