Review - 2

Predicting football matches using Machine Learning

Akshay Murthy	21BCE2837
S Rahul	21BCI0036

Objectives

- Collect match data using web scraping using BeautifulSoup. We will collect 2 years worth of football from English Premier League

- Use it to extract features and predict whether the the home team would win or not using the following Machine Learning algorithms with and without rolling averages
    - Naive Bayes Classifier
    - Logistic Regression
    - Support Vector Machines
    - Random Forest Classifier

- Demonstrate how rolling averages enhances the outcomes of these machine learning algorithms

- Document the best features to consider for the rolling averages

Motivation behind the project

- **Predicting the outcome of football matches has always been a challenging yet fascinating task.**

- **Many factors, including team form, player performance, and match conditions, influence the result.**

- **Football predictions can be used in betting markets, team strategy analysis, and fan engagement.**

- **Season long statistics don’t capture dynamic form. Hence we research the use of rolling averages to capture a team’s dynamic form.**

- **Machine learning models have been increasingly used to predict outcomes based on historical data.**

Literature Survey

Analysing the results of the 12 research articles and researching we found that

- The Hucaljuk,J & Rakipović,A (2011) proposed to build a system that would predict result at the full time of all matches from the Champions League in Europe. The main ones were namely Naive Bayes, k-nearest neighbors, Random forest and Bayesian networks. Results of the k-nearest neighbours were relatively poor.

- O.Fazrin in 2013 used Bayesian networks to predict the results. This time the predictions were based on better features, which were divided into two categories – non physiological and physiological factors. While better features were used, the desired results were not obtained as the dataset was based on just one team

- Darwin, P & Dra, H (2016) proposed a logistic regression model to estimate results of the 2015 season of the English Premier League. Based on [4], they developed this model with the help of data using four significant variables. They concluded that home defense and away defense were the most important variables.

Literature Survey

- Referees [12] can significantly influence the final score in football, often favoring the home team due to bias. Experienced referees tend to make more accurate decisions, but the impact of crowd support can exacerbate referee favoritism.

- Gomes [6] used statistics from more than 4900 games spanning 13 seasons (from 2000/01 to 2012/13) to develop prediction models. Such models were tested in 7 rounds of the 2013/14 season, making a total of 70 games and achieving an accuracy rate of 54.29%.

- An in-depth investigation [7] was carried out to try to understand what led to this surprising victory and to try to understand how future predictions can be made. It was found that the achievement was attained due to the excellent performance of the Leicester goalkeeper and the fact that they were very effective at scoring on counterattacks. Another important factor was that several Leicester players made a large number of interceptions of passes

Literature Survey

- Reddy [8] also included logistic regression in one of the methods he used to predict English Premier League season 2012-2013 matches. His research concluded that the significant variables were number of away goals scored, number of red cards, home team position, and shots on target.

- Berrar et al. (2018) [3] applied two feature engineering methods, the recency feature extraction and the rating feature learning, combined with two machine learning models, a k-NN and an extreme gradient boosted tree (XGBoost), to predict a match outcome

Gap Identification

- While machine learning has been widely applied to predict football match outcomes, many existing studies overlook a crucial aspect of the game—recent team form
- Most approaches rely on static historical data or aggregate season-long statistics, which fail to account for fluctuations in team performance over time.
- Football is highly dynamic, with a team's form being a key factor that can significantly influence match outcomes. However, there has been limited focus on incorporating recent form into predictive models.
- To address this gap, this project introduces the use of **rolling averages** for key offensive and defensive attributes, such as goals, shots on target, and defensive performance metrics, from a team's previous matches.
- By calculating rolling averages, we capture a team's current form leading up to each match, offering a more accurate and dynamic representation of their performance.

Problem Formulation

- Predicting football match outcomes is a complex task influenced by numerous dynamic factors, including team form, player performance, and match conditions.

- Traditional prediction methods often fail to capture the nuanced impact of recent team performance and key match statistics.

- This project aims to develop machine learning models that incorporate rolling averages of offensive and defensive attributes to enhance prediction accuracy.

- The problem is formulated as a classification task, predicting match outcomes based on historical data and calculated rolling averages. We will document the best features to consider for the rolling averages

Rolling averages

It calculates the average of a specific number of recent data points. This is helpful because it:

- **Captures Recent Form:** Rolling averages are particularly useful in sports like football where team performance can fluctuate significantly over time. By focusing on recent matches, they can capture changes in a team's playing style, injuries, or other factors that may influence their current form.
- **Reduces Noise from Historical Data:** Historical data can be noisy, especially for teams with inconsistent performance over the years. Rolling averages can help filter out this noise and focus on the more relevant recent trends.
- **Adapts to Changing Circumstances:** Rolling averages can adapt to changes in the league or individual teams. For example, if a team acquires a new star player or changes their manager, rolling averages can quickly reflect these changes and provide a more accurate prediction.

Project Plan

- Data collection: Scrape detailed football match data from FBref.com using BeautifulSoup
- Feature Extraction: Extract essential features from the scraped data, including team performance metrics, venue details, and opponent statistics, to serve as inputs for machine learning models.
- Rolling averages: : Calculate rolling averages for key offensive and defensive attributes from a team's previous matches to capture recent form and enhance prediction accuracy
- Model Comparison: Apply various machine learning algorithms, including Naive Bayes, Logistic Regression, Support Vector Machines (SVM) and Random Forest.Compare the predictive accuracy of these models with and without the incorporation of rolling averages.
- Performance evaluation: Evaluate and document the performance of each model in terms of prediction accuracy, highlighting the effectiveness of rolling averages in improving model performance.

Project Plan

We will be predicting the result as the target variable in the form of

- 1: Home Win
- 0: No home win

Requirement Analysis

Programming Language: Python: The primary language for implementing data scraping, feature extraction, and machine learning models. We will be using Jupyter notebook as the development environment

Libraries:

- **BeautifulSoup**: For web scraping football data from FBref.com.
- **pandas**: For data manipulation and rolling averages calculation.
- **NumPy**: For numerical operations and data handling.
- **Scikit-learn**: For implementing Naive Bayes, Logistic Regression, SVM, and Random Forest machine learning models.
- **Matplotlib**: For visualizing model performance and results.

Features

Venue:

- The venue plays a crucial role in match outcomes, as home teams often perform better due to familiarity with the environment and support from local fans.

Opponent:

- The strength and recent form of the opposing team heavily influence match results, as teams adjust their strategies based on their opponent’s tactics and abilities.

Referee:

- The referee can impact the flow of the game, with certain referees known for being stricter or lenient, affecting the number of fouls, penalties, and overall match tempo.

Hour:

- Match timing can influence player performance due to factors such as fatigue, recovery time, and even weather conditions, making the hour of the match a relevant predictor.

Day of the Week:

- Matches played on weekends or weekdays can differ in intensity.

Features

Attack Strength (Fixed Rating)

- The attack strength of each team is derived from an external rating system. This fixed rating quantifies the offensive capabilities of a team, including factors like goal-scoring potential, shot creation, and overall attacking performance.

Defense Strength (Fixed Rating)

- Similar to the attack strength, the defense strength rating assesses a team’s defensive capabilities. It factors in the ability to prevent goals, successful tackles, interceptions, and general defensive organization. These fixed ratings help provide an objective measure of a team's form.

Rolling averages

Rolling averages features

Attacking Attributes (Rolling Averages):

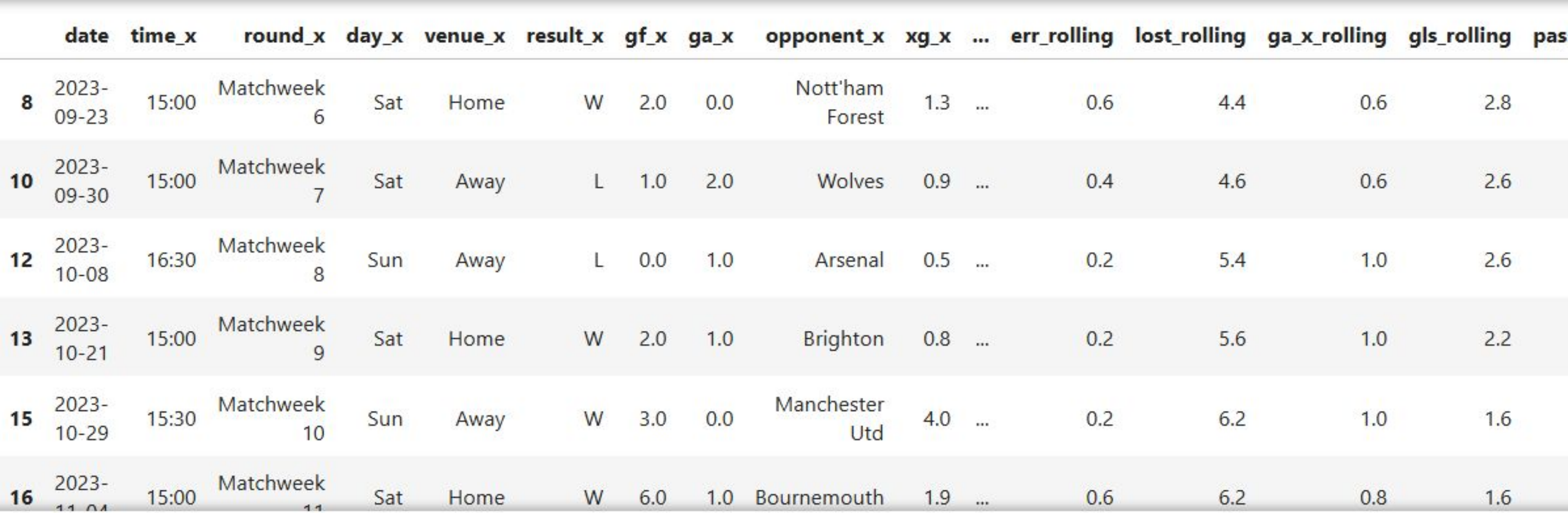
- **Shots on Target:** Measures how often a team is able to place the ball on goal, indicating offensive efficiency.
- **xG (Expected Goals):** Quantifies the quality of chances a team creates, based on shot location and context.
- **Goals Scored:** A direct measure of a team's attacking success and ability to finish chances.

- **Average Shot Distance:** Evaluates the range from which shots are taken, indicating attacking positioning and chance creation quality.
- **Free Kicks Attempted:** Highlights a team's ability to create set-piece opportunities, a key aspect of goal-scoring potential.
- **Penalty Kicks Attempted:** Reflects offensive pressure and penalty box presence, contributing to scoring chances.

Rolling averages features

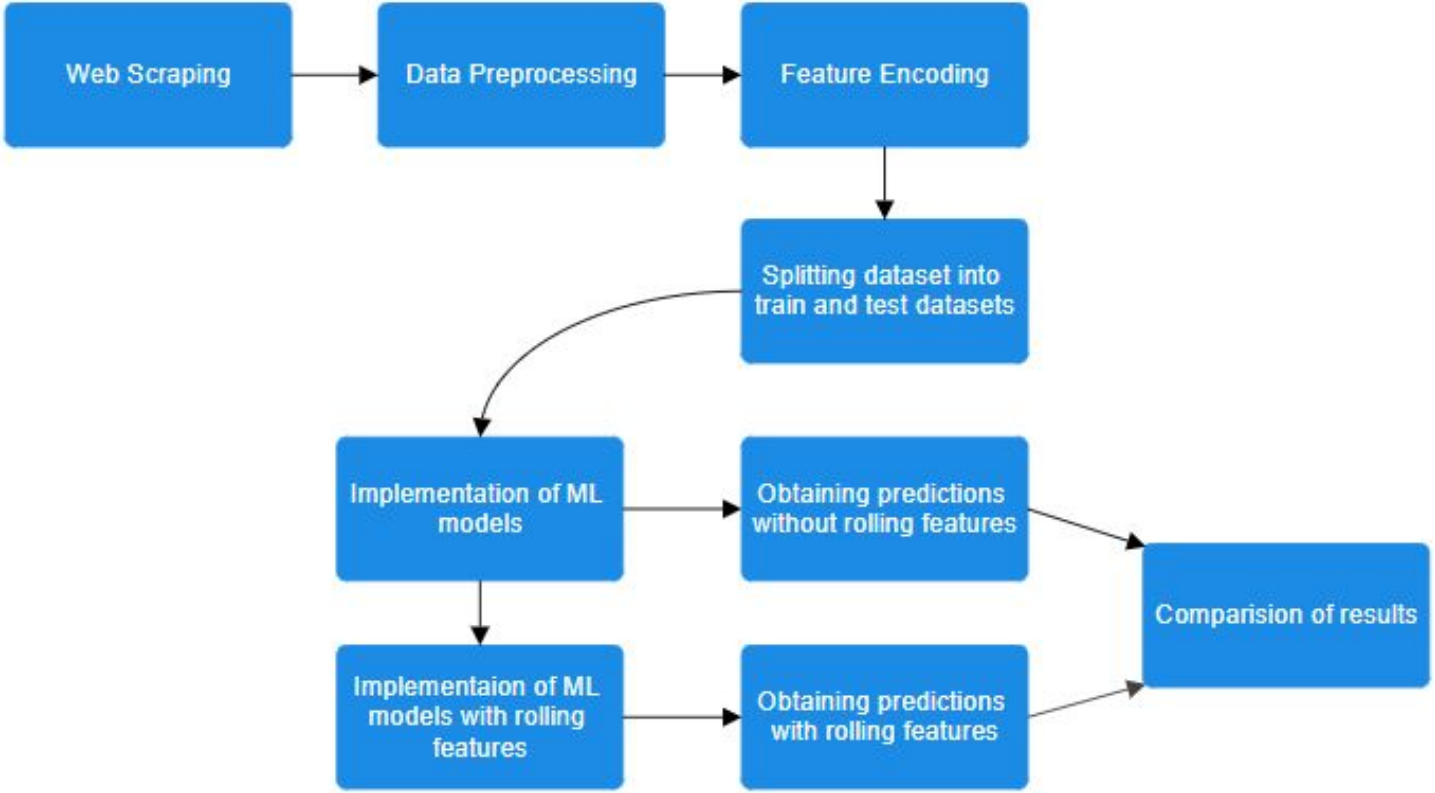
Defensive Attributes (Rolling Averages):

- **Tackles Attempted:** Shows defensive engagement and effort in disrupting the opposition’s play.
- **Tackles Won:** Measures successful defensive interventions, indicating a team's defensive solidity.
- **Dribblers Tackle %:** Indicates how often defenders succeed in dispossessing dribbling attackers.
- **Shots Blocked:** Reflects the defense's ability to prevent goal-scoring opportunities.
- **Passes Blocked:** Shows the effectiveness of the defense in intercepting and stopping opposition passing plays.
- **Clearances:** Reflects how often defenders remove the ball from dangerous areas.
- **Errors Made:** Tracks defensive mistakes that may lead to goal-scoring opportunities for the opponent.
- **Goals Conceded:** Reflects defensive vulnerability, which indirectly impacts a team's attacking mindset.



	date	time_x	round_x	day_x	venue_x	result_x	gf_x	ga_x	opponent_x	xg_x	...	err_rolling	lost_rolling	ga_x_rolling	gls_rolling	pas
8	2023-09-23	15:00	Matchweek 6	Sat	Home	W	2.0	0.0	Nott'ham Forest	1.3	...	0.6	4.4	0.6	2.8	
10	2023-09-30	15:00	Matchweek 7	Sat	Away	L	1.0	2.0	Wolves	0.9	...	0.4	4.6	0.6	2.6	
12	2023-10-08	16:30	Matchweek 8	Sun	Away	L	0.0	1.0	Arsenal	0.5	...	0.2	5.4	1.0	2.6	
13	2023-10-21	15:00	Matchweek 9	Sat	Home	W	2.0	1.0	Brighton	0.8	...	0.2	5.6	1.0	2.2	
15	2023-10-29	15:30	Matchweek 10	Sun	Away	W	3.0	0.0	Manchester Utd	4.0	...	0.2	6.2	1.0	1.6	
16	2023-11-04	15:00	Matchweek 11	Sat	Home	W	6.0	1.0	Bournemouth	1.9	...	0.6	6.2	0.8	1.6	

Design and Implementation



Web Scraping

Data Preprocessing

Feature Encoding

Splitting dataset into train and test datasets

Implementation of ML models

Obtaining predictions without rolling features

Implementaion of ML models with rolling features

Obtaining predictions with rolling features

Comparision of results

Architecture

Web Scraping

- Collecting data from the web to build the dataset.

Data Preprocessing

- Cleaning and preparing the data for analysis.

Feature Encoding

- Transforming categorical data into a numerical format for compatibility with ML algorithms.

Splitting Dataset

- Dividing the data into training and testing datasets.

Implementation of ML Models

- Training and testing machine learning models without rolling features.

Architecture

Obtaining Predictions (Without Rolling Features)

- Using the trained models to make predictions based on the dataset without rolling features.

Implementation of ML Models with Rolling Features

- Training and testing the models with rolling features included in the dataset.

Obtaining Predictions (With Rolling Features)

- Generating predictions using the models trained on datasets with rolling features.

Comparison of Results

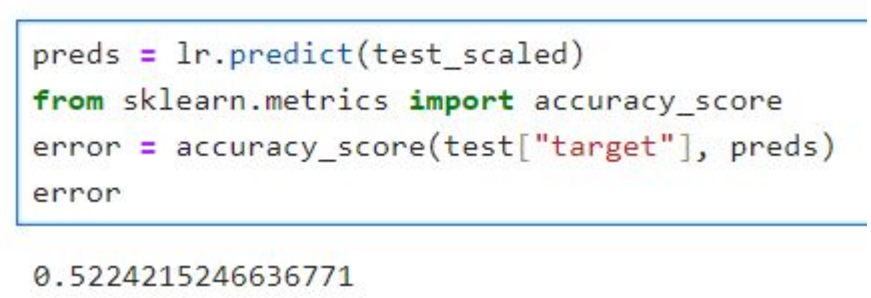
- Evaluating and comparing the model performances with and without rolling features to analyze the impact of rolling features.

Experimental result and analysis

Comparison of accuracy with and without rolling averages

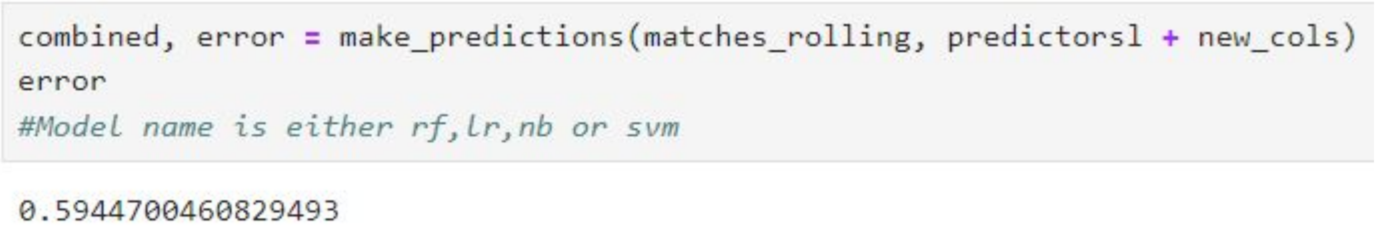
Machine Learning Algorithm	Accuracy without Rolling averages	Accuracy with rolling averages
Random Forest	66%	68%
Naive Bayes	58.5%	64.9%
SVM	62.1%	66.5%
Logistic Regression	52.24%	59.4%

Logistic regression



```
preds = lr.predict(test_scaled)
from sklearn.metrics import accuracy_score
error = accuracy_score(test["target"], preds)
error
```

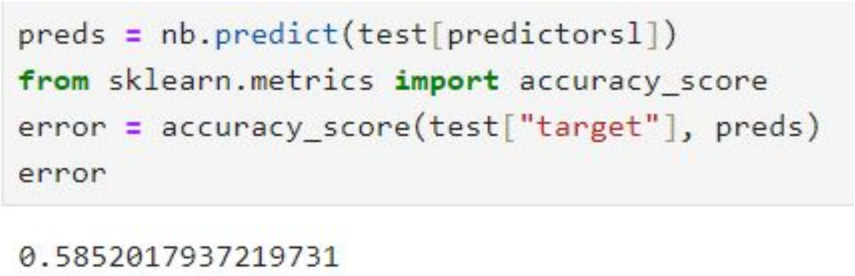
```
0.5224215246636771
```



```
combined, error = make_predictions(matches_rolling, predictorsl + new_cols)
error
#Model name is either rf,lr,nb or svm
```

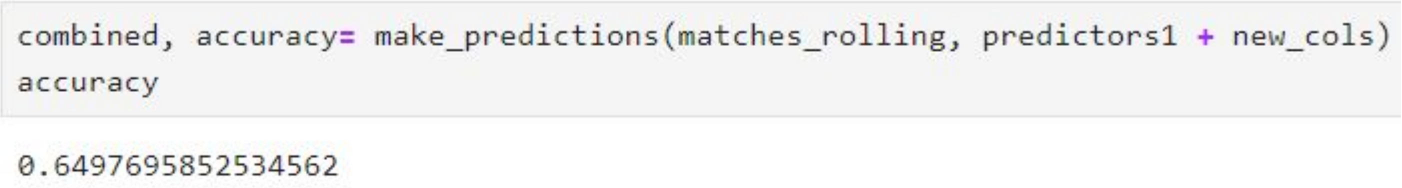
```
0.5944700460829493
```

Naive Bayes



```
preds = nb.predict(test[predictors1])
from sklearn.metrics import accuracy_score
error = accuracy_score(test["target"], preds)
error
```

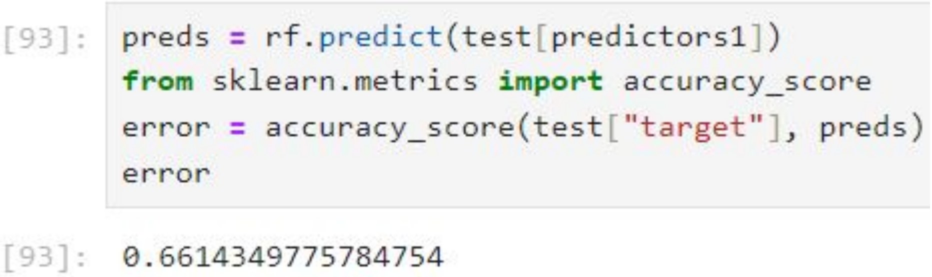
```
0.5852017937219731
```



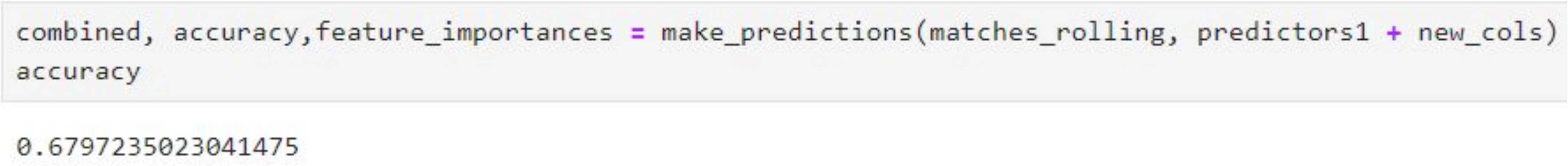
```
combined, accuracy= make_predictions(matches_rolling, predictors1 + new_cols)
accuracy
```

```
0.6497695852534562
```

Random Forest

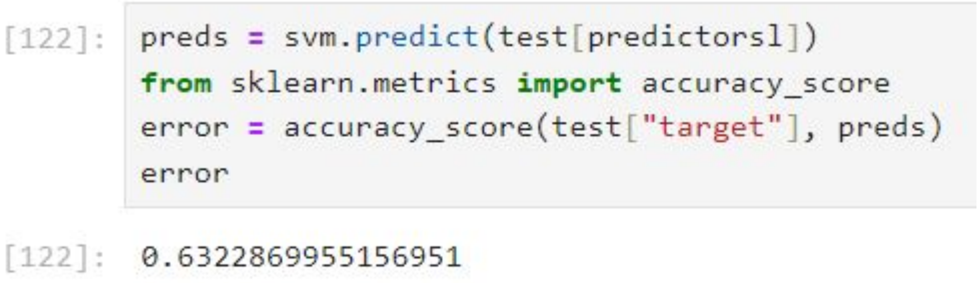


```
[93]: preds = rf.predict(test[predictors1])
      from sklearn.metrics import accuracy_score
      error = accuracy_score(test["target"], preds)
      error

[93]: 0.6614349775784754
```

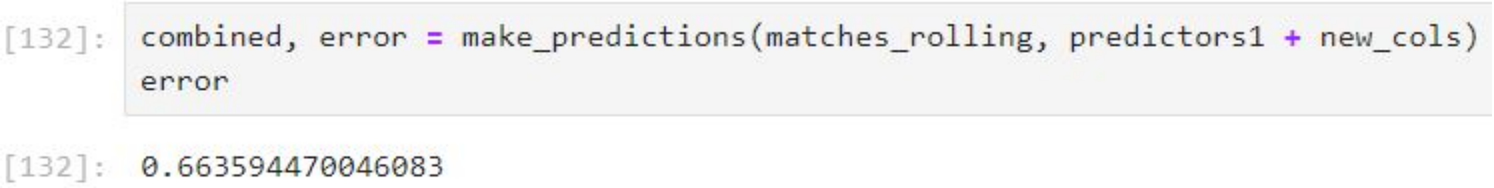


```
combined, accuracy,feature_importances = make_predictions(matches_rolling, predictors1 + new_cols)
accuracy

0.6797235023041475
```

Support vector machines

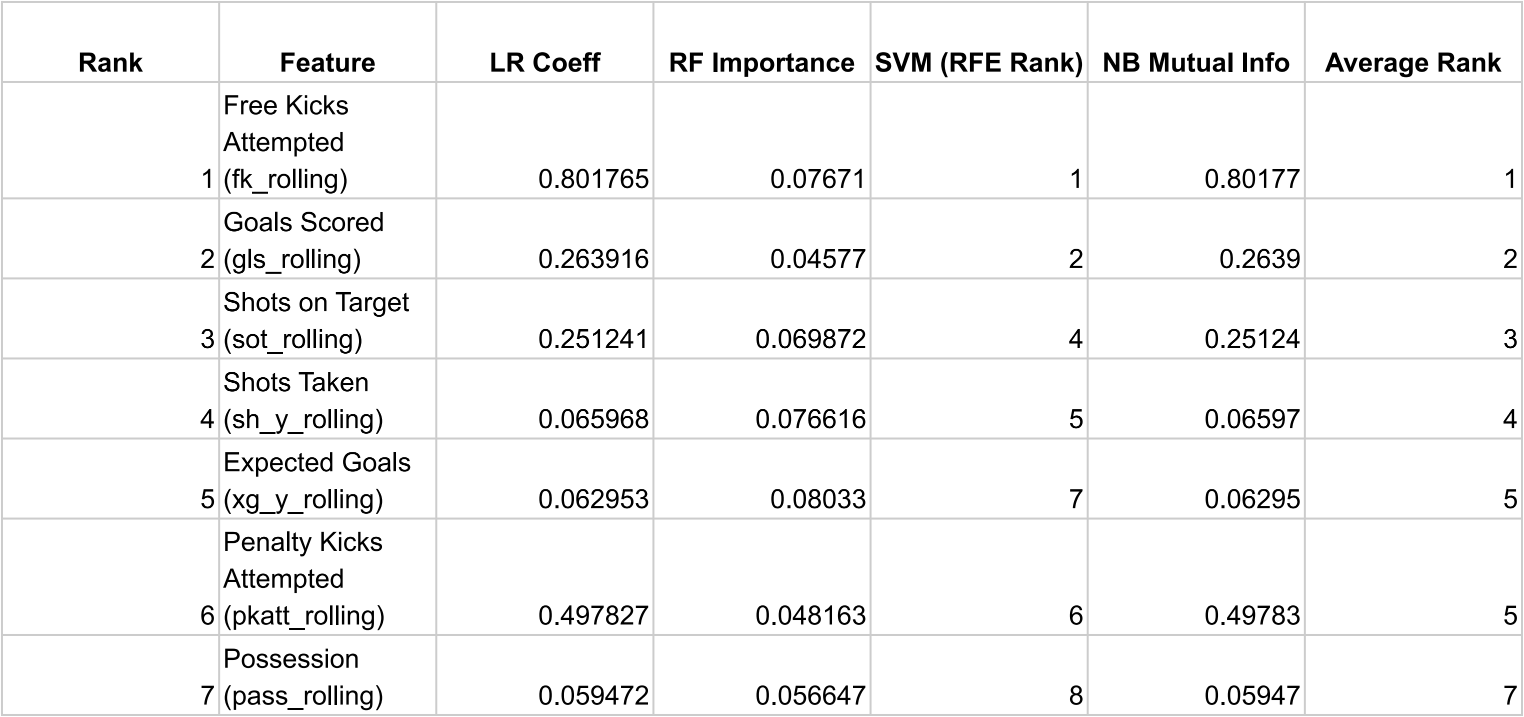


```
[122]: preds = svm.predict(test[predictors1])
       from sklearn.metrics import accuracy_score
       error = accuracy_score(test["target"], preds)
       error

[122]: 0.6322869955156951
```



```
[132]: combined, error = make_predictions(matches_rolling, predictors1 + new_cols)
       error

[132]: 0.663594470046083
```

Highest ranking offensive features



Rank	Feature	LR Coeff	RF Importance	SVM (RFE Rank)	NB Mutual Info	Average Rank
1	Free Kicks Attempted (fk_rolling)	0.801765	0.07671	1	0.80177	1
2	Goals Scored (gls_rolling)	0.263916	0.04577	2	0.2639	2
3	Shots on Target (sot_rolling)	0.251241	0.069872	4	0.25124	3
4	Shots Taken (sh_y_rolling)	0.065968	0.076616	5	0.06597	4
5	Expected Goals (xg_y_rolling)	0.062953	0.08033	7	0.06295	5
6	Penalty Kicks Attempted (pkatt_rolling)	0.497827	0.048163	6	0.49783	5
7	Possession (pass_rolling)	0.059472	0.056647	8	0.05947	7

Highest ranking defensive features

Rank	Feature	LR Coeff	RF Importance	SVM (RFE Rank)	NB Mutual Info	Average Rank
1	Defensive Errors (err_rolling)	0.127877	0.039102	3	0.12788	1
2	Tackles Attempted (tkl_rolling)	0.076003	0.058076	8	0.076	2
3	Goals Against (ga_x_rolling)	0.060905	0.046184	9	0.06091	3
4	Clearances (clearances_rolling)	0.062953	0.056647	11	0.06295	4
5	Interceptions (interceptions_rolling)	0.059472	0.056647	10	0.05947	5
6	Blocks (blocks_rolling)	0.039102	0.039102	12	0.0391	6
7	Dribbles Tackled (drib_tklw_rolling)	0.091476	0.04577	13	0.09148	7

Best Machine Learning Model

- **Random Forest consistently achieved the highest accuracy among the models due to its ability to handle complex interactions and non-linear relationships between features, which are common in football match data.**

- **Random Forest effectively handles large datasets with many features, such as rolling averages, venue, and opponent stats,**

- **Random Forest provides feature importance metrics, helping to identify key predictors (e.g., rolling averages of attacking and defensive stats),**

Conclusion and Future Development

- Machine learning models like Naive Bayes, Logistic Regression, SVM, and Random Forest were successfully applied to predict football match outcomes using historical match data.

- The incorporation of rolling averages for offensive and defensive attributes significantly improved the predictive accuracy by dynamically capturing team form.

- Feature importance analysis and validation techniques ensured a robust model while identifying the most influential rolling average attributes contributing to match outcomes.

- The identified rolling average features can serve as a foundation for future studies to capture team form more effectively and further refine prediction models for various sports analytics applications.

References

[1] Hucaljuk and Rakipovic, “Predicting football scores using machine learning techniques,” 2011 Proceedings of the 34th International Convention MIPRO, pp. 1623–1627, Jan. 2011, [Online]. Available: http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000005967321

[2] F. Owramipur, P. Eskandarian, and F. S. Mozneb, “Football Result Prediction with Bayesian Network in Spanish League-Barcelona Team,” International Journal of Computer Theory and Engineering, pp. 812–815, Jan. 2013, doi: 10.7763/ijcte.2013.v5.802.

[3] D. Prasetio and Dra. Harlili, “Predicting football match results with logistic regression,” N/A, Aug. 2016, doi: 10.1109/icaicta.2016.7803111.

References

[4] H. Arntzen and L. M. Hvattum, “Predicting match outcomes in association football using team ratings and player ratings,” Statistical Modelling, vol. 21, no. 5, pp. 449–470, Jul. 2020, doi: 10.1177/1471082x20929881.

[5] “Decision Support System for predicting Football Game result,” book, 348AD.

[6] H. Ruiz, P. Power, X. Wei, and P. Lucey, “‘The Leicester City Fairytale?,’” N/A, Aug. 2017, doi: 10.1145/3097983.3098121.

References

[7] G. Chakraborty and Vandana Reddy & Sai Vijay Kishore Movva, The Soccer Oracle: Predicting Soccer Game Outcomes Using SAS® Enterprise Miner™. 2014. [Online]. Available: https://www.researchgate.net/publication/279530821

[8] M.-C. Malamatinos, E. Vrochidou, and G. A. Papakostas, “On Predicting Soccer Outcomes in the Greek League Using Machine Learning,” Computers, vol. 11, no. 9, p. 133, Aug. 2022, doi: 10.3390/computers11090133.

[9] D. Berrar, P. Lopes, and W. Dubitzky, “Incorporating domain knowledge in machine learning for soccer outcome prediction,” Machine Learning, vol. 108, no. 1, pp. 97–126, Aug. 2018, doi: 10.1007/s10994-018-5747-8.

[10] F. Rodrigues and Â. Pinto, “Prediction of football match results with Machine Learning,” Procedia Computer Science, vol. 204, pp. 463–470, Jan. 2022, doi: 10.1016/j.procs.2022.08.057.

References

[11] J. D. Rose, M. K. Vijaykumar, U. Sakthi, and P. Nithya, “Comparison of Football Results Using Machine Learning Algorithms,” 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), Jul. 2022, doi: 10.1109/icses55317.2022.9914265.

[12] M. Fontenla and G. M. Izón, “The effects of referees on the final score in football,” *Estudios Económicos*, vol. 35, no. 70, pp. 79–98, Jun. 2018, doi: 10.52292/j.estudecon.2018.1098.

K Fold Cross Validation and LeaveOneOut cross validation

Machine Learning Algorithm	K fold cross validation	LeaveOneOut
Random Forest	63%	66.9%
Naive Bayes	60.4%	59%
SVM	63.2%	60.4%
Logistic Regression	60.3%	59.9%