Prediction on Future Matches and Top Scorers for Season 20-21 in English Premier League (EPL)

by Jin Young Park, Taehyeon Lee

I. Summary of Research Questions and Results

A. For the first part of the project, we will collect match histories from the previous seasons and teams' overall ratings to predict the future matches. With the collected datasets, we will calculate Head-to-Head ratio and All-Time win ratio for each match and team. We will train a machine learning model with these ratios based on match histories, and overall team ratings, and use it to compute probabilities in future matches and to calculate the game points that the teams will earn at the end of the season. With the prediction, we will create a graph showing the current standing of each team and future standing based on the calculation.

B. For the next part, we will collect top scorer standing and data like how many goals scored, shot made per game, goal made per shot, and minutes they played in this season. With the data, we will calculate the total goal scored per player at the end of the season, the player's contribution to the team, and the efficiency of each player.

C. Who is going to win the next match?
   1. Compare match histories for every team for the current season and previous seasons including essential match stats such as goals, passes, and etc to create a machine learning model to predict who is going to win the next match of the current league. It will predict every game left in the season and also provide approximate total points at the end of the season. Based on the machine learning model it was able to derive the winner by Hometeam, Awayteam, or Draw

D. Which team will be the Champion of the Season 20-21?
   1. With the prediction of the result of each match, calculate the points(win:3, draw:1, lose:0) and add the points to the current points of the teams. The team with the highest points will be the champion of the season.
      a) The Manchester City was the champion of the season by 80 points

E. Who will be the top scorer for the season?
   1. Compute how many goals will a top scorer for the season score by the end of this season. It will pick every player in the team who is an offensive player (striker and attacking midfielders) and compute the average goal per playing time rate for the season. Create a machine learning model to compute how many goals potential-top-scorers would produce and apply the rate for the current leftover matches and add them up to the current score they made in the season so far to finalize the top scorer.
      a) Salah was top scorer by 22points

II. Motivation and background

Even if the world of sports is constant competition, behind the scenes, money to organize and manage a team plays a significant role. In order to get sufficient funds and budget, the team manager should perform well in the league and come up with a winning strategy. To set up a successful plan, we believe, collecting and analyzing appropriate data is essential. As a result, computing winning rates for each match and various statistics to predict the winner of a future match, champion of the league, and top scorer can support building the strategy.

When predicting which team is going to win for each game based on the previous match data and overall score of the team, the manager can come up with a special plan or strategy for the opponent. While it predicts the probability of winning, it can also easily find out the possible points earned for the season and estimate where the team would be placed by comparing the points. The result of the league is a very big outcome for the team and manager because where the team placed at the end of season influences the budget and fund for the team. European soccer clubs are playing several different tournaments along with English Premier League as well. Even though every match is crucial, if the team is suffering in the Premier League but there is a possibility to win another tournament, then the manager of the team should prepare a strategy and organize a running time for players. Based on the prediction of the champion of the league and placement, the manager can prepare where the team should focus and pursue a trophy cup for other leagues because it can bring more funds for the team management by selling goods, advertisement and other sources. Moreover, predicting the top scorer and scores the player will make can be crucial data for the team to consider when the trade season comes because the team and manager can know the exact value of the player and offer reasonable salary or trade for other players who they might need most. Thus, in order to run and manage the team smoothly with sufficient funds analysis and prediction for a next match, prospect champion and top scorer can be very useful.

III. Dataset

Our datasets are from various sources and from these following links:
Seasonal data: https://www.football-data.co.uk/englandm.php
FIFA overall data: https://www.fifaindex.com/teams/?league=13&order=desc
Current Standing: https://www.msn.com/en-us/sports/soccer/premier-league/standings
Top Scorer data:
https://www.msn.com/en-us/sports/soccer/premier-league/player-stats/sp-s-gs
Current season schedules: https://fixturedownload.com/results/epl-2020

We got total 8 seasonal data from 2013-14 to 2020-21 seasons as csv files. Then we found FIFA overall data from the link and scraped overall scores for each team the corresponding seasons via BeautifulSoup library from python. We used BeautifulSoup again to scrape current standing data and top scorer data from the links above. Lastly, we found the whole schedule of the current season to make predictions for future matches.

Seasonal data contains from teams to many other match stats such as full time home team goal, away team goal, result and etc..

Div	Date	Time	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HTHG	HTAG	HTR	Referee	HS	AS	HST	AST
E0	12/9/20	12:30	Fulham	Arsenal	0	3	A	0	1	A	C Kavanagh	5	13	2	6
E0	12/9/20	15:00	Crystal Palac	Southampto	1	0	H	1	0	H	J Moss	5	9	3	5
E0	12/9/20	17:30	Liverpool	Leeds	4	3	H	3	2	H	M Oliver	22	6	6	3
E0	12/9/20	20:00	West Ham	Newcastle	0	2	A	0	0	D	S Attwell	15	15	3	2
E0	13/09/2020	14:00	West Brom	Leicester	0	3	A	0	0	D	A Taylor	7	13	1	7
E0	13/09/2020	16:30	Tottenham	Everton	0	1	A	0	0	D	M Atkinson	9	15	5	4
E0	14/09/2020	20:15	Brighton	Chelsea	1	3	A	0	1	A	C Pawson	13	10	3	5
E0	14/09/2020	18:00	Sheffield Un	Wolves	0	2	A	0	2	A	M Dean	9	11	2	4
E0	19/09/2020	12:30	Everton	West Brom	5	2	H	2	1	H	M Dean	17	6	7	4
E0	19/09/2020	15:00	Leeds	Fulham	4	3	H	2	1	H	A Taylor	10	14	7	6
E0	19/09/2020	17:30	Man United	Crystal Palac	1	3	A	0	1	A	M Atkinson	13	14	4	5
E0	19/09/2020	20:00	Arsenal	West Ham	2	1	H	1	1	D	M Oliver	7	14	3	3
E0	20/09/2020	12:00	Southampto	Tottenham	2	5	A	1	1	D	D Coote	14	9	7	6
E0	20/09/2020	14:00	Newcastle	Brighton	0	3	A	0	2	A	K Friend	6	13	0	6
E0	20/09/2020	16:30	Chelsea	Liverpool	0	2	A	0	0	D	P Tierney	5	18	3	6
E0	20/09/2020	19:00	Leicester	Burnley	4	2	H	1	1	D	L Mason	14	16	6	5
E0	21/09/2020	18:00	Aston Villa	Sheffield Un	1	0	H	0	0	D	G Scott	18	4	2	1
E0	21/09/2020	20:15	Wolves	Man City	1	3	A	0	2	A	A Marriner	10	14	1	9
E0	26/09/2020	12:30	Brighton	Man United	2	3	A	1	1	D	C Kavanagh	18	7	5	3
E0	26/09/2020	15:00	Crystal Palac	Everton	1	2	A	1	2	A	K Friend	8	10	1	5
E0	26/09/2020	17:30	West Brom	Chelsea	3	3	D	3	0	H	J Moss	9	22	3	10
E0	26/09/2020	20:00	Burnley	Southampto	0	1	A	0	1	A	A Marriner	10	5	2	1
E0	27/09/2020	12:00	Sheffield Un	Leeds	0	1	A	0	0	D	P Tierney	14	17	4	9
E0	27/09/2020	14:00	Tottenham	Newcastle	1	1	D	1	0	H	R Bankes	23	6	12	1

FIFA overall data has team and score for each team

Team	Overall
Man City	85
Liverpool	85
Tottenham	82
Man United	82
Chelsea	82
Leicester	80
Arsenal	80
Everton	79
Wolves	79
West Ham	78
Southamptor	77
Aston Villa	77
Newcastle	76
Crystal Palac	76
Burnley	76
Leeds	75
Brighton	75
Fulham	74
West Brom	74
Sheffield	74

IV. Methodology

First, we searched and gathered all the datasets we needed. We gathered total 8 seasons of individual season data, current schedule data, and scraped using BeautifulSoup to gather overall score, top scorer and standing.

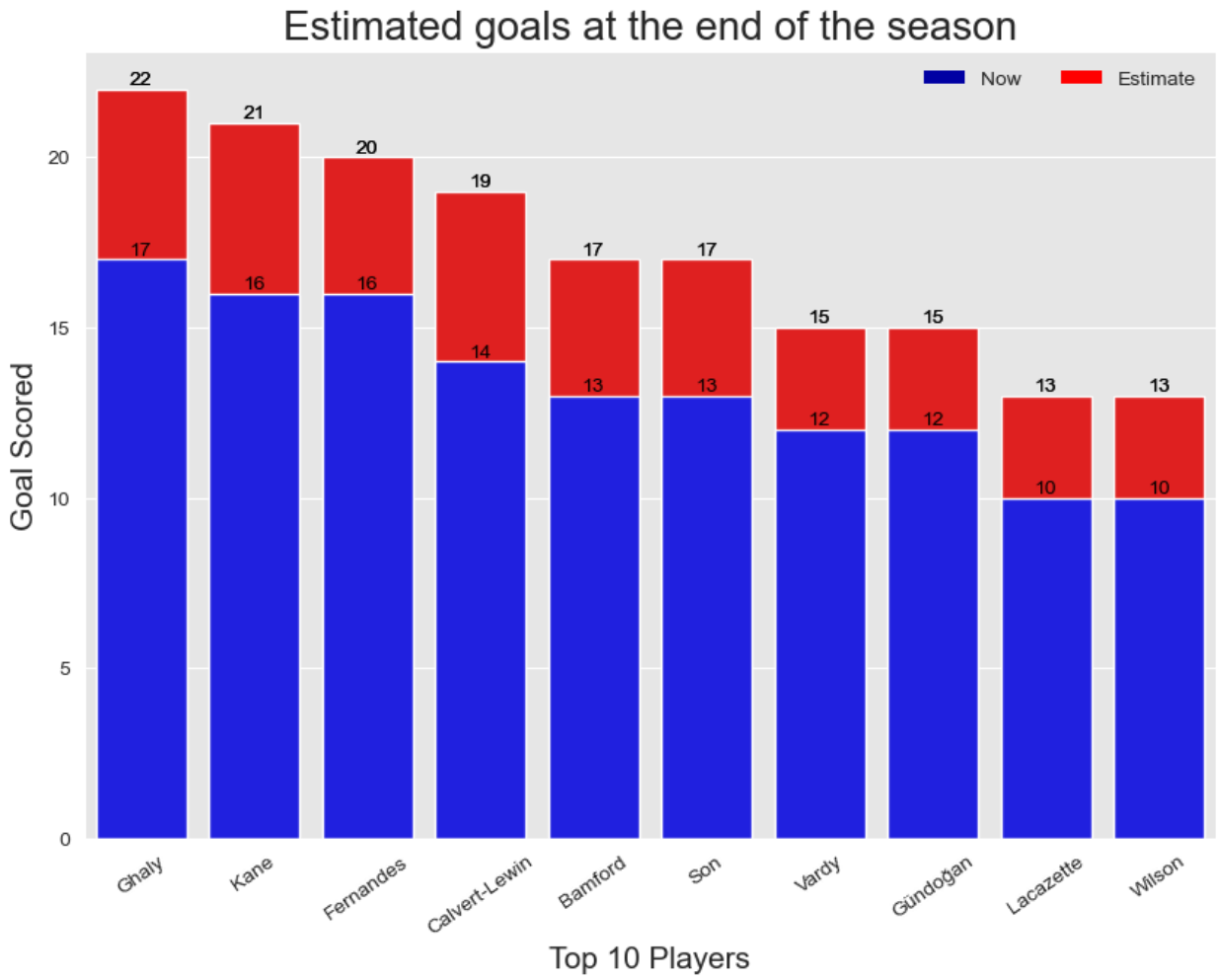
After we all gathered data we started to clean and filter data. We first unify team names so we can merge or concatenate data by name. We changed team names on FIFA

overall datasets to match teams in seasonal data and merged corresponding years. Then we concatenated all seasons but current season to create one dataframe and dropped unnecessary columns. In order to prepare machine learning features, we then calculated head-to-head ratio which looks at matching history with the individual opponent and calculated the winning rate over the team. We also computed overall winning ratio for home and away teams where it finds the number of all wins over the total number of matches. Furthermore, we merged two dataset: top scorer and current team standing and dropped players under top 100 and unnecessary columns.

To predict the next round winner and champion of the season, we used a machine learning model: RandomFroestClassifier. The training data was all seasons but current season and test data was current season upto where the match occurred. We set label for column called ‘FTR’ which is full time result and rest of other columns were features. We first imported the model and trained model based on train features and train label. After the train our model was given the test dataset of features and label and the accuracy for the model was about 52%. Afterward, we gave model the future matches and predicted match result out of it. Based on the prediction we calculated the total points and added to the current standing point.

V. Results

Based on the model we figured out the champion of the season was Manchester City by 80 points and top scorer was Salah by 22 goals.



VI.

VII. Challenge Goals

1. **Multiple datasets: The dataset represents different stats such as match histories and each team's overall rating per season. Thus we need to combine multiple datasets to compute and run the machine learning model.**
2. **Messy Data: The dataset for calculating how many seasons to take to yield a great strike pairing is not presented in a csv file. We need to gather information and stats from articles and websites for their performances, nationalities, and number of seasons played in the same team.**
3. **Machine Learning: Make predictions on which team would win for the next match based on the previous histories and match stats and who would be a top scorer for the current season. In order to go further, we will use hyperparameters to find which model derives the best results.**

VIII. Work Plan Evaluation

1. **Clean and filter the data (3 hours)**
2. **Collect dataset we need from multiple sources**
3. **Clean and filter down to the data columns we need**

1. **Work on research questions 1, 2, 3 (16 hours)**
2. **Have a zoom meeting every time we work on code**
3. **Calculate the winning rates of each team based on past and present seasons**
4. **Select potential top scorers from attacking players of each team and calculate average goal per playing time rate**
5. **Use a classifier model to compute the winning team of each match, the champion of the season, and top scorer of the season.**
6. **Prepare report - 3 hours**

IX. Testing

We tested our machine learning model through our created testing dataset.

X. Collaboration

We used many concepts that were not taught in the class such as RandomTreeClassfier and beautiful soup. For RandomTreeClassifier, we used google and kaggle to look over the usage of it and used it in our code. Same with other materials, we mainly used google and stackoverflow.