FootStats - An Application for Statistics and Predictive Analytics in Football

A Project Report

Submitted by

RUSHANG PHIRA
ALOK RAJPUROHIT
ANIKET RAMAN
SHAURYA RAWAT

Under the Guidance

DR. PRACHI NATU

In partial fulfilment for the award of the

degree of

BACHELORS OF TECHNOLOGY
COMPUTER ENGINEERING

At

MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT AND ENGINEERING

FEBRUARY 2022

DECLARATION

I, **Rushang Phira, Alok Rajpurohit, Aniket Raman, Shaurya Rawat** Roll No. **E011, E014, E015, E018** of B. Tech (Computer Engineering), VIII semester understand that plagiarism is defined as anyone or combination of the following:

1. Un-credited verbatim copying of individual sentences, paragraphs or illustration (such as graphs, diagrams, etc.) from any source, published or unpublished, including the internet.

2. Un-credited improper paraphrasing of pages paragraphs (changing a few words phrases, or rearranging the original sentence order)

3. Credited verbatim copying of a major portion of a paper (or thesis chapter) without clear delineation of who did wrote what. (Source: IEEE, The institute, Dec. 2004)

4. I have made sure that all the ideas, expressions, graphs, diagrams, etc., that are not a result of my work, are properly credited. Long phrases or sentences that had to be used verbatim from published literature have been clearly identified using quotation marks.

5. I affirm that no portion of my work can be considered as plagiarism and I take full responsibility if such a complaint occurs. I understand fully well that the guide of the seminar/ project report may not be in a position to check for the possibility of such incidences of plagiarism in this body of work.

Name: Rushang, Alok, Aniket and Shaurya
Roll No.: E011, E014, E015, E018
Place: NMIMS, Mumbai
Date: 22/03/2022

CERTIFICATE

This is to certify that the project entitled **"An Application for Statistics and Predictive Analytics in Football"** is the bonafide work carried out by **Alok Rajpurohit, Rushang Phira, Aniket Raman and Shaurya Rawat** of B. Tech (Computer Engineering), MPSTME (NMIMS), Mumbai, during the VIII semester of the academic year 2021-2022, in partial fulfillment of the requirements for the award of the Degree of Bachelors of Engineering, Integrated as per the norms prescribed by NMIMS. The project work has been assessed and found to be satisfactory.

Dr. Prachi Natu
Internal Mentor

Examiner 1

Examiner 2

Director

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Data Analytics in Sports:

The field of analytics in sport has been very instrumental due to the competitive nature of sport. Players, team managers, coaches and fans often rely on sports analytics before making decisions or developing strategies to win games. Sports are competitive so it’s no wonder many types of statistics are meticulously kept on file to see which players or teams can beat records. Data science isn’t just used in sports to fuel competition between professional players; it also plays a key role in improving game quality, fan experience and player safety. Thanks to these emerging applications, students who earn advanced data science degrees may find themselves providing crucial services to their favourite teams and helping to evolve sports for a new generation. Recognizing the scope of data that could be handled in the sports world, using data analytics would be invaluable. Day by day, the world of sports keeps on improving its capabilities of using sports analytics as a tool to improve their win rate.

Sports analytics can be explained as using data related to any sports or game. Like statistics of players, Weather conditions, Team’s recent wins/lose, etc. With this data, we can create predictive machine learning models to make informed decisions on behalf of the management. The main objective of sports analysis is to improve team performance and enhance the chances of winning the game. The value of a win speaks volumes and takes on different forms like trickles down to the fans filling the stadium seats, television contracts, fan store merchandise, parking, concessions, sponsorships, enrolment, retention, and local pride.

1.2 Role of Analytics in Football:

For most of modern history, coaches had little more than pen, paper, and video to help them analyse play. During the ’90s, Manchester United became one of the pioneering teams to adopt analytics in their decision-making; by 2010, analytics were becoming widely adopted by teams in leading international leagues. In recent years, football experienced rapid technological advances, with platforms able to capture and analyse data from training, match play, internet of things (IoT) devices, and wearables. Coaches now rely heavily on metrics to guide their decision-making and help their teams excel.

There are plenty of software available to measure various stats in football. The use of data and technical analysis has become essential for today's football clubs, and it is no longer confined to the biggest and richest teams. Smaller clubs are making use of it as well thanks to software being cheaper and more readily available than in the past. One such tool is optical tracking, which can be used to pinpoint the position of players on the pitch 25 times a second, in relation to the ball, opposition and teammates. There is also ball-related data such as passes, shots and turnovers, while more advanced tools can analyse defensive stability, pitch control and off-ball scoring opportunities.

The most valuable intelligence for coaches shows what happens in real time. This means coaches can use this data to change the shape of their teams and behaviour to increase their likelihood of winning. These types of insights are mainly gathered from playing logs, video, and GPS tracking, and spatially related data.

Meanwhile, team data can also be analysed as a network, in which nodes represent players and the lines between the nodes represent interactions, such as passes between teammates. Coaches can identify different types of interactions and encode different types of events. This data allows them to identify, change, and test the effectiveness of typical passages of play.

Data analytics is also highly relevant when it comes to youth development. The reason is having objective and measurable feedback that could help both coaches and players speed up the learning processes and create virtuous development cycles. So, data analytics becomes a tool to help predict and cultivate players' potential.

1.3 Predictive Analytics:

Predictive analytics take analysis into the future and can be used to help coaches understand the consequences of altered team formations and other changes. Coaches can tailor training, strategy, and individuals' roles according to data about their next opponents. Business managers can use data in the same way to tailor their approaches to different customers, staffing levels, inventory, and more.

A metric that teams have discovered has predictive power is expected goals (xG), which measures the quality of players' shots in attacking play and the probability the shots will result in goals. xG uses algorithms that account for factors such as distance from goal, angles, and more. Coaches use this metric to try to predict the best positioning of players and patterns of play to optimise scoring opportunities. In this case, analytics provide insights into the most effective strategies to apply in different situations. With that in mind, data can be put to work to help recruit the right team members for future success or fast-track others. Teams use a wealth of data about players' strengths and weaknesses to gain a deeper insight into individual performance to help them and the club succeed. What's required to harness the power of all this data is an analytics platform that can handle huge and growing sets of data points from a multitude of live and cached sources, then visualise it all in ways that can provide fast, comprehensible, and actionable insights.

2. LITERATURE REVIEW

As for any good major project, a lot of research papers were looked at carefully related to the Machine Learning in Sports Analytics field. Some of them were more generic than others, some a little more in depth and they covered a wide range of topics, all mentioned in our features and functionalities for the final project. The summary and some of the learnings from these papers were as follows:

Title	**Author**	**Area of Coverage**	**Algorithms mentioned in paper**	**Important Findings and Methodology**
An examination of expected goals and shot efficiency in soccer	Alex Rathke	Expected Goals	o Logistic Regression o Simple Probability Function	In this paper, the authors look to examine the credibility of XG and whether it maps well against real goals. The dataset used is of shot locations (x and y coordinates) across the 12-13 season from the Premier League and the Bundesliga (18218 values). They then split the pitch into 8 zones, each indicating a shot location and examined the probability if a shot in that zone would be a goal. The expected and actual

				goals have a high relationship value and hence the statistic is credible.
Creating a Model for Expected Goals in Football using Qualitative Player Information	Pau Madrero Pardo	Expected Goals	o Concept Creation o Logistic Regression o Cross Validation o ANN	This paper combines data from two sources (from OPTA for stats and from FIFA for player ratings) to create a sophisticated model. Concepts were created manually and with the eye test and validated with a flowchart. The next thing to do was classify shots as type of shots.. All the variables used in the dataset were listed. All of them were given a value/weight based on some of the dataset to predict further data.
Machine Learning in Football Betting: Prediction of Match Results Based on	Johannes Stubinger	Betting in Sports	o Linear Regression o Random Forest o Boosting o ANN	This paper tried to predict games and thereby set odds on these games. The parameters used are player performance, team performance, whether it is a home or away game. Based

Player Characteristics				on the given data, we predict which team has what probability of winning. The probability is calculated based on data from all football game results (home win/away win/draw), matches between the two teams in question, whether any player from either side has a decent record vs the other etc.
Sports Analytics algorithms for performance prediction	Konstantinos Apostolou, Christos Tjortis	Player Performance Prediction	o Linear Regression o ANN	This paper aims to propose a prediction approach on the football data by implementing Machine Learning algorithms. The predictions are made for player's position, number of goals per season, and number of shots per match. The predicted results are visualized by bar charts and compared to the real results. It was observed that the accuracy of the predicted results was good in all the

				three experiments.
A study of Prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market	Dibyanshu Patnaik, Harsh Praharaj, Kartikeya Prakash, Krishna Samdani	Footballer value prediction	o Multilevel Regression Analysis o MLP Neural Network	The paper examines the various approaches to predict the value players. It evaluates all the approaches and concludes the best one. The various data modelling approaches are stated. The crowd-based estimation model works on the parameters on which the public judges the value of the player. Next approach involves multilevel regression analysis, which involved the implementation of several models each considering different features as significant. The third approach involves the use of an index score to assess the player's value. An MLP neural network was implemented on the players' data with various variations of the hyperparameters. In the end, it is concluded that

				the multilevel regression approach is the best among the others.
Towards data-driven football player assessment	Rade Stanojevic, Laszlo Gyarmati	Footballer value prediction	o Linear Regression o Reverse Linear Regression	This paper a model assesses a player’s value based on performance data. 2 data sources used: one provides the player game stats of and the other provides the estimated market value. A regression model is developed after performing feature extraction where the stats, physical data and team-related information is included. The performance is compared to other methods. The paper also attempts to check the vice versa effect as to whether the market value affects the performance on the field but no conclusion is stated in this case.
Beat the Bookmaker – Winning	Julian Knoll	Betting in Sports	o SVM o Random Forest	In this paper, a data-driven framework was developed in order to predict the outcome

Football Bets with Machine Learning			o Boosting	of football matches which would generate meaningful profits by betting. This paper describes a way to estimate the outcome of football matches based on large data sources and common machine learning algorithms.
Support Vector Machine–Based Prediction System for a Football Match Result	Chinwe Peace Igiri	Football score prediction	o SVM o Neural Networks	The time-series data compose a sequence of data that is collected at regular intervals over a period of time. In this case, it is a set of data built from football match history. Players' performance and manager indices were gathered from the 2014-2015 season of the English Premier League. Only 53.3% accuracy was obtained using SVM, while an 85% was attainable using Neural Networks, according to the paper. Thus, it concludes that SVM is not an appropriate method for this purpose.

A Data Science Approach to Football Team Player Selection	Dr P Rajesh, Bharadwaj Dr Mansoor Alam, Dr Mansour Tahernezhadi	Scouting Footballers		This paper presents a data science approach to minimize the time taken in selecting a player for a team by considering the cost and player's skills as constraints It presents statistical analysis of player performance based on abilities and skills for a new team using powerBI and Python Pandas by minimizing the cost. The results show that it leads to improved business profits through a systematic enhancement to football data sets.
Modeling of Football Match Outcomes with Expected Goals Statistic	Adan Partida , Anastasia Martinez , Cody Durrer , Oscar Gutierrez and Filippo Posta	Expected Goals	o Logistic Regression o Simple Probability Function	The study in this paper explored the possibility of using publicly available data to create straightforward mathematical models to predict the most likely outcome of a football match. In their approach, goals scored by a team are

				predicted almost independently from a team’s opposition. We only used game related data instead of trying to incorporate too many variables such as form, injuries, line-up, game importance, etc.

3. PROPOSED SYSTEM

- The end product of our project is going to be a web app.

- This detailed web app will provide a list of ML based functionalities from the world of football.

- This list of functionalities included in the app included but are not restricted to – Predication of player's market value, Form Prediction/Analysis, Expected Goals Prediction in a match situation and suggesting a player's best position based on his skills and attributes.

- The primary aim of the app is to show stats on how teams have performed in betting markets, to help the user to find trends or make betting selections. Although this means the app does not include every possible feature relating to football, this app is so far ahead of the market in this area, providing a view of fixtures that isn't available elsewhere, that this is a vital download for football fans and bettors.

- The app also includes predictions, in most of the main betting markets, and value bets.

- A unique app. With the data powered by live Stats, you can review and analyse an enormous number of data points, including goals, shots, dribbles, passes and many more. These data points are available for matches in 11 competitions, including the Premier League, La Liga, Serie A, Bundesliga, Champions League and Europa League.

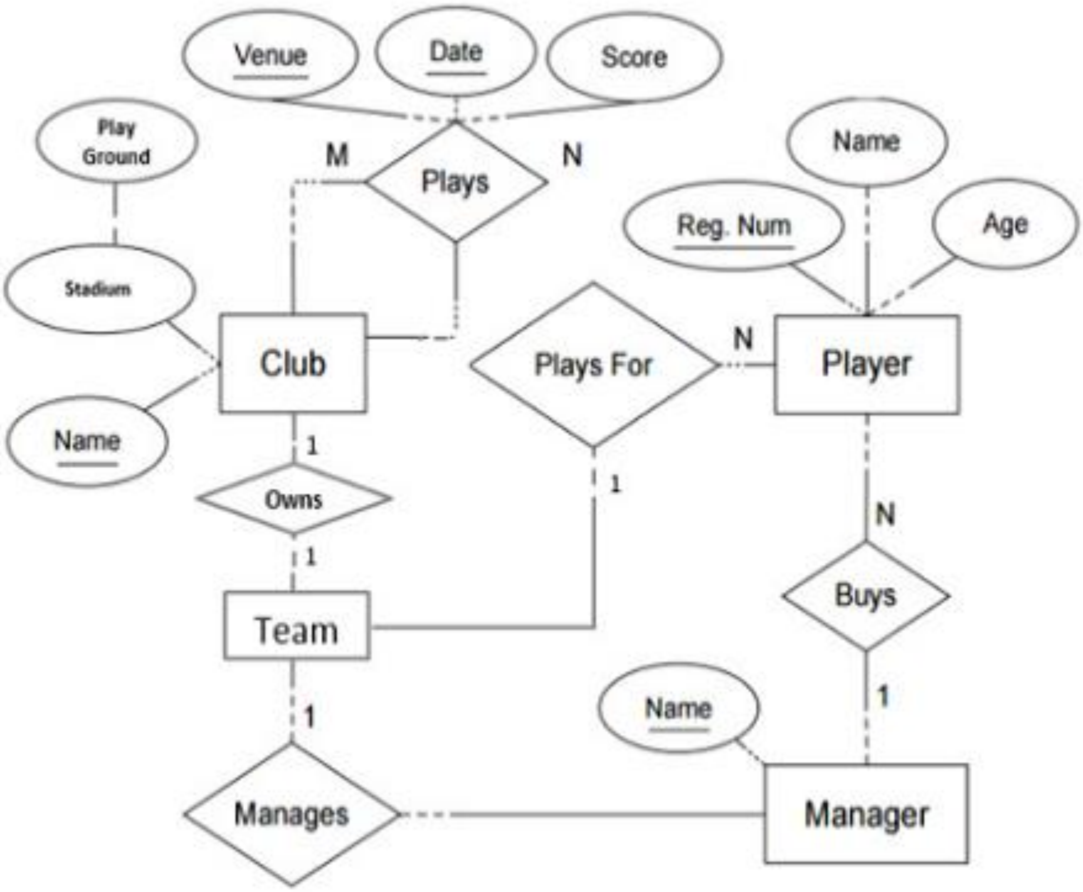
- The app aims to give the user a more detailed indication of which teams are in and out of form. The app is bright and well laid out, with helpful instructions pointing out where the hidden and locked parts of the app are, and what specific icons mean.

- The app will perform analysis on team and player, which represents their ability based on past performances, using its own statistical modelling and algorithm. This system is an interesting approach and at a glance, gives the user a good indication of the two team's form leading up to a game. The app also includes other features, including game insights and head-to-head record; with standout stats.

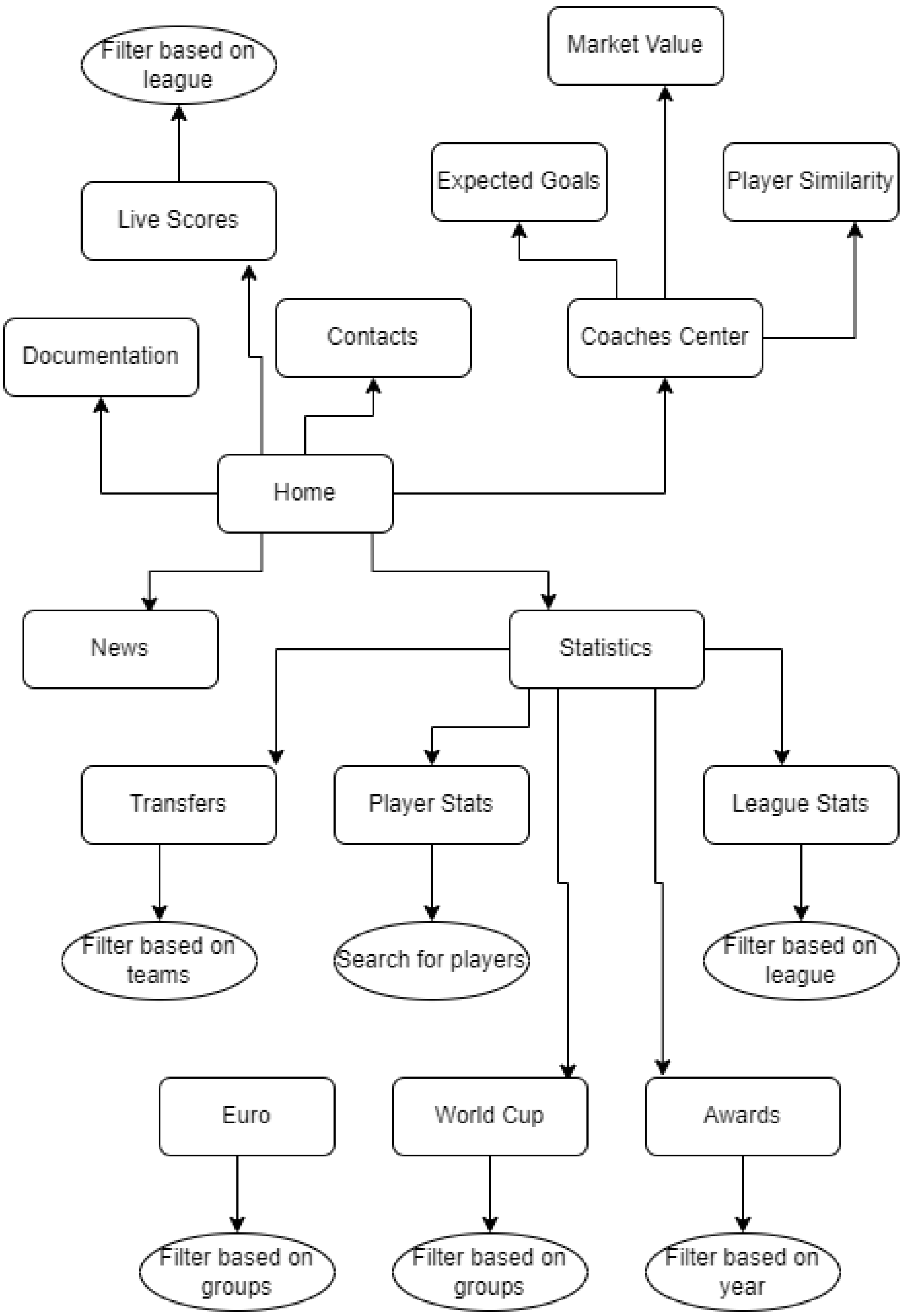
- The app will be free and open-source

4. ARCHITECHTURE AND DESIGN

4.1 Diagrams:



ER DIAGRAM

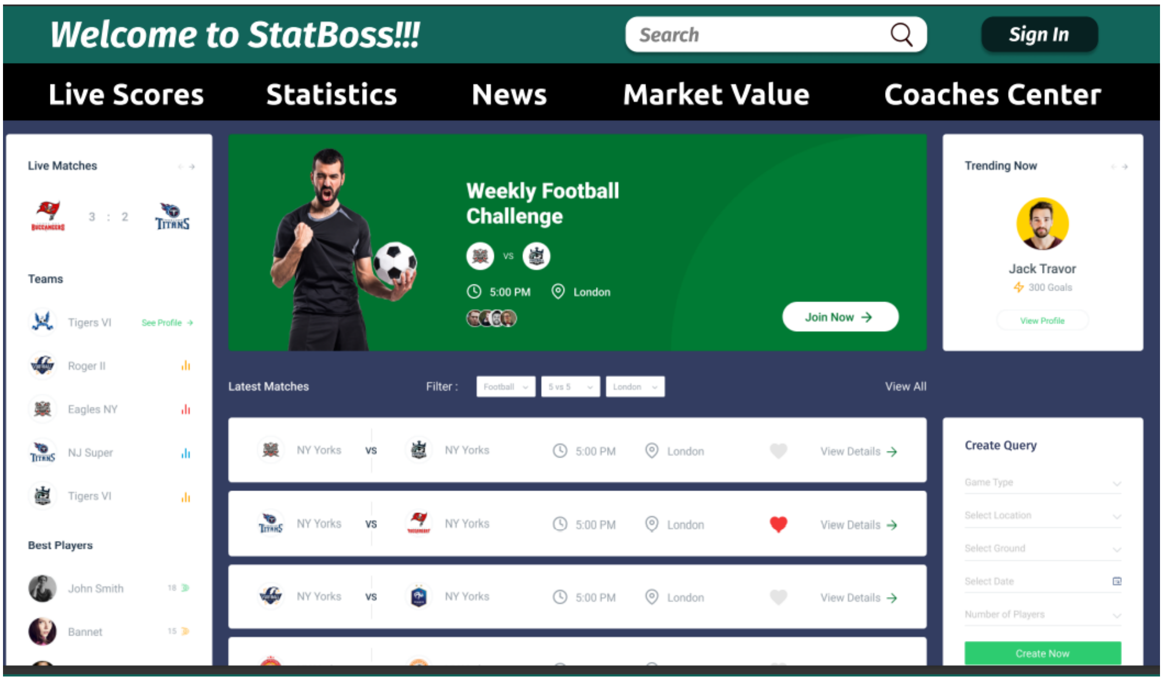


MODULE DIAGRAM

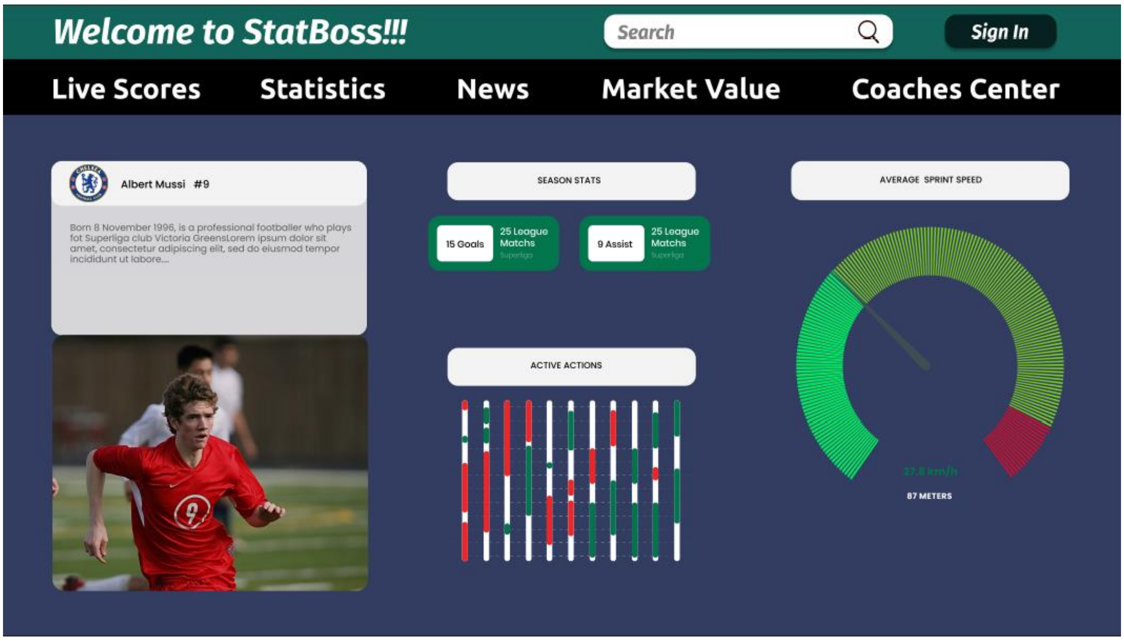
4.2 Wireframes:

The screenshots below are a mockup of the proposed web application.

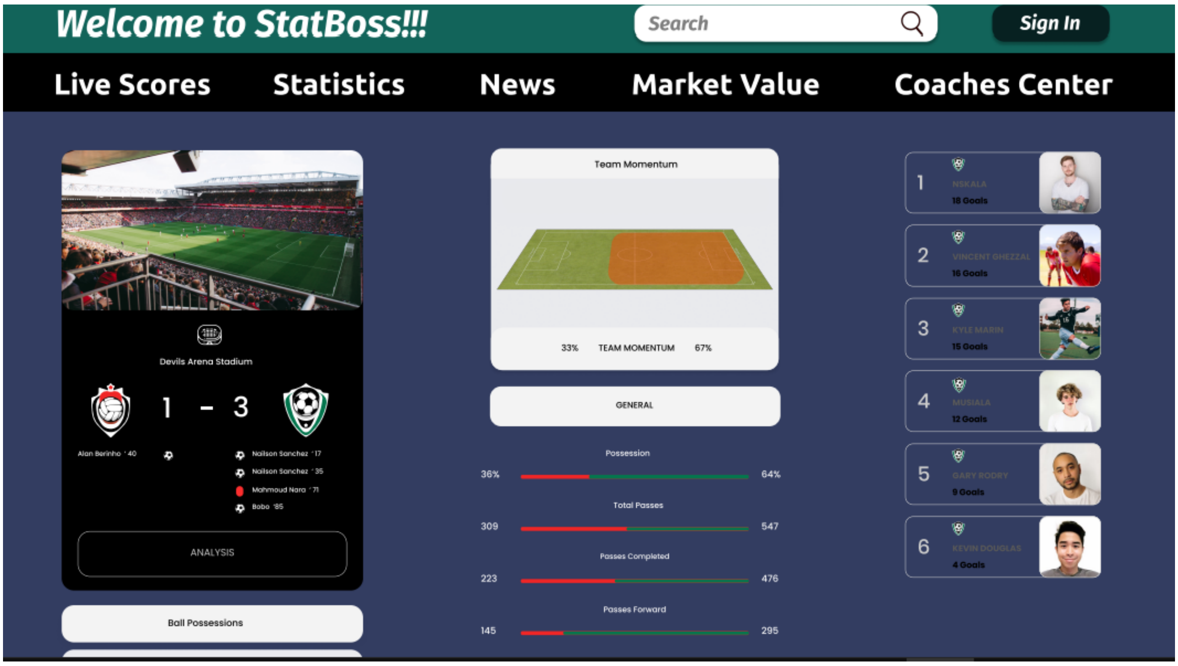
HOME MODULE



LIVE SCORES MODULE



PLAYER STATS MODULE



TEAM STATS MODULE

5. DEVELOPED ML MODELS

5.1 Market Value:

Background

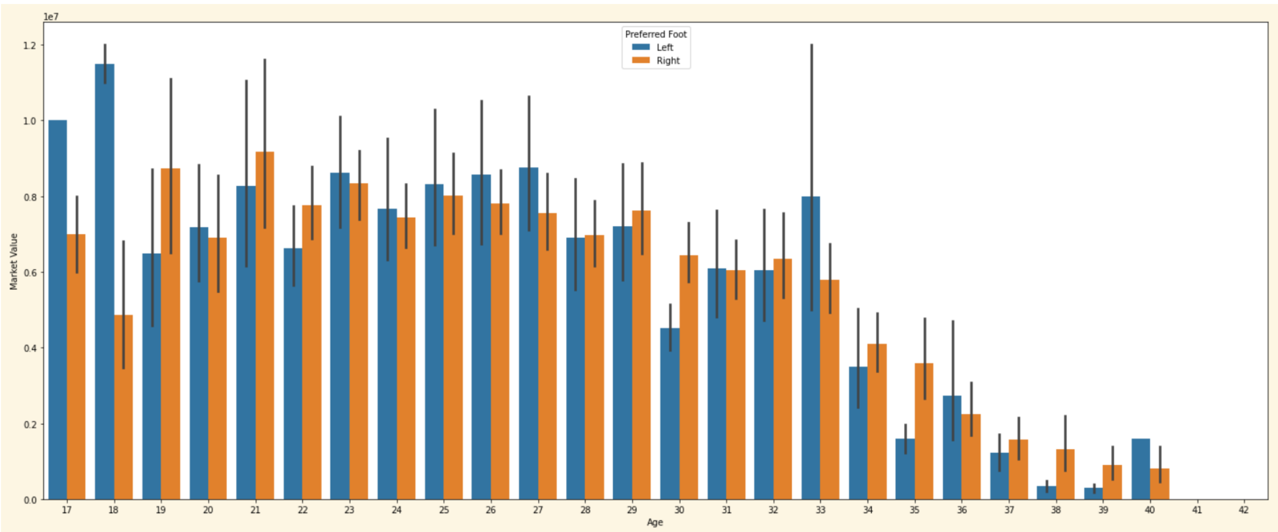
An important part of this project would be predicting the market value of players based on various statistics
Typically, a younger player will cost more than an older one. If the player is already closer to retirement, a football player has fewer years left to keep performing at the highest level thus depreciating their value whereas a young player is perceived to be a long-term investment. Many clubs make a lot of their income by developing young players and then selling them to a bigger club for a higher amount. Thus, the outcomes of predicting the market value is valuable in the industry. Two other equally important factors are the players skills and the position on the field they prefer to play in. Attacking players, and those who score a lot of goals or give a lot of assists, tend to be more expensive. Higher-skilled players will have larger transfer fees. Research shows that players who get higher average ratings tend to have higher transfer fees. Thus, we made use of a dataset which encapsulates the points mentioned above.
As a part of the analysis to be done, an extract is presented here. The dataset used consisted of roughly 6000 players and information on their skill level based on the videogame “FIFA”. It consisted of statistics such as their interpretive statistics such as overall skill level, strength, salary (speculated) as well as real life metrics like their weight, age, height, dominant foot.

EDA and Preprocessing:

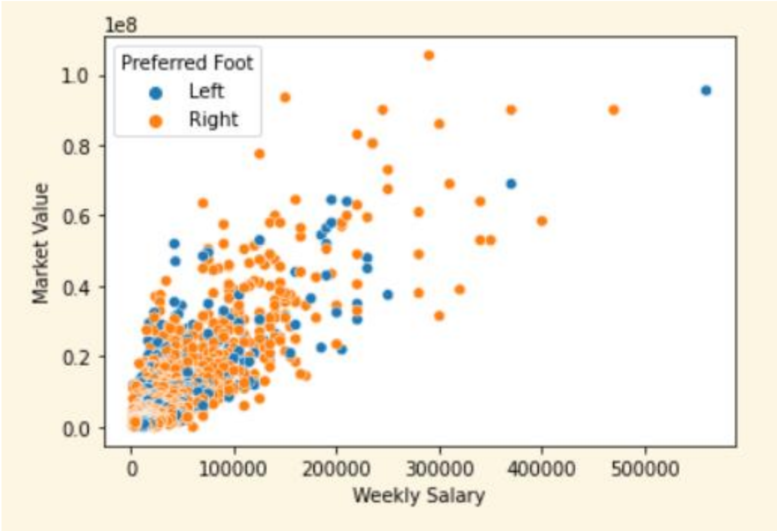
Given below is a snippet of the dataset.

Player	Overall Score	Potential Score	Market Value	Weekly Salary	Height	Weight	Age	Preferred Foot	Ball Skills	Defence	Mental	Passing	Physical
0	Lionel Messi	94	94	95500000	560000	170	72	33	Left	96.5	32	77.83	90.6
1	Cristiano Ronaldo	93	93	58500000	400000	187	83	35	Right	90.5	28	76.67	81.3
2	Neymar Jr	92	92	105500000	290000	175	68	28	Right	95.5	31.33	75	8
3	Virgil van Dijk	91	92	90000000	245000	193	92	29	Right	73.5	90.67	77.33	71.6
4	Jan Oblak	91	93	77500000	125000	188	87	27	Right	21	19	47.5	3
5	Kevin De Bruyne	91	91	90000000	370000	181	70	29	Right	89	61	83.83	92.3
6	Robert Lewandov	91	91	86000000	300000	184	80	32	Right	86.5	31.67	78.17	74.3
7	Eden Hazard	91	91	90000000	470000	175	74	29	Right	94.5	27.67	75.33	84.3
8	Alisson	90	91	64500000	160000	191	91	27	Right	28.5	16.67	45	35.3
9	Mohamed Salah	90	90	80500000	235000	175	71	28	Left	89.5	40.67	79.17	79.3
10	Sadio Mané	90	90	80500000	235000	175	69	28	Right	90	40.67	77.33	77.3
11	Marc-André ter S	90	93	67500000	250000	187	85	28	Right	25.5	16	50.33	47.3
12	Sergio Agüero	90	90	69000000	310000	173	70	32	Right	88.5	27.67	74.83	72.3
13	Kylian Mbappé	89	95	93500000	150000	178	73	21	Right	89.5	33.33	74.17	76.6
14	N'Golo Kanté	89	89	59500000	230000	168	70	29	Right	79.5	88.67	85.17	78.3
15	Harry Kane	89	91	83000000	220000	188	89	27	Right	82	43.33	78.67	79.6
16	Antoine Griezma	89	89	69000000	370000	176	73	29	Left	89	54	79.83	83.3
17	Toni Kroos	89	89	64000000	340000	183	76	30	Right	84.5	67.67	79.5	90.6
18	Luka Modrić	89	89	39000000	320000	172	66	35	Right	89.5	71	81.5	88.3
19	Luis Suárez	89	89	53000000	350000	182	86	33	Right	84.5	46.67	80.17	79.6
20	Manuel Neuer	89	89	36000000	160000	193	92	34	Right	38	12.67	49.5	4
21	Sergio Ramos	89	89	31500000	300000	184	82	34	Right	74	87.33	81.17	76.3
22	Ederson	88	91	54500000	185000	188	86	27	Left	31.5	17.33	52	45.3
23	Raheem Sterling	88	90	73000000	250000	170	69	25	Right	89	49	72.17	7
24	Roberto Firmino	88	88	63000000	220000	181	76	28	Right	90.5	55.33	80.83	78.6
25	Kalidou Koulibaly	88	90	60000000	140000	187	89	29	Right	70	88.67	71.33	54.6
26	Casemiro	88	89	61000000	280000	185	84	28	Right	73.5	86.33	83.5	75.3

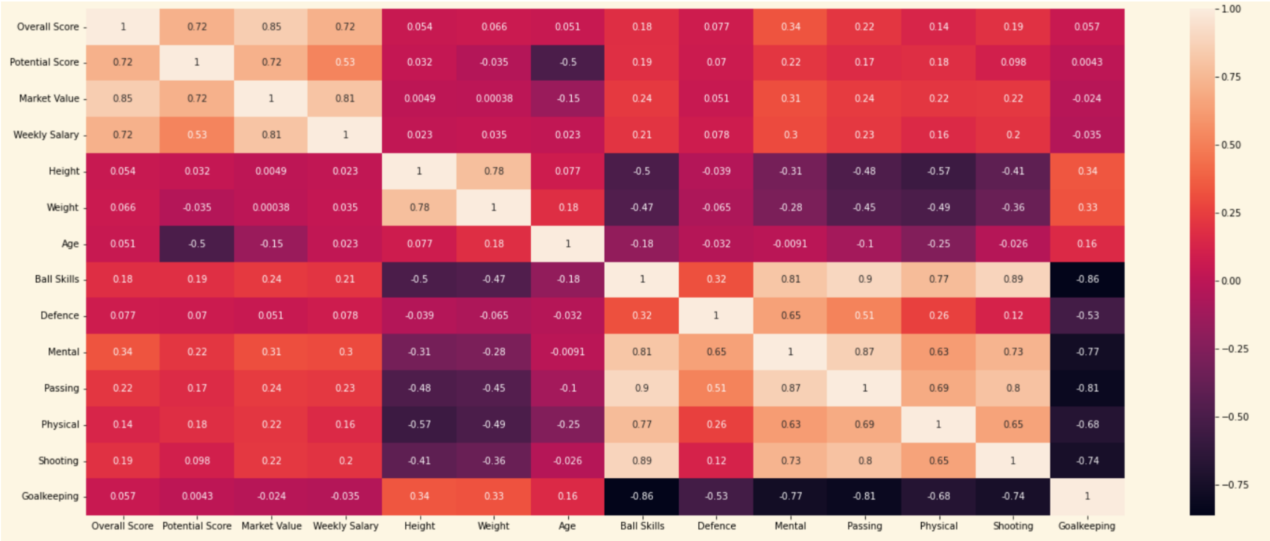
Below are the results of working on a dataset analyzing the details of various players -



Distribution of the dominant foot of players sorted by age.



A scatterplot showing the distribution of the dominant foot of players sorted by their weekly salary

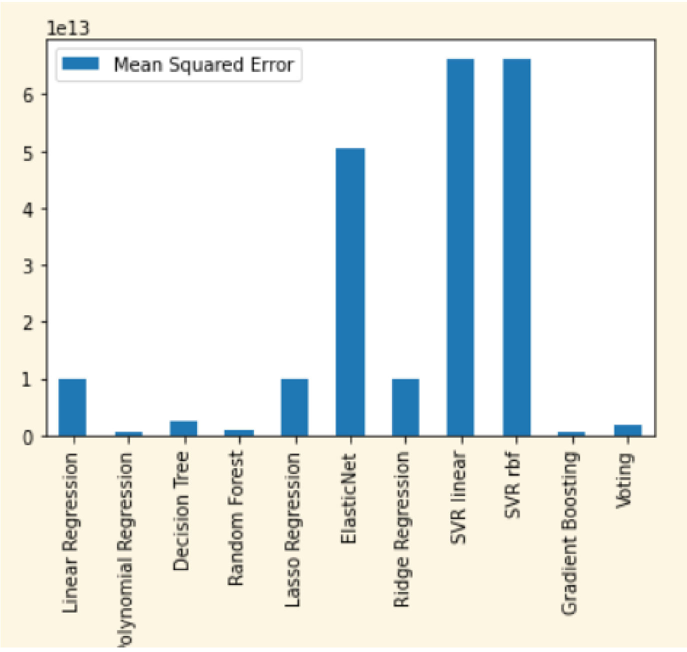


Correlation heatmap

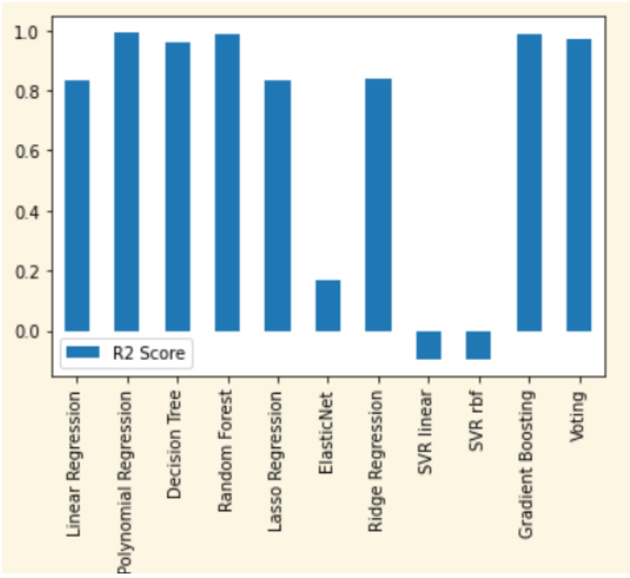
The correlation heatmap above provides an interesting insight into the significance of the factors that play a role in the market value of a player.

Algorithms and Results:

We applied several regression algorithms were applied on the dataset, primarily in order to predict the market values of players. To do so, we dropped attributes which did not contribute towards this goal such as 'Height', 'Weight', 'Preferred Foot', 'Market Value' and 'Player'. Preprocessing steps were also applied like scaling through the MinMaxScaler. The metrics for measuring the efficacy were the mean squared error and R2 score. The results are visualized below:



Graph comparing the mean squared error of the algorithms. We find that Polynomial regression, random forest and gradient boosting have the least mean squared error values. To get the best algorithm, R2 score comparison is also conducted.



Graph comparing the R2 scores. We find that all the 3 best algorithms have the same R2 score. These 3 algorithms could be implemented.

5.2 Player Similarity:

Description of concept: Seeking out similar alternatives to certain players is a key element of scouting. Often times, a targeted player may be under contract at another club which makes acquiring them nearly impossible. In these cases, finding alternate players with a high similarity to these targets would be highly beneficial.

Dataset: For this implementation, we used a dataset containing information of over 15000 players collected from the FIFA dataset. It involves attributes of players in detail.

Below are some sample screenshots of the dataset

	short_name	player_positions	overall	potential	wage_eur	age	club_contract_valid_until	nationality_name	preferred_foot	weak_foot	skill_moves	release_clause_eur	pace
0	L. Messi	RW, ST, CF	93	93	320000	34	2023	Argentina	Left	4	4	144300000	85
1	R. Lewandowski	ST	92	92	270000	32	2023	Poland	Right	4	4	197200000	78
2	Cristiano Ronaldo	ST, LW	91	91	270000	36	2023	Portugal	Right	4	5	83300000	87
3	Neymar Jr	LW, CAM	91	91	270000	29	2025	Brazil	Right	5	5	238700000	91
4	K. De Bruyne	CM, CAM	91	91	350000	30	2025	Belgium	Right	5	4	232200000	76
...	...	...	...	...	...	...	...	...	...	...	...	...	...
19234	Song Defu	CDM	47	52	1000	22	2021	China PR	Right	3	2	114000	58
19235	C. Porter	CM	47	59	500	19	2021	Republic of Ireland	Right	3	2	193000	59
19236	N. Logue	CM	47	55	500	21	2021	Republic of Ireland	Right	3	2	175000	60
19237	L. Rudden	ST	47	60	500	19	2021	Republic of Ireland	Right	3	2	239000	68
19238	E. Lalchhanchhuaha	CAM	47	60	500	19	2025	India	Right	3	2	217000	68

19239 rows × 54 columns

skill_dribbling	skill_curve	skill_fk_accuracy	skill_long_passing	skill_ball_control	movement_acceleration	movement_sprint_speed
96	93	94	91	96	91	80
85	79	85	70	88	77	79
88	81	84	77	88	85	88
95	88	87	81	95	93	89
88	85	83	93	91	76	76
...	...	...	...	...	...	...
45	33	38	48	49	56	60
41	53	31	50	42	60	58
47	37	37	49	49	60	60
42	36	34	33	45	69	67
48	38	32	49	38	70	67

To preprocess this dataset, we drop columns not affecting results in any way. We are most concerned about attributes like the preferred foot of the player, their age and various other metrics. We drop columns like nationality, release clause value, etc. Apart from their ability, information like their current market value is also crucial for detecting similar players.

After filtering out the dataset, we normalize values and make sure they're all in the int64 type.

```
overall                      int64
potential                    int64
wage_eur                     int64
age                          int64
weak_foot                    int64
skill_moves                  int64
release_clause_eur           int64
pace                         int64
shooting                     int64
passing                      int64
dribbling                    int64
defending                    int64
physic                       int64
attacking_crossing           int64
attacking_finishing          int64
attacking_heading_accuracy   int64
attacking_short_passing      int64
attacking_volleys            int64
skill_dribbling              int64
skill_curve                  int64
skill_fk_accuracy            int64
skill_long_passing           int64
skill_ball_control           int64
movement_acceleration        int64
movement_sprint_speed        int64
movement_agility             int64
movement_reactions           int64
movement_balance             int64
```

Player similarity can be used to find players of a similar profile, especially if the player's age and rating are far apart. In can be used to find players who are significantly younger but have a similar player build as the player being compared to. It can also be used to dismiss the fact that just because two players are of a similar caliber, they must be of a similar type and either of them could replace the other or play in their position.

In Data Mining, similarity measure refers to distance with dimensions representing features of the data object, in a dataset. If this distance is less, there will be a high degree of similarity, but when the distance is large, there will be a low degree of similarity.
Some of the popular similarity measures are –Euclidean Distance, Manhattan Distance, Jaccard Similarity, Minkowski Distance, Cosine Similarity.

Cosine similarity: Cosine similarity is a metric, helpful in determining, how similar the data objects are irrespective of their size. Cos(x, y) = x . y / ||x|| * ||y||.
Adjusted cosine similarity measure is a modified form of vector-based similarity where we take into the fact that different users have different ratings schemes;

in other words, some users might rate items highly in general, and others might give items lower ratings as a preference.
The cosine similarity is beneficial because even if the two similar data objects are far apart by the Euclidean distance because of the size, they could still have a smaller angle between them. Smaller the angle, higher the similarity.
When plotted on a multi-dimensional space, the cosine similarity captures the orientation (the angle) of the data objects and not the magnitude.

Results:

Below, we have compared the best-rated player on FIFA 22 with one of the worst players. To the casual observer, they might conclude that there should be no similarity. However, in reality, despite the disparity in rating, there is a resemblance between them, given that the worse player is young has room to improve their attributes. They also play in similar positions on the pitch, making an 80% similarity rating justified.

```
from scipy import spatial

dataSetI = new_data.iloc[0:1,:]
dataSetII = new_data.iloc[19234:19235,:]
dataSetIII= new_data.iloc[1:2,:]
cosine_result_1 = 1 - spatial.distance.cosine(dataSetI, dataSetII)
cosine_result_1

0.8001408605559942
```

Below is another example of cosine similarity. The players compared are Neymar Jr. and Noa Lang. Although Neymar Jr. is rated considerably higher, they are almost identical to each other. This is due to the fact that both play the same wide positions on the field. Also, their stats relative to their rating are extremely similar in terms of pace, dribbling ability, passing and more. This is why there is a high cosine similarity between them. In the real world, data like this could cause scouts to earmark Lang as a possible replacement or successor to Neymar Jr., thus helping out in club’s transfer dealings.

Comparing Neymar Jr. and Noa Lang

```
[ ] data1SetI = data.iloc[2:3,:]
    data1SetII = data.iloc[837:838,:]
    data1SetIII= data.iloc[19234:19235,:]
    cosine_result_1 = 1 - spatial.distance.cosine(data1SetI, data1SetIII)
    cosine_result_1
```

```
0.9984635475822238
```

5.3 Expected Goals (xG) Analysis

For analysis of expected goals, we made use of a data-set containing shot events for various matches during the years 2012-2017. xG could give teams a better understanding of their performance, regardless of the result. This is because results are not always reflective of performance. Most publicly available football statistics are limited to aggregated data such as Goals, Shots, Fouls, Cards. When assessing performance or building predictive models, this simple aggregation, without any context, can be misleading. A team that produces 10 shots on target from long range, for example, has a lower chance of scoring than a club that produces the same number of shots from inside the box. Metrics derived from this simple count of shots, however, will equally evaluate the two teams. This data set was obtained from Kaggle and as a consequence of web scraping and integration of various data sources. There are 11 different sorts of events, as well as the primary and secondary players involved in those events, as well as numerous more statistics. The data set provides a granular view of 9,074 games, totaling 941,009 events from the biggest 5 European football (soccer) leagues: England, Spain, Germany, Italy, France from the 2011/2012 season to 2016/2017 season as of 25.01.2017. A football game generates much more events and it is very important and interesting to take into account the context in which those events were generated.

```
Confusion Matrix:
[[70773   921]
 [ 6225  2279]]
Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     71694
           1       0.71      0.27      0.39      8504

    accuracy                           0.91     80198
   macro avg       0.82      0.63      0.67     80198
weighted avg       0.90      0.91      0.89     80198
```

Results for XGBoost

Now that we have our data set, we can apply the xG model. We started by using one-hot encoding to all the shot parameters (whether a goal was scored, the type of attack, the area the shot was taken from, etc). This would help us learn the data well. We then split the data using 35\% for validation and the rest for training. XGBoost, being an ensemble technique, gave us the best results for market value predictions and was also used here. We got a very good accuracy of 91\%. Looking at the confusion matrix to our precision, recall, and F1 score were all above 90\% too. Similarly, we applied Logistic Regression and Neural Networks too with the former giving us an accuracy of a much lower 72\% but still above 90\% scores on the confusion matrix.

```
Confusion Matrix:
[[70820   874]
 [ 6265  2239]]
Report:
              precision    recall  f1-score   support

           0       0.92      0.99      0.95     71694
           1       0.72      0.26      0.39      8504

    accuracy                           0.91     80198
   macro avg       0.82      0.63      0.67     80198
weighted avg       0.90      0.91      0.89     80198
```

Results for Logistic regression

Neural networks had an accuracy of 91.10\%, slightly better than XGBoost. We saw there was a massive correlation between xG and goals which suggests our model was working.

	goals_scored	expected_goals
goals_scored	1.000000	0.976885
expected_goals	0.976885	1.000000

Having analyzed how good each technique was, we took our xG model to the player level to analyze them better. We analyzed the best finishers, players with high xG, players with high xG - goals, the best headers, left-footed players, right-footed players, and outside-of-the-box shooters. We also took our model to the passing side of the game and analyzed players with the best passes, crosses, and the unluckiest players.

	n_shots	goals_scored	expected_goals	difference
player				
lionel messi	914	205.0	147.461680	-57.538321
luis suarez	433	96.0	65.747676	-30.252324
zlatan ibrahimovic	774	153.0	123.051116	-29.948884
gonzalo higuain	552	118.0	88.149569	-29.850431
cristiano ronaldo	1190	198.0	170.513715	-27.486285
robert lewandowski	633	124.0	98.861785	-25.138215
alexandre lacazette	391	88.0	64.810171	-23.189829
alexis sanchez	445	80.0	59.443467	-20.556533
antoine griezmann	493	80.0	60.042045	-19.957955
karim benzema	434	85.0	65.082137	-19.917864
diego costa	410	93.0	73.842607	-19.157393
franck ribery	226	40.0	22.224773	-17.775227
eden hazard	332	62.0	44.373842	-17.626158
mario mandzukic	329	67.0	50.691988	-16.308012
wissam ben yedder	367	69.0	53.423314	-15.576687
marco reus	408	65.0	49.832598	-15.167402
mauro icardi	323	72.0	56.946932	-15.053068
fernando llorente	282	56.0	41.077089	-14.922911
gareth bale	301	50.0	35.579504	-14.420496
alexander meier	291	57.0	42.844460	-14.155540

A list of statistically best finishers

	n_passes	goals_scored_from_passes	xGoals_from_passes	difference	xGoals_per_pass
player2					
luis suarez	185	40.0	31.000877	-8.999123	0.167572
gareth bale	109	21.0	17.690061	-3.309939	0.162294
angel di maria	211	43.0	32.743670	-10.256330	0.155183
lionel messi	350	68.0	52.762740	-15.237260	0.150751
neymar	179	25.0	26.439148	1.439148	0.147705
karim bellarabi	109	18.0	16.092844	-1.907156	0.147641
henrikh mkhitaryan	148	21.0	21.404461	0.404461	0.144625
marco verratti	113	21.0	16.328394	-4.671606	0.144499
pierreemerick aubameyang	130	21.0	18.465793	-2.534207	0.142045
cesc fabregas	264	53.0	37.000010	-15.999991	0.140152
patrick herrmann	120	20.0	16.772269	-3.227731	0.139769
jimmy briand	120	22.0	16.732658	-5.267343	0.139439
zlatan ibrahimovic	270	36.0	37.132533	1.132533	0.137528
javier pastore	185	28.0	25.120704	-2.879296	0.135788
nolito	179	18.0	24.195035	6.195035	0.135168
cristiano ronaldo	222	45.0	29.930214	-15.069786	0.134821
alvaro negredo	109	9.0	14.667663	5.667663	0.134566
arjen robben	159	20.0	20.938858	0.938858	0.131691
roberto soldado	101	19.0	13.297210	-5.702790	0.131656
alexis sanchez	248	35.0	32.259163	-2.740837	0.130077

A list of statistically best passers

6. DEVELOPED APPLICATION

As previously mentioned, this study aims to create an application that uses ML and AI to predict team and player statistics as well as attributes. In the previous section, the machine learning models for the same were built. In this section, we have developed the application and integrated the models in it along with various other functionalities. It is a completely responsive web application built using ReactJS and Flask.

6.1 Front-End:

The front-end was created using react. This interactive react app is displayed to the end-user. The app includes 5 modules including home, live scores, news, stats, and coaches center.

Live scores from matches across the world are displayed to the user in the live scores modules. These scores can be filtered by leagues. The latest football news is displayed under the news module and various live league and player stats are shown in the stats module. These modules use the data fetched/received from the server through API calls. The coaches center module shows the results from the prediction models developed on the server using machine learning.

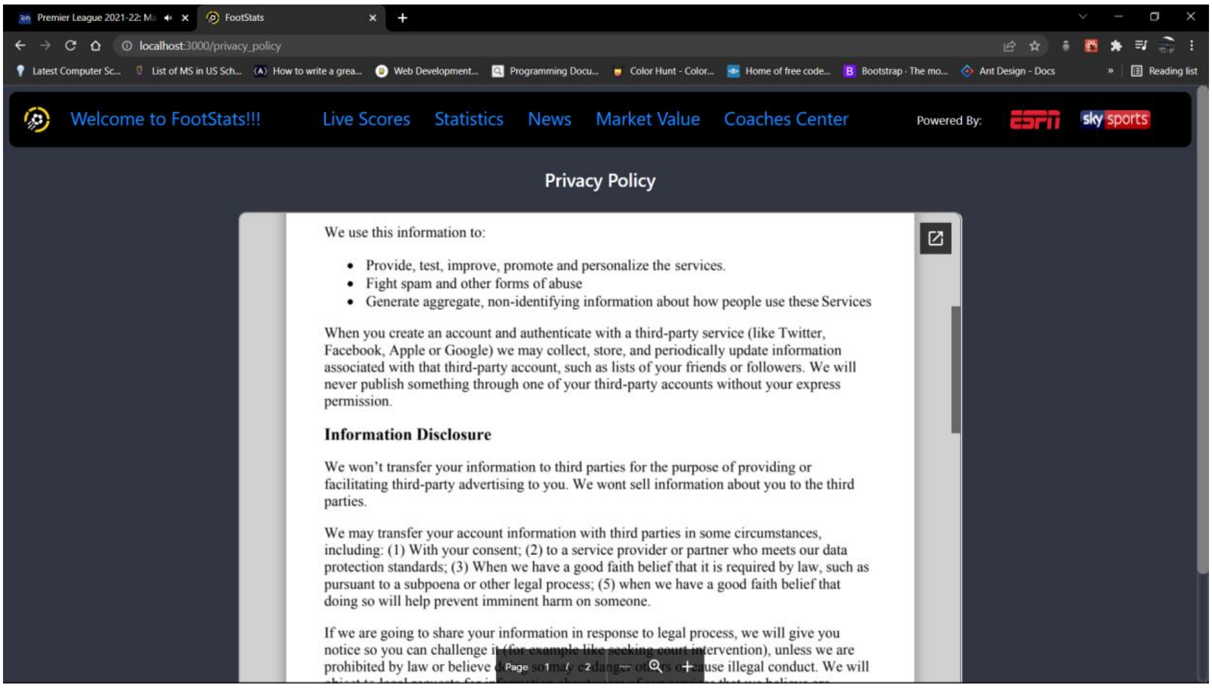
The home module contains three sections - the news section, the scores section, and the transfers section. The news section provides the latest football news from the news page, the scores section provides the latest football scores from the live scores page and the transfers section provide the latest football transfers from the statistics page. All this information is presented in the form of cards - which is the basic component of this application.

The news module shows the latest news from the world of football including results, pundit opinions, discussions, transfer rumours, etc. The live scores module shows the latest fixtures and live scores from all the football matches around the world. These matches or cards can be filtered by league. Almost all leagues recognized by FIFA are available in this module. The cards used in this module are bigger and contain more information as compared to the ones used in the home module.

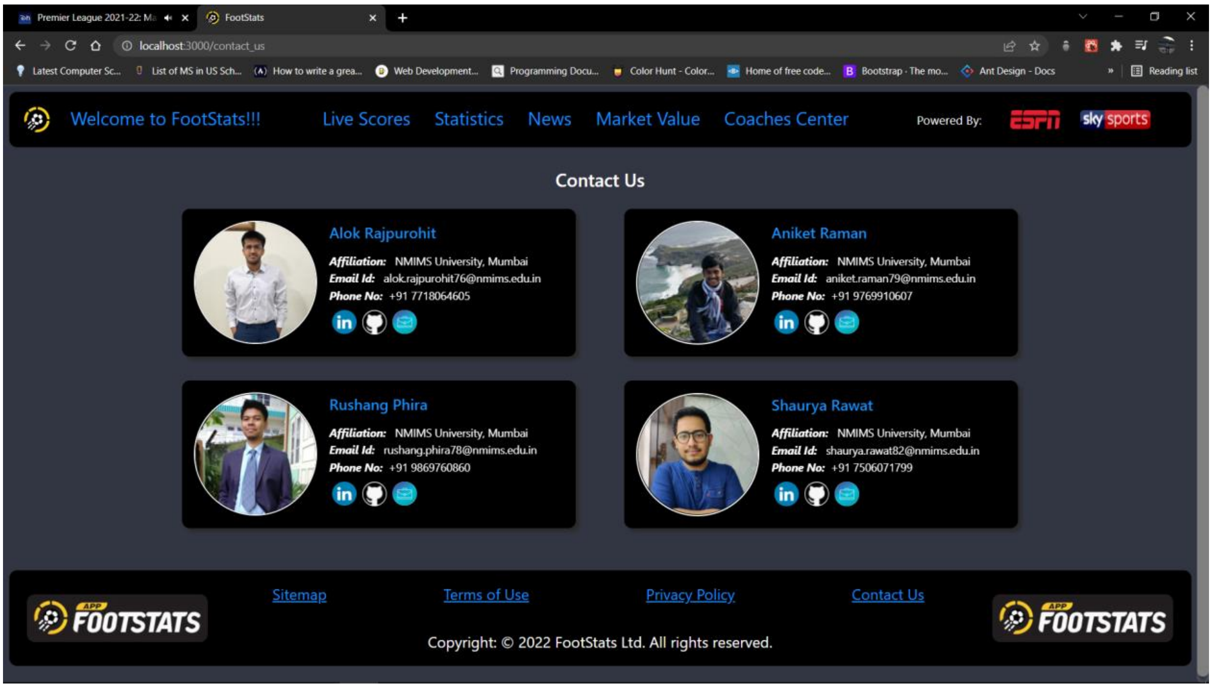
The stats module contains 5 different smaller modules - player stats, league stats, awards, transfers, and international competitions. In the player stats module, all the statistics on every player registered with FIFA are available. These stats include name, age, dob, preferred foot, nationality details, club details, positions, ratings, traits, attributes, etc. Users can search for any player of their choice using the search functionality available in this module. More than 17,000 players are available in this module. The league stats module shows the latest and live tables from the top leagues of the world. Apart from this it also shows the top scorers and top assists from those leagues. The awards module shows all the Ballon d'Or award winners and their details and statistics. The transfers module shows all the transfers from the current season or the latest transfer window. Filter functionality is available in all the above-mentioned modules. The international competitions module shows all the statistics, details, and journeys of the two most followed international tournaments - the latest FIFA World Cup and the latest UEFA Euro.

The coaches center module shows the results and stats from the 3 machine learning models developed in the previous section. These include predicting any player's market value, predicting the similarity between 2 players, and analyzing/predicting the expected goals of a match. The models themselves as well as their working is also shown in this module.

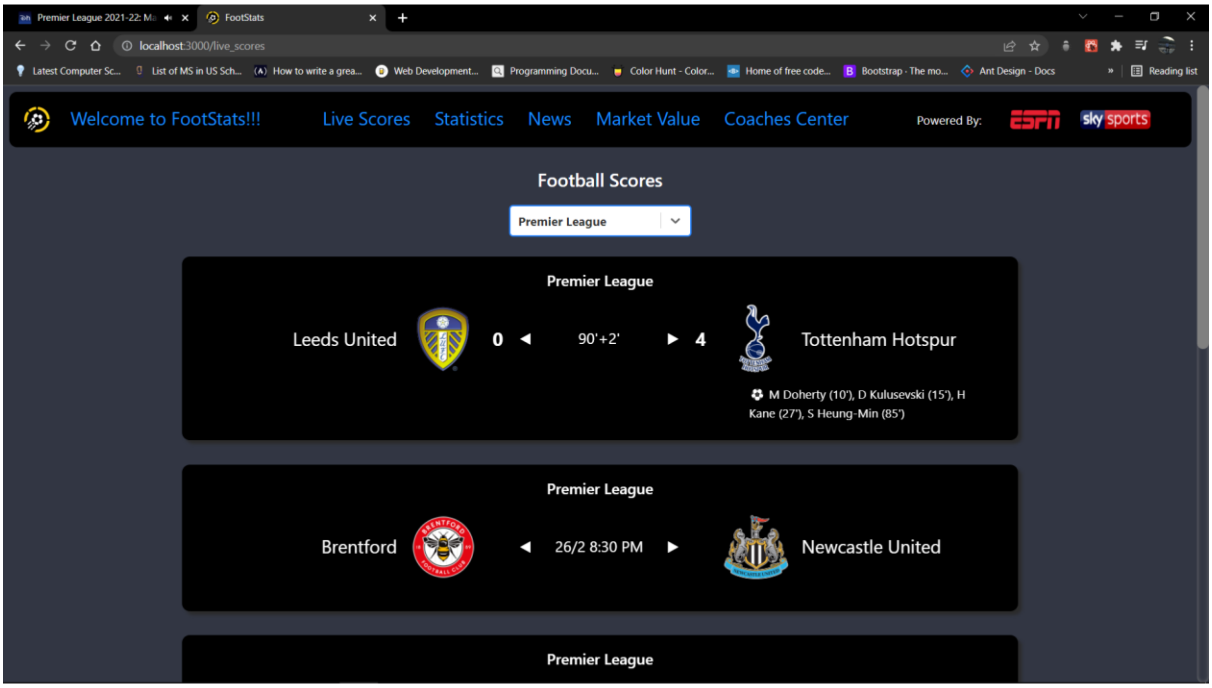
Home Page



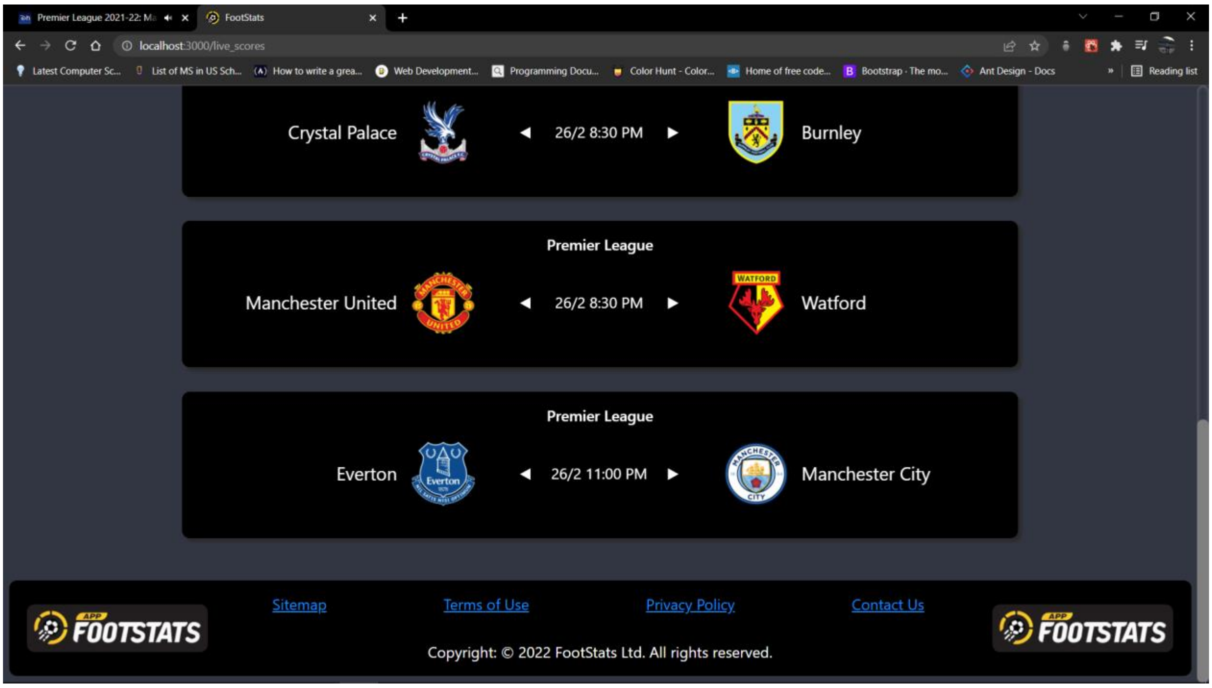
Documents Page



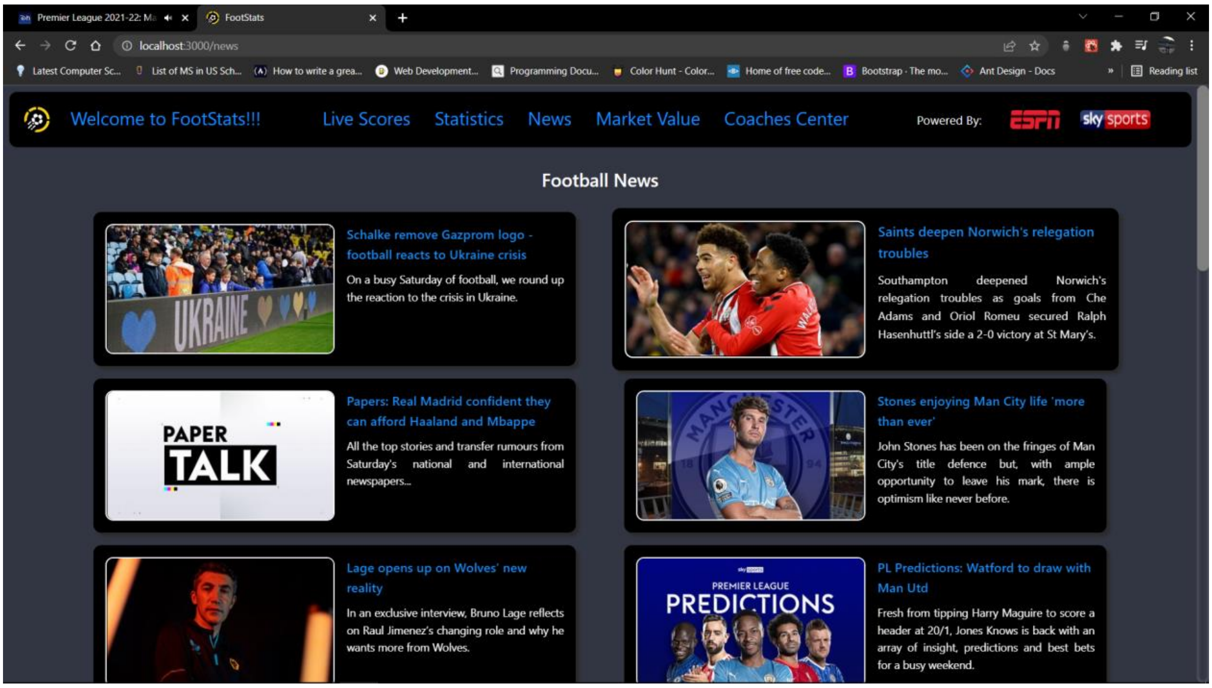
Contact Us Page



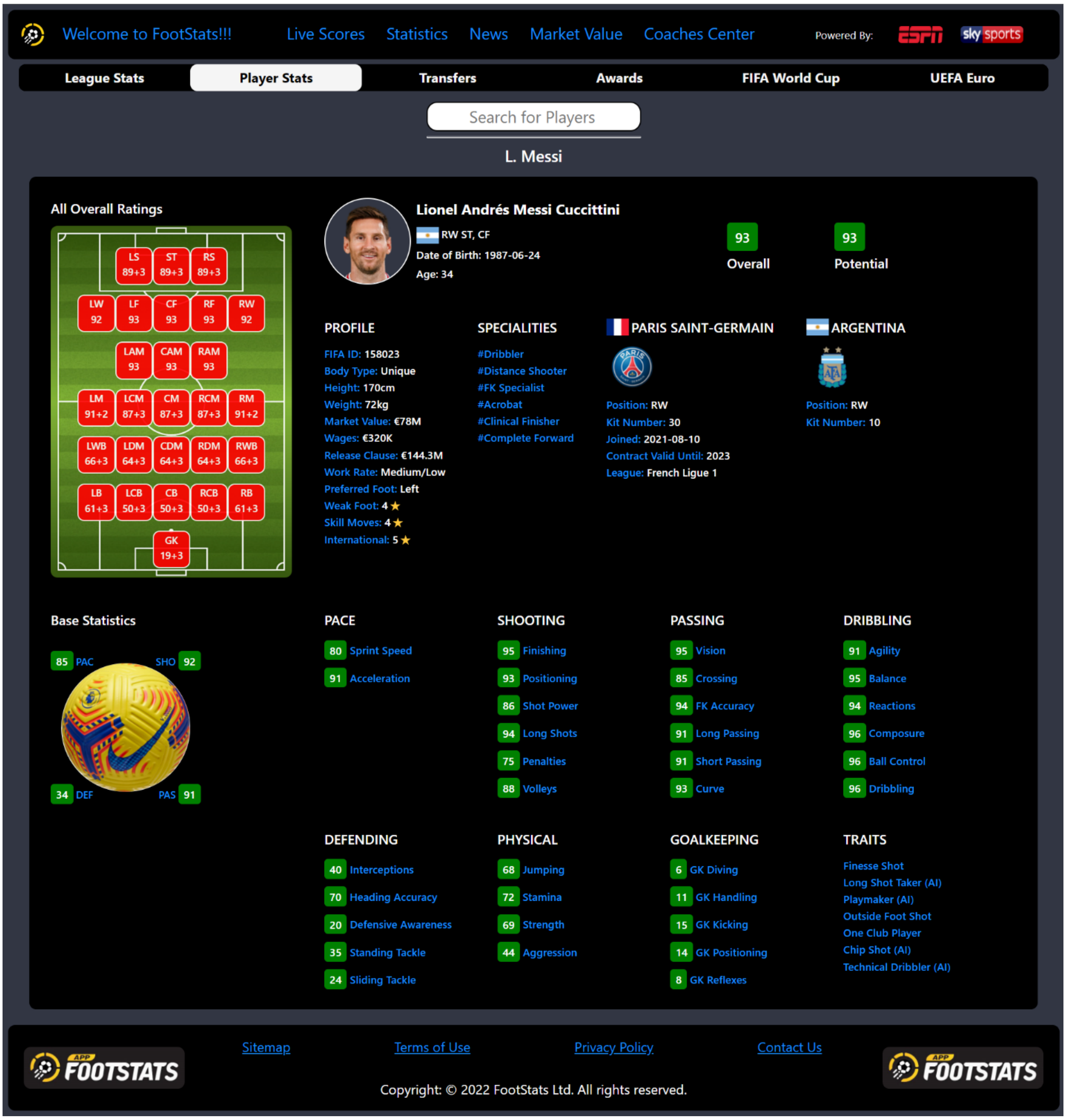
Live Scores Page top



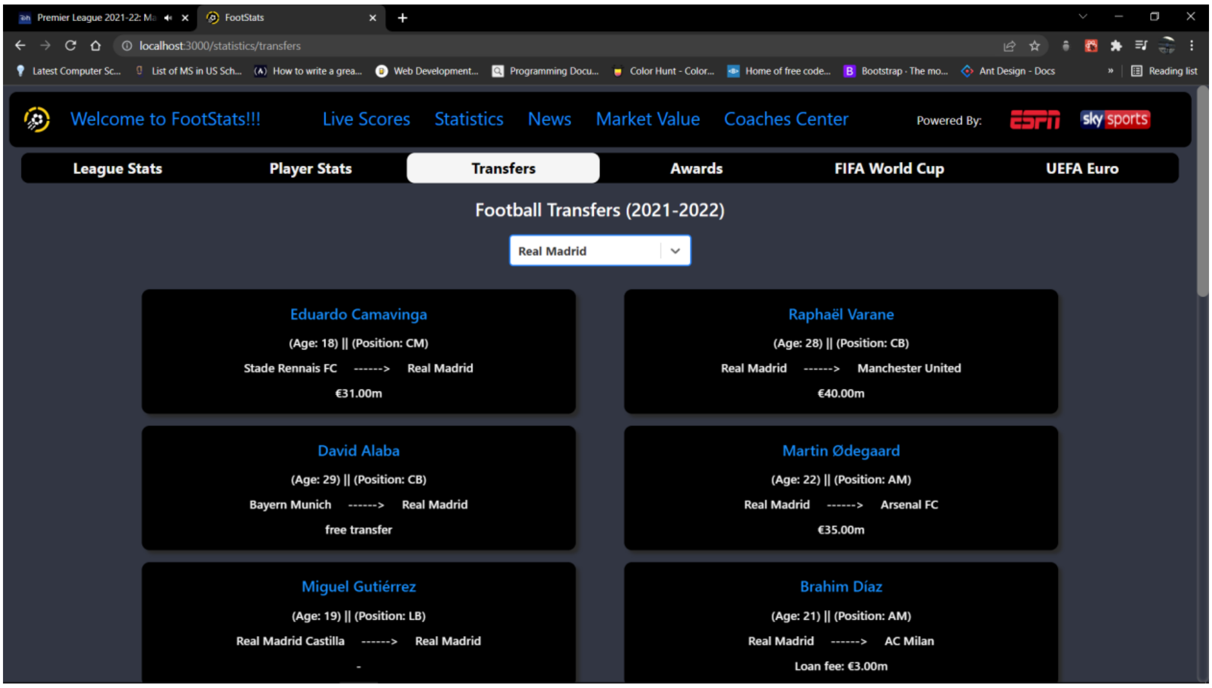
Live Scores Page bottom


FootStats
localhost:3000/news
Welcome to FootStats!!!
Live Scores
Statistics
News
Market Value
Coaches Center
Powered By:
Football News
Schalke remove Gazprom logo - football reacts to Ukraine crisis
On a busy Saturday of football, we round up the reaction to the crisis in Ukraine.
Saints deepen Norwich's relegation troubles
Southampton deepened Norwich's relegation troubles as goals from Che Adams and Oriol Romeu secured Ralph Hasenhuttl's side a 2-0 victory at St Mary's.
Papers: Real Madrid confident they can afford Haaland and Mbappe
All the top stories and transfer rumours from Saturday's national and international newspapers...
Stones enjoying Man City life 'more than ever'
John Stones has been on the fringes of Man City's title defence but, with ample opportunity to leave his mark, there is optimism like never before.
Lage opens up on Wolves' new reality
In an exclusive interview, Bruno Lage reflects on Raul Jimenez's changing role and why he wants more from Wolves.
PL Predictions: Watford to draw with Man Utd
Fresh from tipping Harry Maguire to score a header at 20/1, Jones Knows is back with an array of insight, predictions and best bets for a busy weekend.

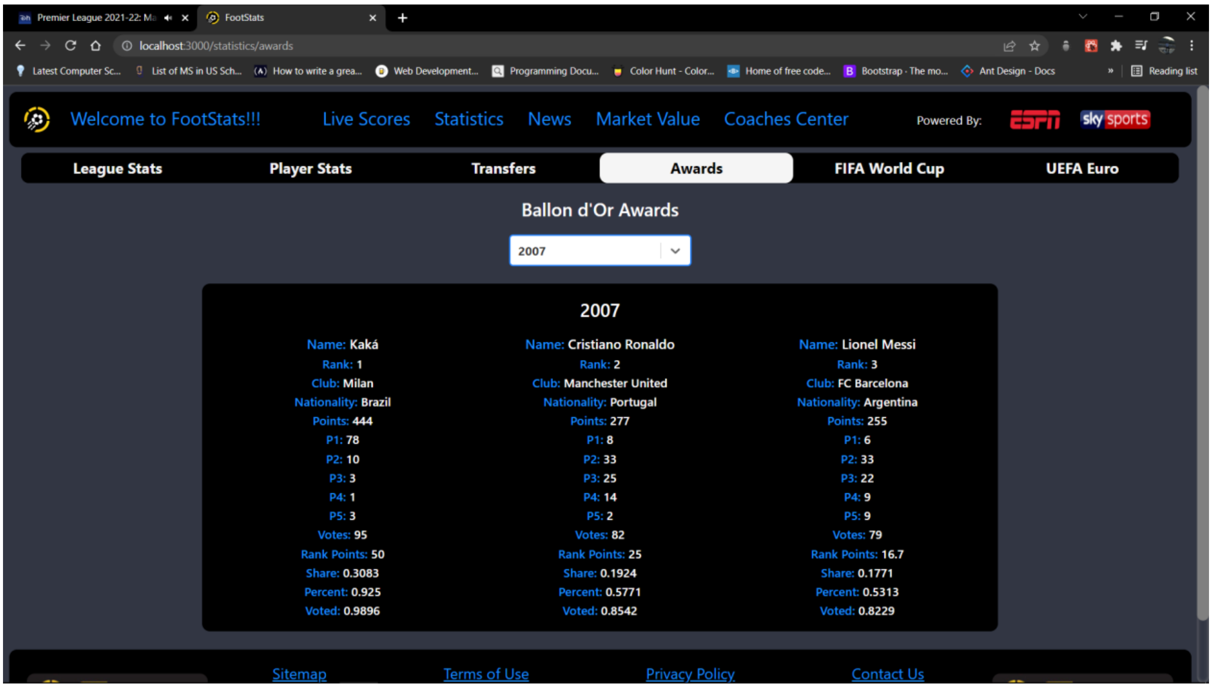
News Page



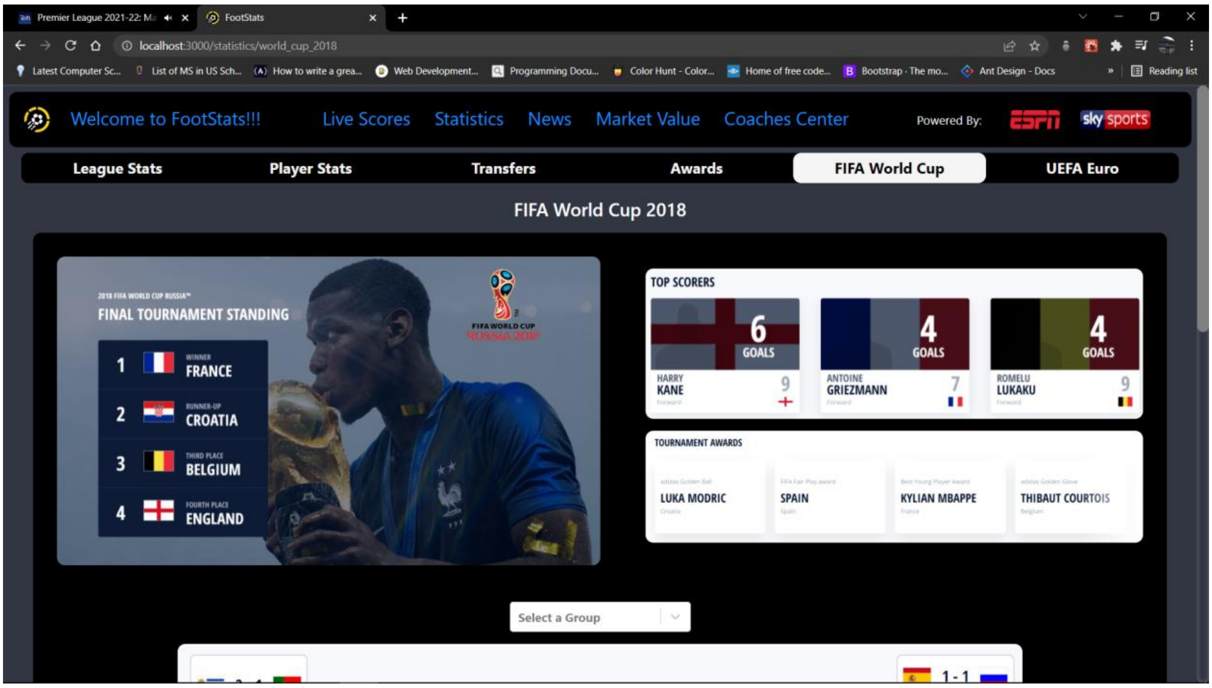
Player Stats Page



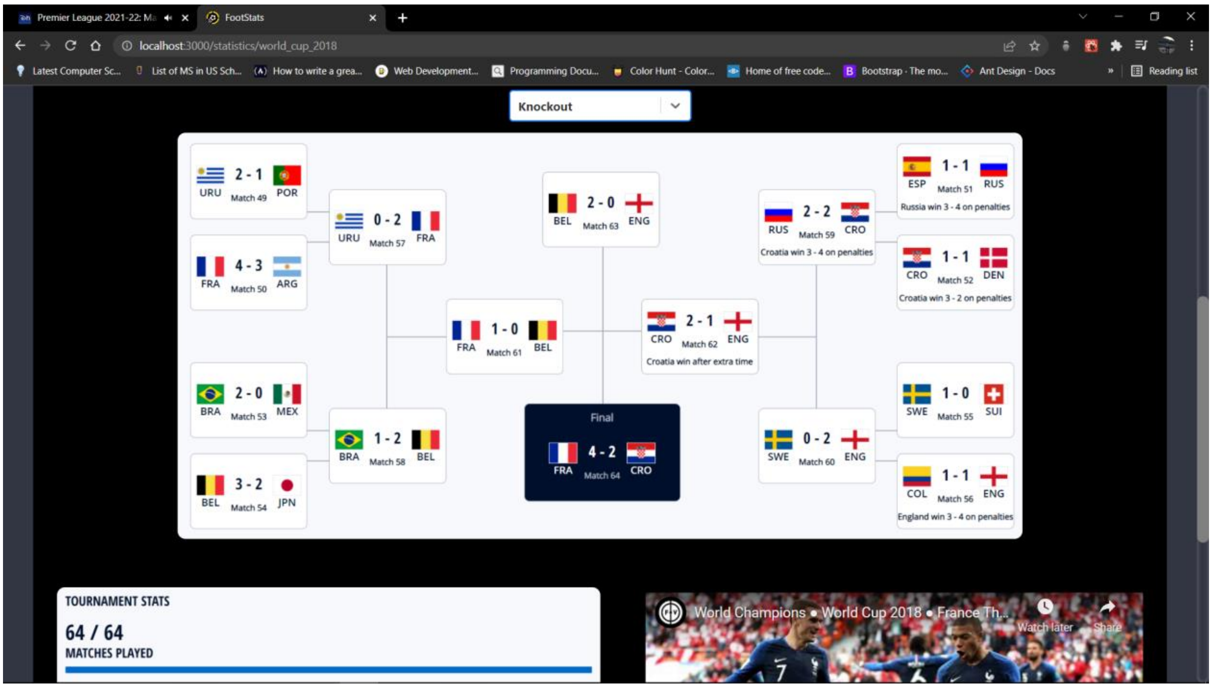
Transfers Page



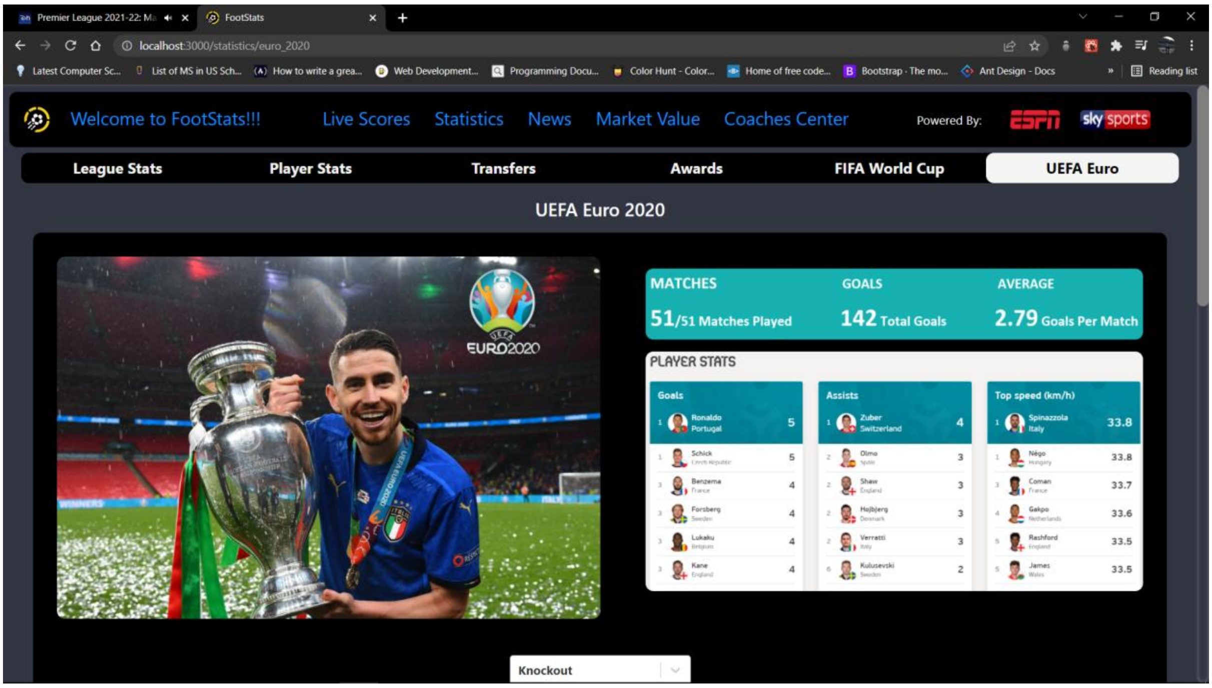
Awards Page



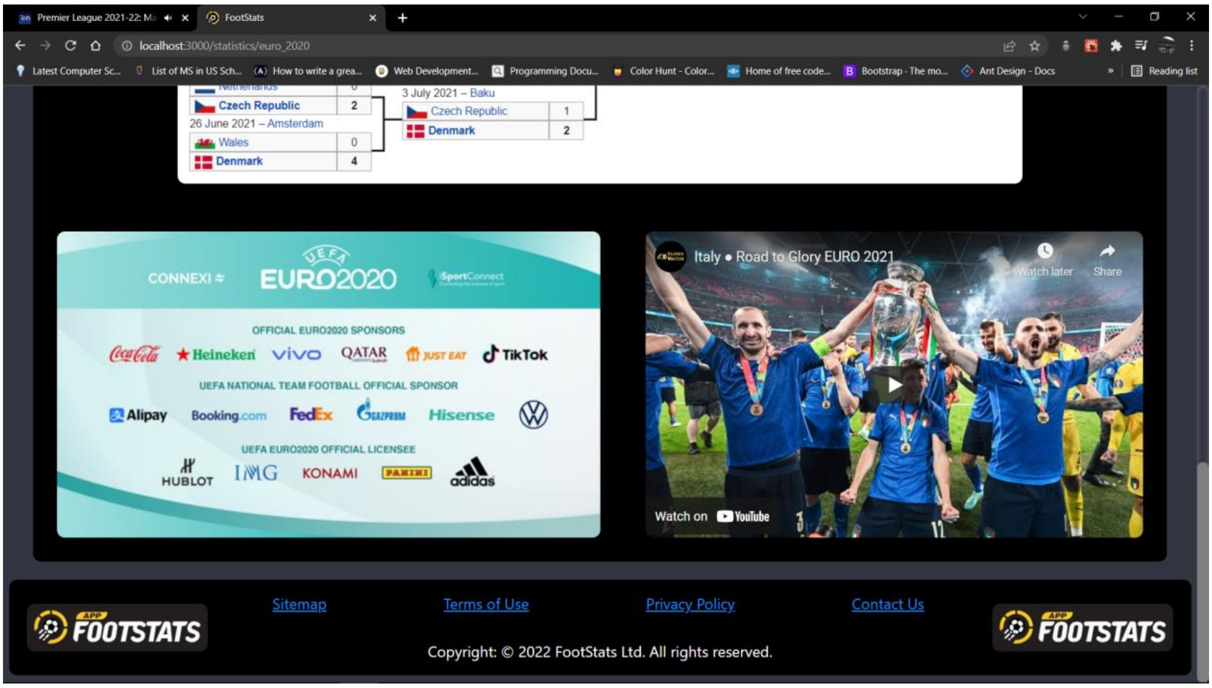
World Cup Page top



World Cup Page bottom



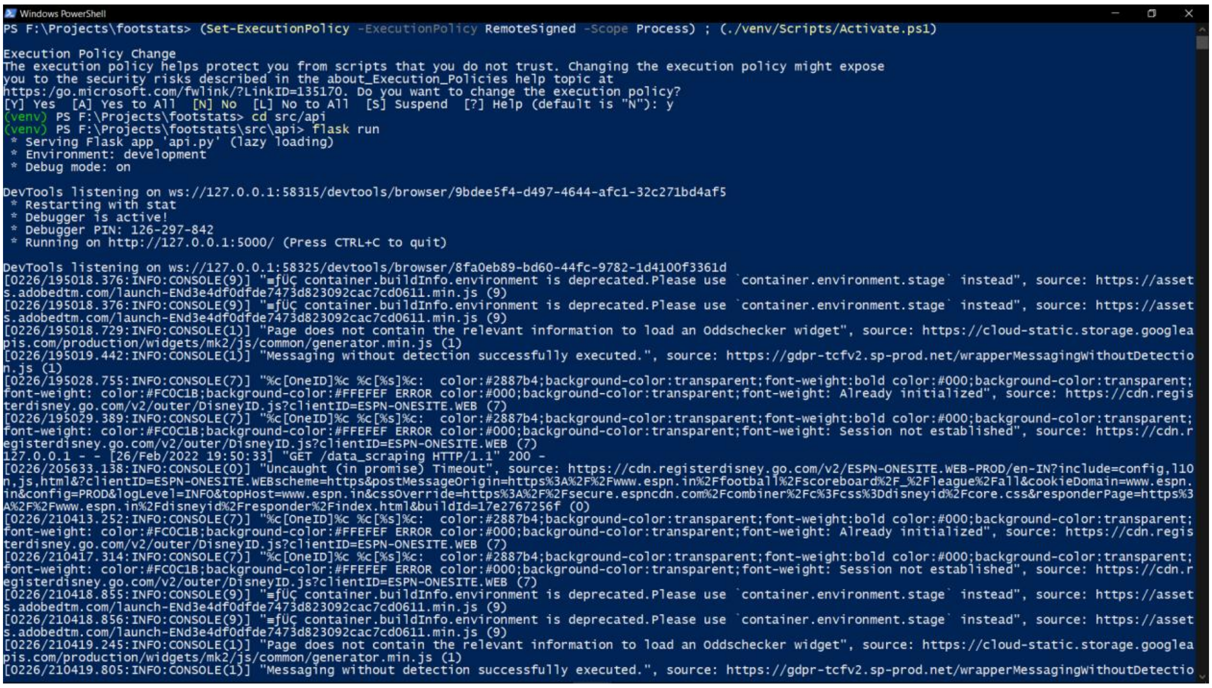
Euro Page top



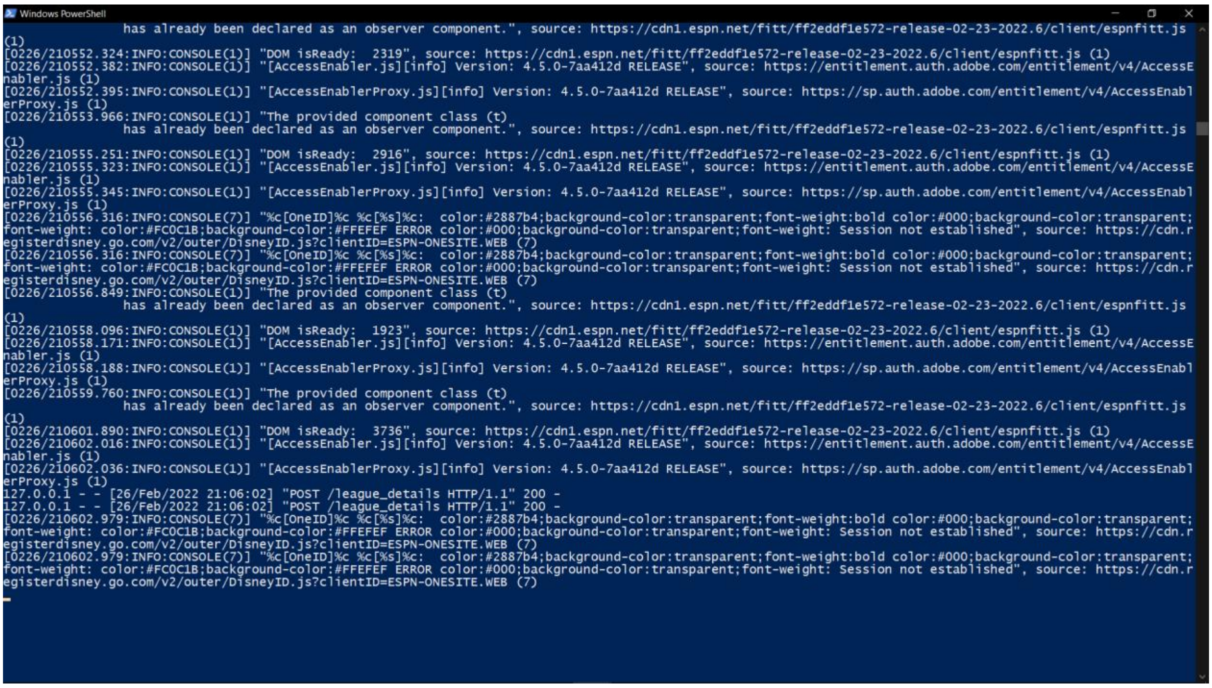
Euro Page bottom

6.2 Back-End:

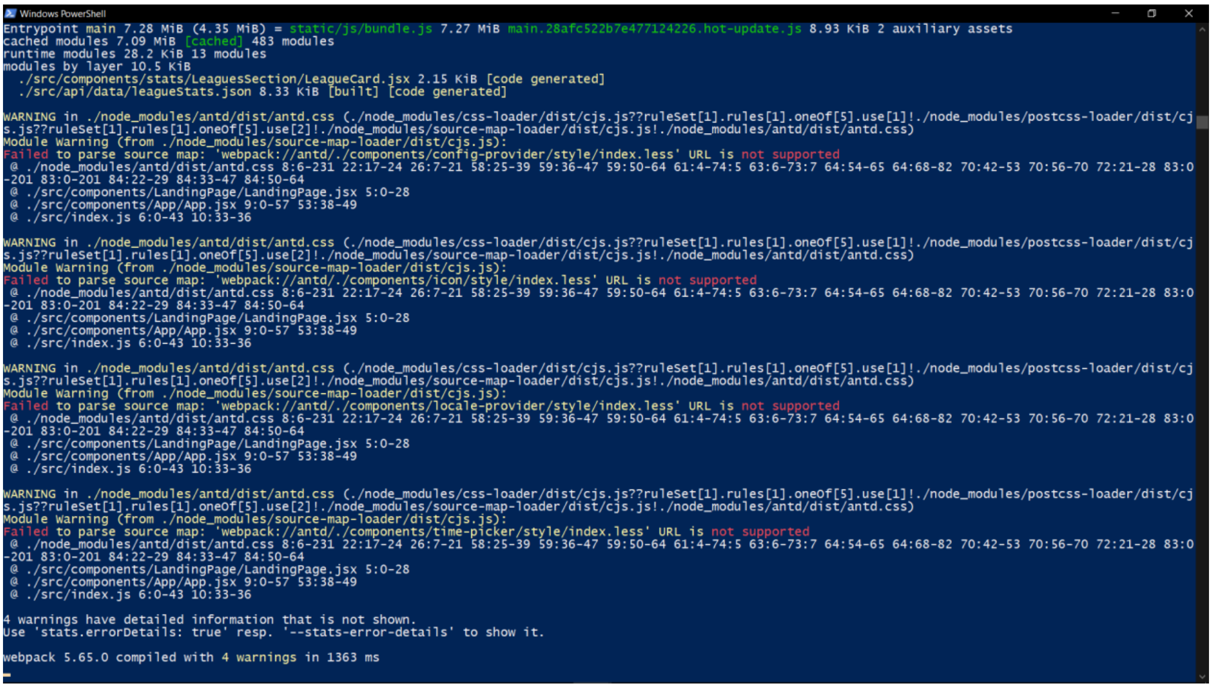
The server for the app is created using flask. Data scraping is performed on the server, after which we use our own APIs to create/fetch JSON data sets. Using these data sets we have created the machine learning models to predict player similarity, market value, expected goals, etc using python and its various packages. We have deployed these models on the server. Every time the application loads up or any page within it is refreshed, the server is also refreshed as it is connected to the front-end. This provides the live functionality to the news, scores, league stats, and various other features within the application. This is because every time the server is refreshed, the API will be called, web scraping will be performed and the JSON files containing the latest data will be created or loaded.



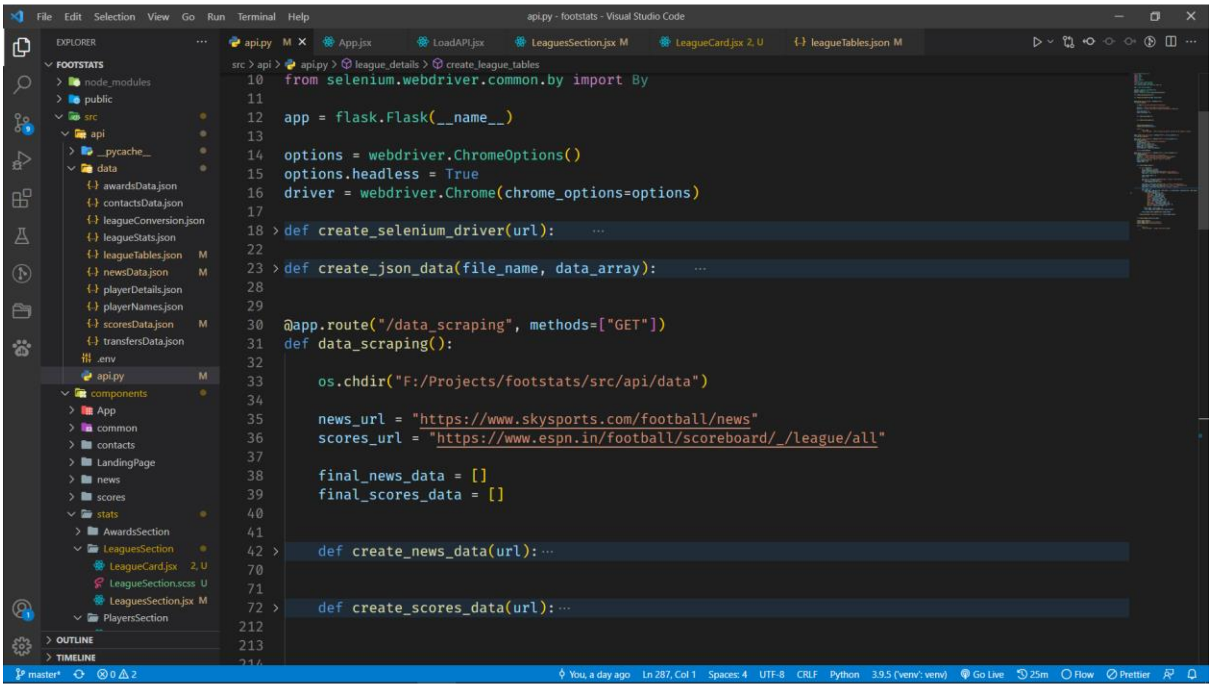
Flask server (1)



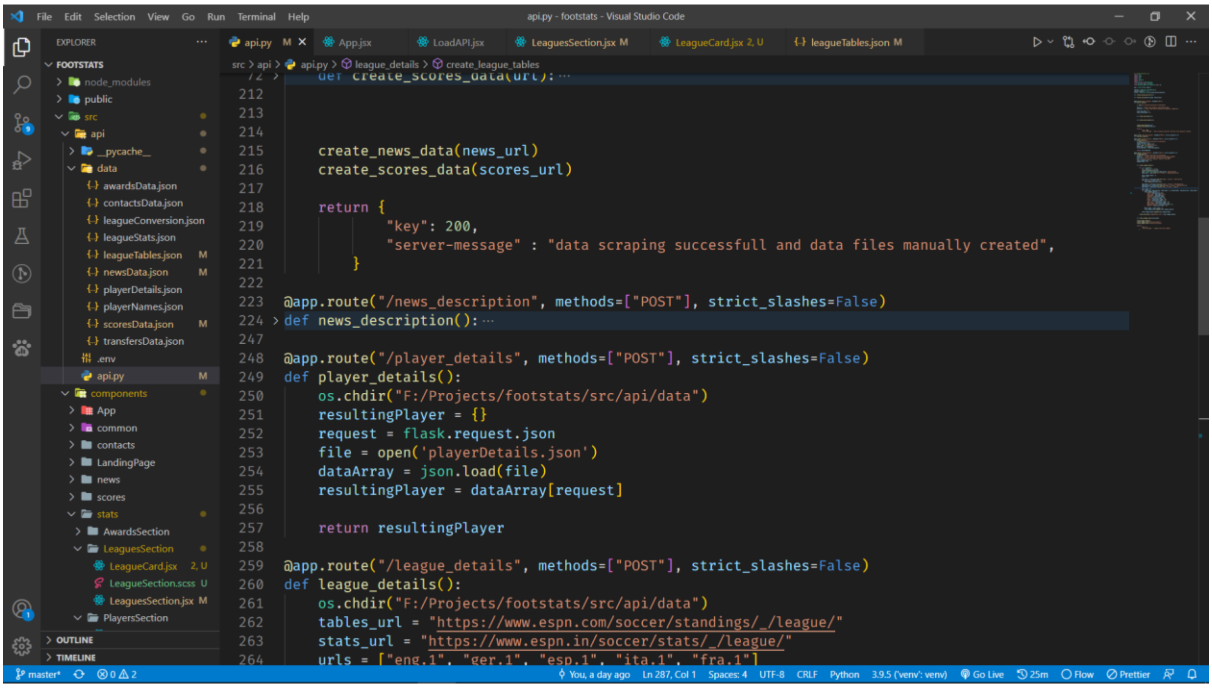
Flask server (2)



Node server



API creation and control (Flask code) (1)

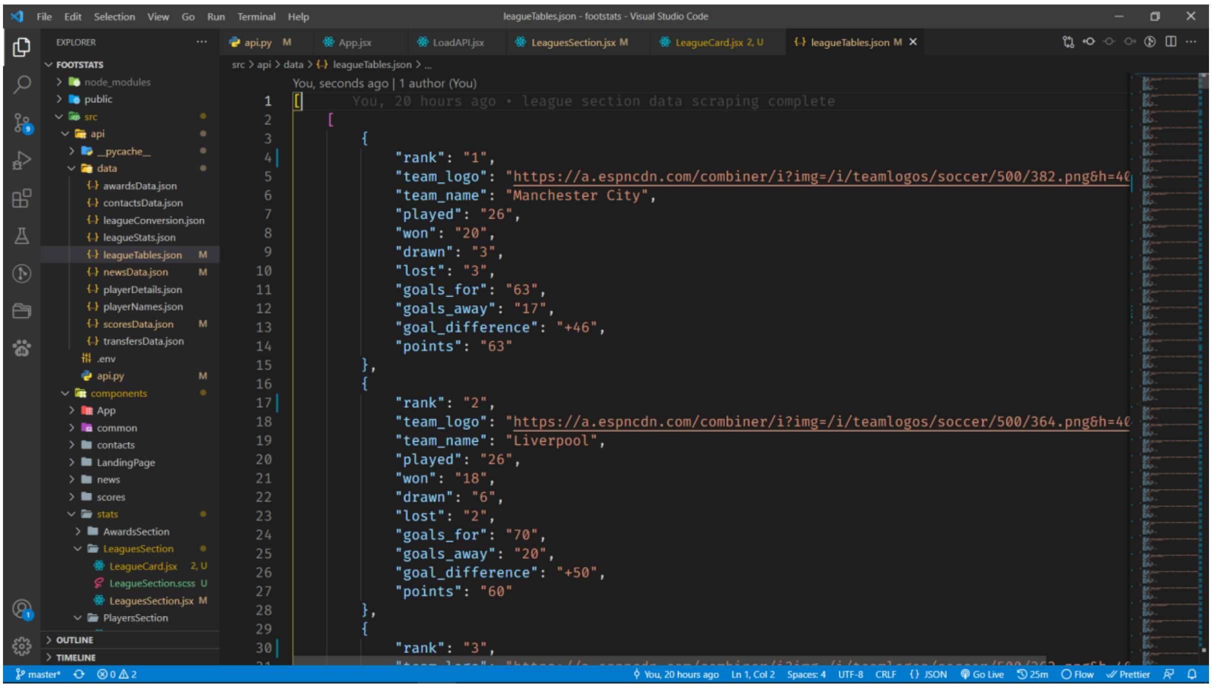


API creation and control (Flask code) (2)

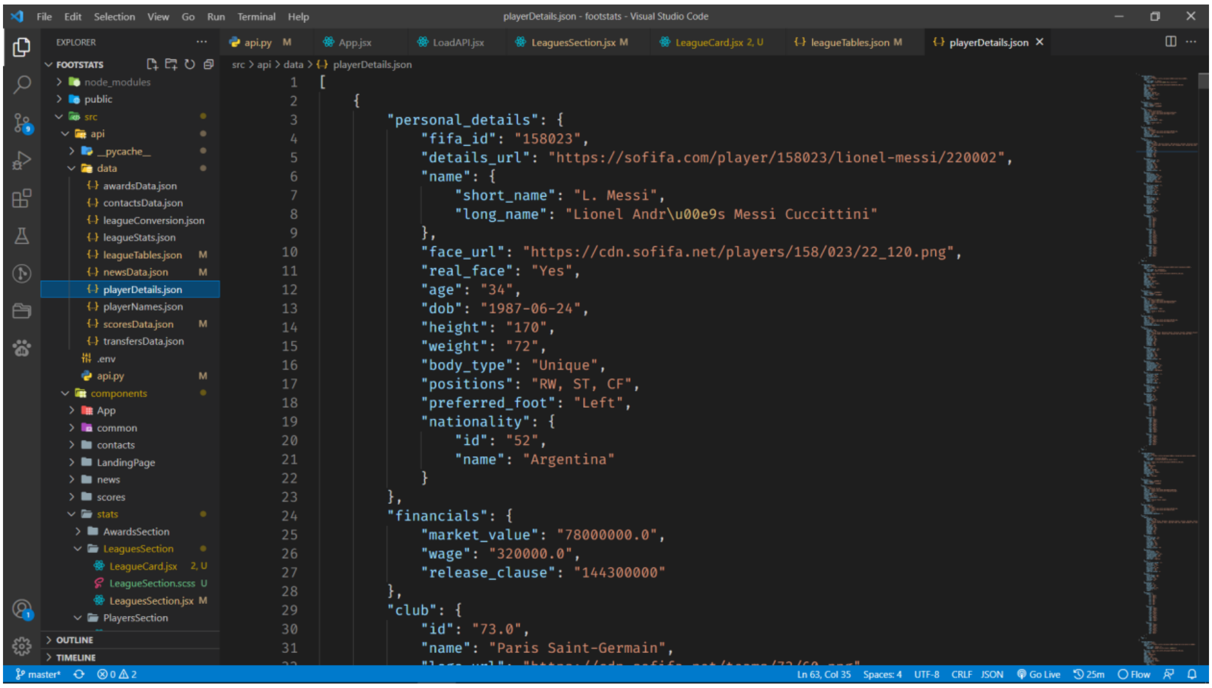
7.3 Database:

The database for this software consists of data sets and JSON files. There are a total of 10 such data files within the database out of which some are manually created by web scraping every time the application is loaded while the rest are predefined. These data files include:

1) newsData - to store data about the latest football news.
2) scoresData - to store data about the latest football scores.
3) leagueTables - to store data about the league table of the top leagues around the world.
4) leagueStats - to store the information about the top scorers and top assists of the top leagues of the world.
5) leagueConversion - to store the key-name pairs of all the leagues around the world.
6) PlayerNames - to store the names of all the players registered with FIFA.
7) PlayerDetails - to store all the statistics about all the players registered with FIFA.
8) awardsData - to store all the data about the Ballon d'Or award from the start of the award to the current year.
9) transfersData - to store the information about all the transfers of the current season or window.
10) contactsData - to store all the contact information about the developers of this application.



Data Files and API's (1)



Data Files and API's (2)

6.4 Technologies Used:

This software contains approximately 4500 lines of code in total, and the code is modular in nature, meaning it is divided into modules or components for re-usability within the application. This reusability of code and components is very important in modern-day software development. ReactJS helps in this by providing the option of components in an application. When measured using lizard.ws, the average complexity score of all modules was 6. A simple search algorithm based on the Greedy Approach was used to retrieve data from the database on the server.

The languages used for the application include HTML, CSS, and JavaScript - front-end; Python and JSON for the back-end. The frameworks used are ReactJS, Node.js, and Flask. The libraries and packages used in the application include jupyter, sklearn, BeautifulSoup, Selenium and AntDesign.

6.5 Testing and Deployment:

White box testing was carried out following the completion of the software implementation. A large number of test cases were created for the same purpose. Approximately 80% of the test cases were successfully passed after white box testing. The remaining tests all yielded unfavorable outcomes in some way. The software code was changed several times until the bugs were fixed and the test cases were successfully cleared in subsequent white box testing rounds. Before deployment/release, the application was also subjected to black-box testing by distributing it to a few friends and coworkers and asking them to try it. All of the bugs they reported were fixed in subsequent code refactoring rounds. Few of the bugs were skipped as they didn't impact the application in any significant way for the release. The entire testing and bug report can be found in the documentation of the software.

The application is currently not deployed on any platform on the internet. But it is hosted on GitHub with its own public repository (URL - https://github.com/alokpurohit18/footstats) that contains all the source code, documentation, packages, and releases. To check the application out, users can download it from this repository and install it on their own systems. The instructions for the installation are mentioned in the readme.md file of the documentation. The application and the GitHub repository is free and open source but under an MIT license.

6.6 Advantages:

A possible use of this app through its statistics is for users to gauge how they have performed in betting markets, to detect trends, or make betting selections. Although this means the app does not include every possible feature relating to football, this app is ahead of the market in this area, providing a view of fixtures that is not available elsewhere.

Another advantage of it is being able to predict the similarity between players. This will allow the coach to identify an ideal and suitable replacement for any player who leaves the club during the transfer window. The xG prediction and market value prediction done by the app will help the coach pick the best 11 for each match and also prevent the club from paying any extra money for any transfer. The player statistics on the other hand will help the club scout players around the world on an initial level before they go ahead and invest in physical scouting.

7. ACTION PLAN

- August 2021: Topic Approval
- September 2021: Literature Survey
- October 2021: Data Collection
- November 2021: Preprocessing
- December 2021: Developing Algorithm and Model
- January-February 2022: Developing Web App, integrating Model
- March 2022: Deployment and Testing

8. CONCLUSION AND FUTURE SCOPE

This project offers something rarely seen in related applications on the market currently. Applications and websites like ”Transfermarkt”,” OneFootball” and more do provide comprehensive details about the world of football, predictive analytics is notably missing from these. This makes it an application encompassing several important areas. The ability to look up any and all results on almost all top leagues quickly thanks to the speed of ReactJS in deployment mode makes this a convenient application to use. This combined with an in-depth analysis of players’ abilities completes the statistics part of the application. What makes it really unique, however, is our additions of predicting the market value of players, checking player similarity, and analysis of expected goals which provide a different dimension to an already robust application.

There is significant potential in this application, particularly in the domain of Deep Learning. The metrics for expected goals are evolving constantly. Developing our own xG algorithm using deep learning is an area that could be explored. Using stills from live matches for each shot, we can develop a novel method to determine expected goals based on our preferred metrics. For example, some xG models consider the position of opposition players on the pitch while some consider only the shot angle and distance from the goal. The contentious nature of this concept makes it an interesting challenge to tackle in the future. All these can be provided in future updates and releases of this software through GitHub.

9. REFERENCES

[1] P. Cintia, F. Giannotti, L. Pappalardo, D. Pedreschi, and M. Malvaldi, “The harsh rule of the goals: Data-driven performance indicators for football teams,” in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10, IEEE, 2015.

[2] K. Tuyls, S. Omidshafiei, P. Muller, Z. Wang, J. Connor, D. Hennes, I. Graham, W. Spearman, T. Waskett, D. Steel, et al., “Game plan: What ai can do for football, and what football can do for ai,” Journal of Artificial Intelligence Research, vol. 71, pp. 41–88, 2021.

[3] U. Lichtenthaler, “Mixing data analytics with intuition: Liverpool football club scores with integrated intelligence,” Journal of Business Strategy, 2020.

[4] M. Mondello and C. Kamke, “The introduction and application of sports analytics in professional sport organizations a case study of the tampa bay lightning,” vol. 6, pp. 1–15, SAGAMORE PUBL LLC 1807 N FEDERAL DR, URBANA, IL 61801 USA, 2014.

[5] D. Memmert, “Data analytics in football: positional data collection, modeling, and analysis,” Journal of Sport Management, vol. 33, no. 574, pp. 2019–0308, 2019.

[6] A. Pearson, “How real-time analytics changes the face of the sports betting industry,” vol. 5, pp. 61–75, Henry Stewart Publications, 2017.

[7] V. De Silva, M. Caine, J. Skinner, S. Dogan, A. Kondoz, T. Peter, E. Axtell, M. Birnie, and B. Smith, “Player tracking data analytics as a tool for physical performance management in football: A case study from chelsea football club academy,” Sports, vol. 6, no. 4, p. 130, 2018.

[8] I. Behravan and S. M. Razavi, "A novel machine learning method for estimating football players' value in the transfer market," vol. 25, pp. 2499–2511, Springer, 2021.

[9] R. Pariath, S. Shah, A. Surve, and J. Mittal, "Player performance prediction in football game," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1148–1153, IEEE, 2018.

[10] A. Rathke, "An examination of expected goals and shot efficiency in soccer," vol. 12, pp. 514–529, Universidad de Alicante, 2017.

[11] R. Rein and D. Memmert, "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science," vol. 5, pp. 1–13, SpringerOpen, 2016.

[12] P. Madrero Pardo, "Creating a model for expected goals in football using qualitative player information," 2020.

[13] J. Stubinger, B. Mangold, and J. Knoll, "Machine learning in football ¨ betting: Prediction of match results based on player characteristics," vol. 10, p. 46, MDPI, 2019.

[14] K. Apostolou and C. Tjortjis, "Sports analytics algorithms for performance prediction," in 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–4, IEEE, 2019.

[15] D. Patnaik, H. Praharaj, K. Prakash, and K. Samdani, "A study of prediction models for football player valuations by quantifying statistical and economic attributes for the global transfer market," in 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–7, IEEE, 2019.

[16] R. Stanojevic and L. Gyarmati, "Towards data-driven football player assessment," in 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), pp. 167–172, IEEE, 2016.

[17] J. Stubinger and J. Knoll, "Beat the bookmaker–winning football bets ¨ with machine learning (best application paper)," in International Conference on Innovative Techniques and Applications of Artificial Intelligence, pp. 219–233, Springer, 2018.

[18] C. P. Igiri, "Support vector machine–based prediction system for a football match result," vol. 17, pp. 21–26, 2015.

[19] P. Rajesh, M. Alam, M. Tahernezhadi, et al., "A data science approach to football team player selection," in 2020 IEEE International Conference on Electro Information Technology (EIT), pp. 175–183, IEEE, 2020.

[20] A. Partida, A. Martinez, C. Durrer, O. Gutierrez, and F. Posta, "Modeling of football match outcomes with expected goals statistic," vol. 10, 2021.

[21] M. He, R. Cachucho, and A. J. Knobbe, "Football player's performance and market value.," in Mlsa@ pkdd/ecml, pp. 87–95, 2015