CAN DATA BE USED TO GAIN AN ADVANTAGE IN SPORTS BETTING?

David Mooney, CCT College Dublin, November 2024

ABSTRACT

Sports betting has always been a prominent past time in Ireland. The first gambling laws were established here in 1926 and after several amendments, most recently in 2015, all forms of regulated sports betting are now legal. *(Brent, H., 2024)* According to (*Health Research Board, August, 2023),* online gambling in Ireland has increased by 300% since 1998.

My intention from the outset of this project is to identify particular markets where the book maker edge is the lowest such as over/under goals per game. The area in particular I will focus on is football. Using historical football data, I will analyse patterns and hopefully make predictions about upcoming games and potential goals scored. By studying a team's goal scored and goals conceded I will be able to calculate the expected goals per game. Originally, I use the English Premier League as there is an abundance of data available relating to it but moving forward, I intend to apply my algorithm to different football league and maybe sports.

Objectives

Market Identification: Research and identify particular markets where the bookies edge is at its lowest. Knowing this will give me a better insight into what kind of data I will need to gather.

Create Predictive Models: Applying machine learning algorithms will enable me to develop predictive models. In their implementation, these models will attempt to predict match results and individual player stats. They will give me an insight into team performances while also casting light on how certain teams perform against other particular teams.

Testing Models: Ensure the model is performing accurately by acquiring the F1 score. Once the model is adequately tested, new data will be applied to ensure it can retain accuracy.

Applying Other Metrics: Extracting other metrics from the data, such as expected goals, will provide a more in-depth analysis of how each team performs on a weekly basis. These metrics will help create an estimation of how many goals a team should score and/or concede, thus aiding in the prediction of the outcome.

Develop Betting Strategy: To ensure profitability, I will develop a betting strategy based off the results of the data acquired. This will most likely come in the form of a spreadsheet that details stake size, edge percentage of particular markets and market fluctuations. *I will not be doing any real gambling. All bets will be fictional.*
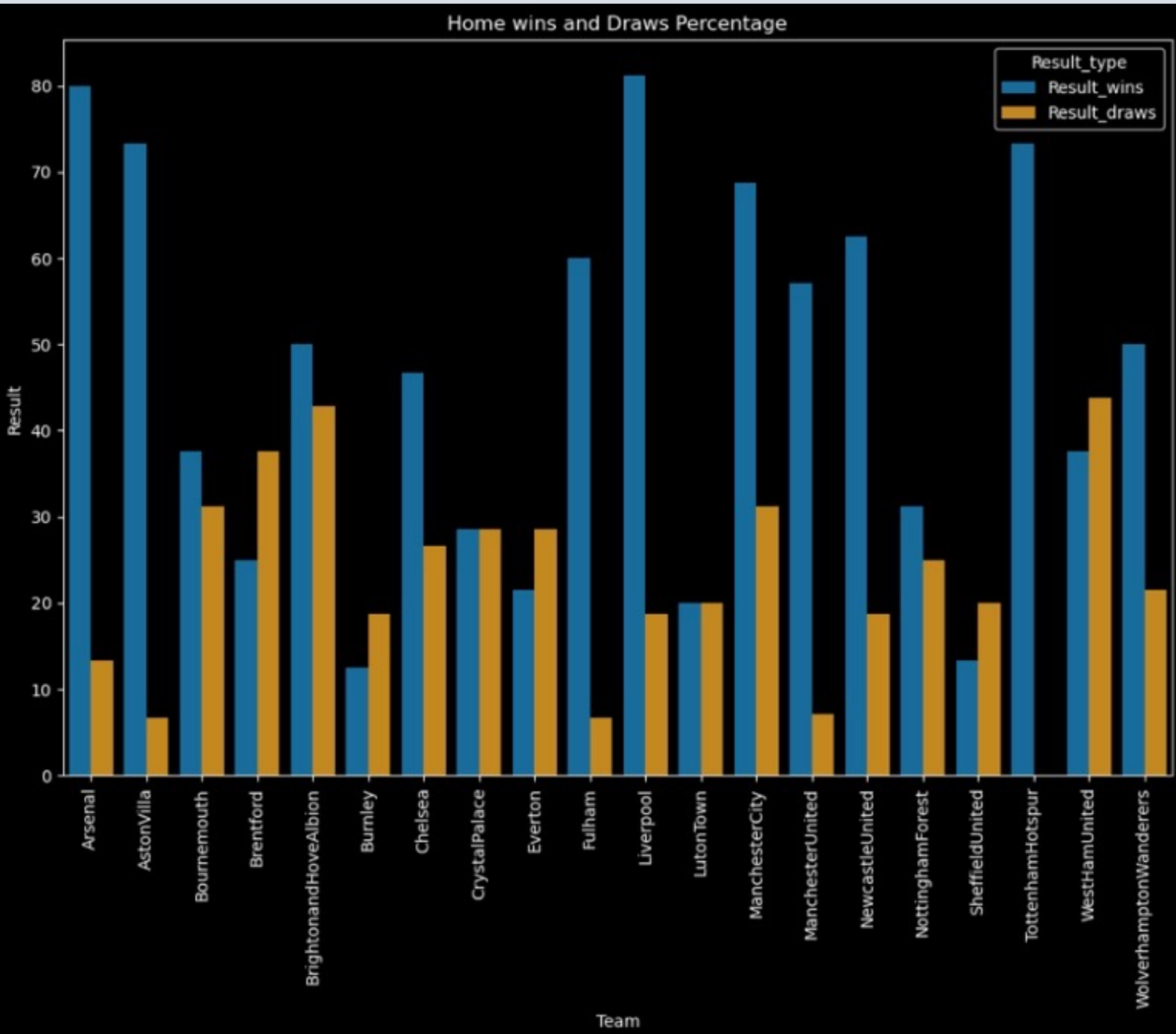

The diagram above shows the percentage of each team's home wins and draws

Statisical Analysis

Teams record against each other – All the times the home team has played the away team and the relevant stats that pertain to each particular meeting.

Possession to shots/shots on target/ expected goals/goals for – I wanted to see how a team's possession relates to their form in front of goal. In theory, a team with more possession should have more shots which leads to more shots on goal which should lead to more expected goals and finally which leads to more goals.
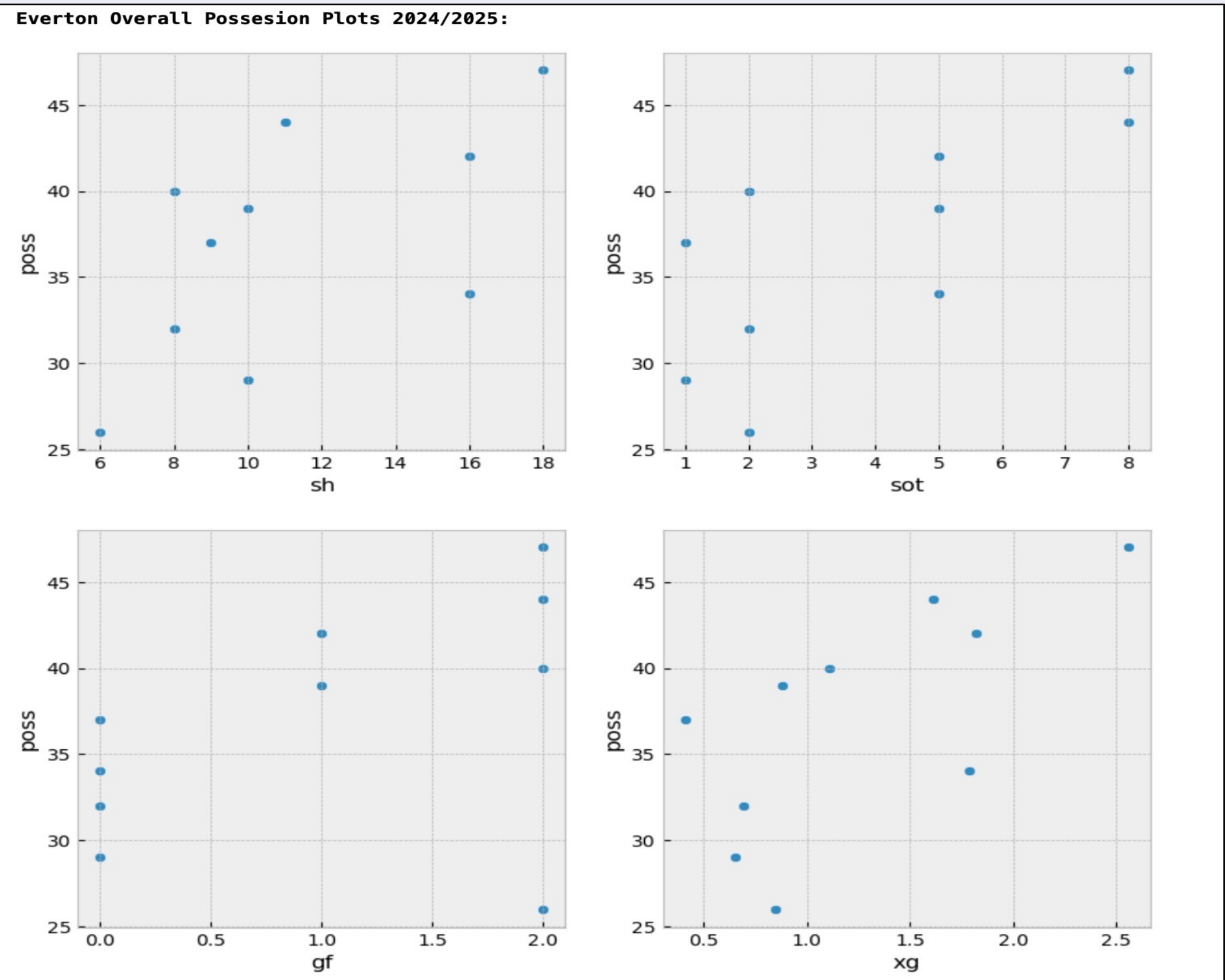
The diagram above shows the shot per goals /xG etc

How a team performs against other opponents (in terms of last year's league position) – who the opponent is plays a huge part in how a team plays. By knowing where the opponent finished each season, I can gauge their level in terms of opposition. The level of opponents are split into four groups of teams.

Shots to Goal Comparisons – This season's stats on percentage of shots on target and percentage that shot on target was a goal.
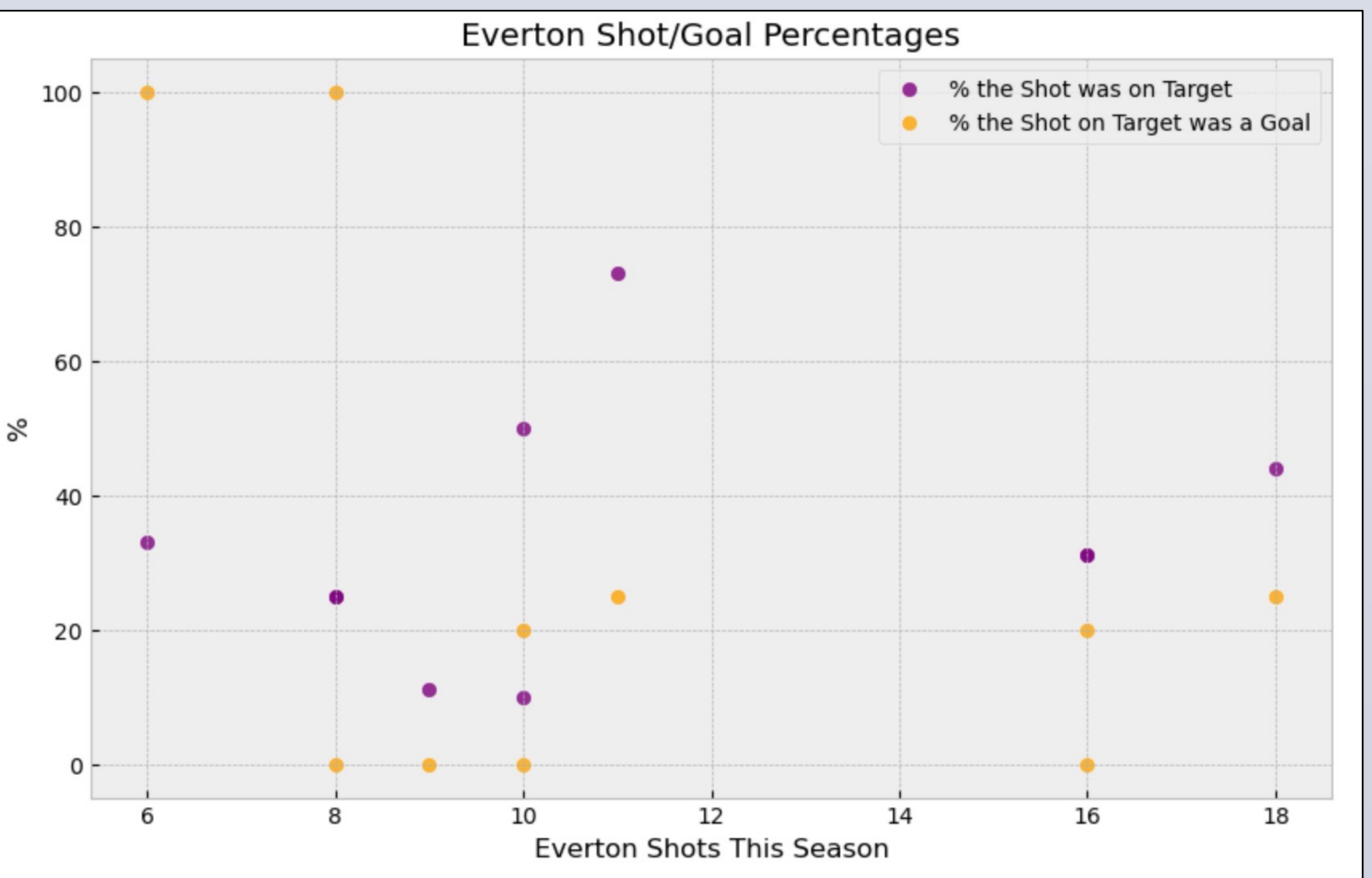

The table above is a plot of the shot – shots on target percentage and shots on goal - goal percentage.

Building Models

Linear or Random Forest

I built and tested a few different models. I tried a Random forest classifier on the result feature and linear/random forest regression task on some numerical features. After figuring what my needs were, I opted for a combination of linear/random forest regression regression on the goals for the home team.
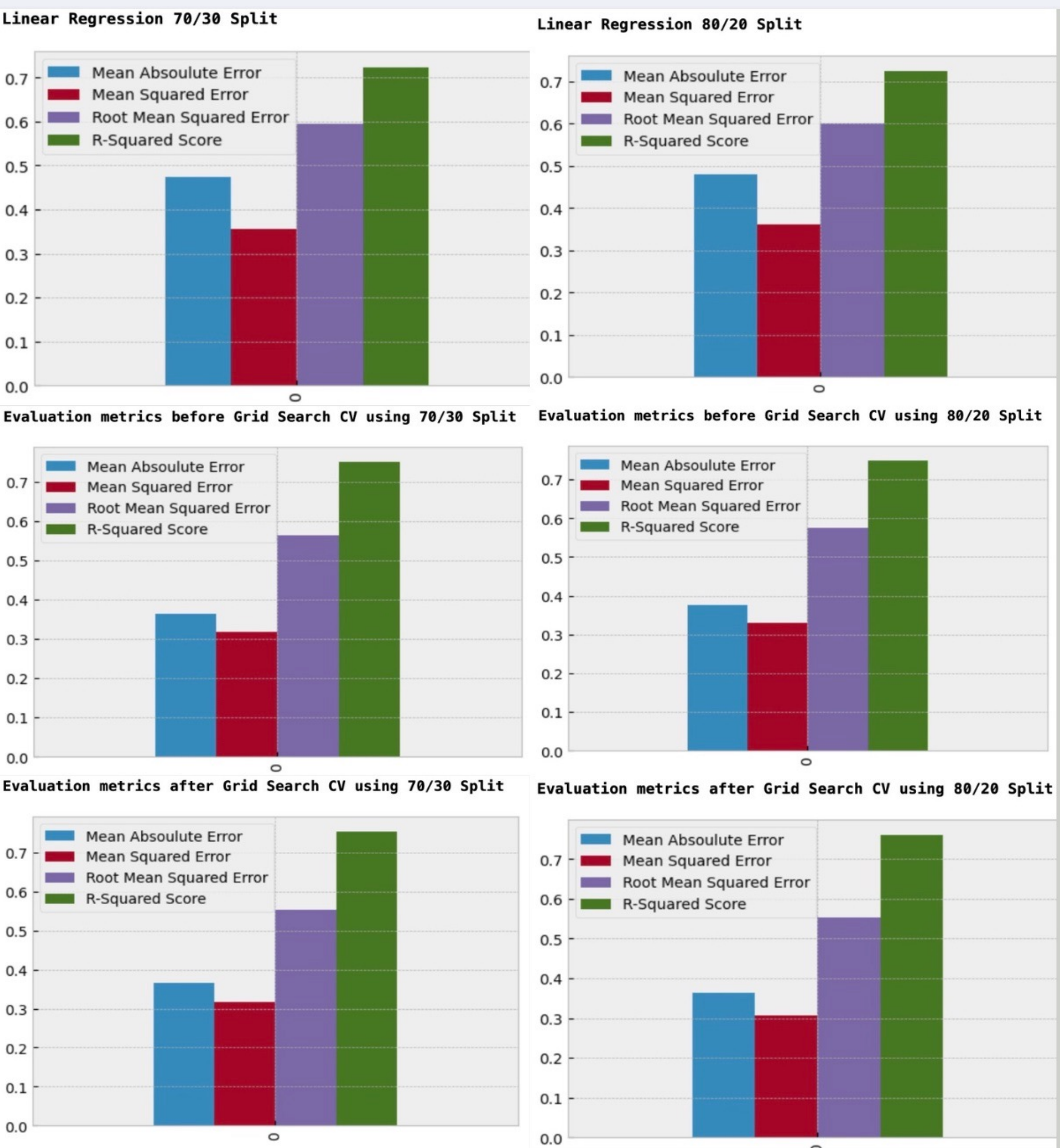
The diagram above shows the evaluation metrics for each model tested on the goals for the home team.

Having looked at the results across all the splits, I can see the results are very close. However, the model with the best performance is the random forest before grid search on a 70/30 split. It has a mean absolute error of 0.36. This means every prediction is out by 0.36 goals from the actual values. The difference between Linear and Random Forest Regressors was 2/3 percent
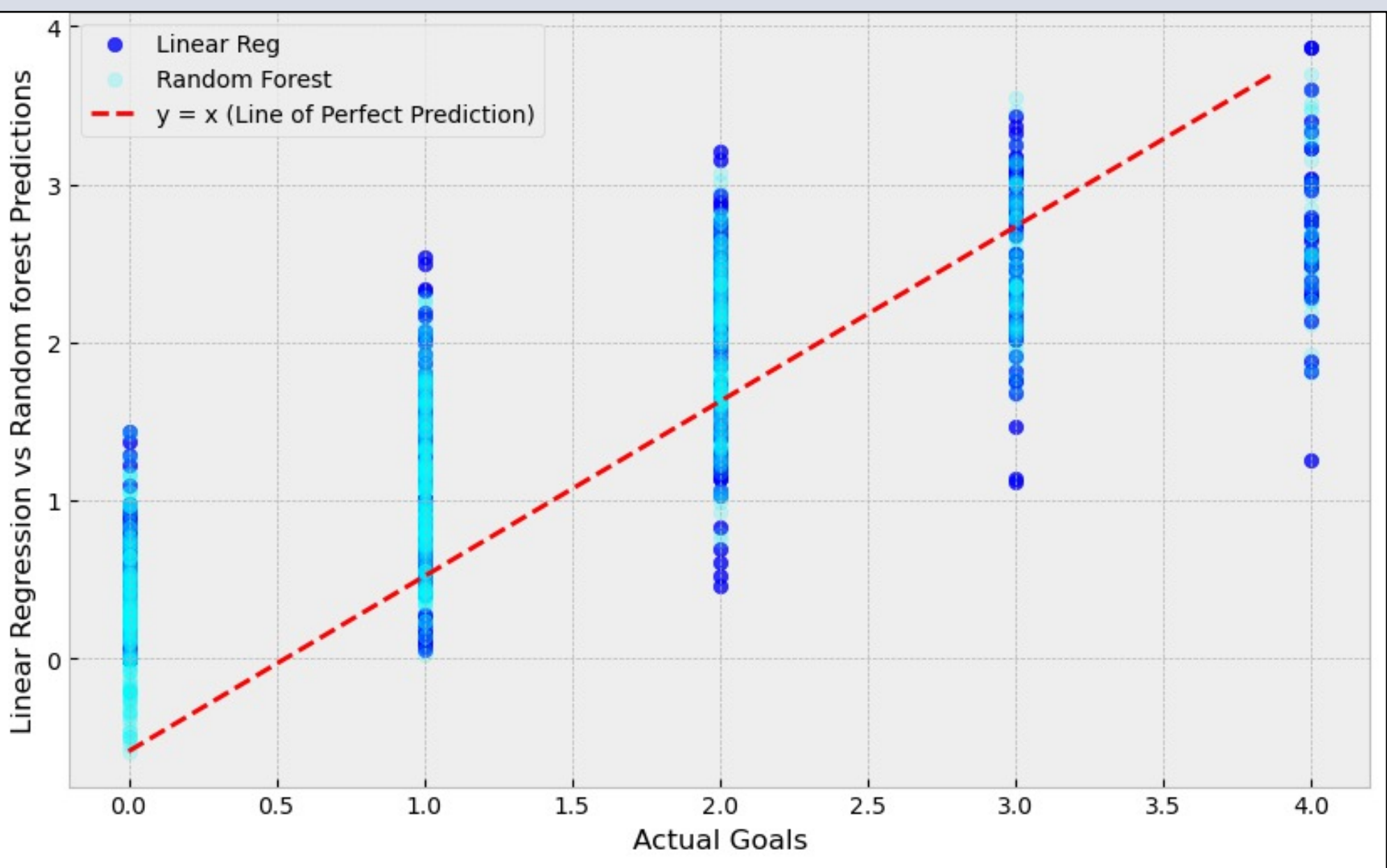
The diagram above shows prediction comparisons for both models

Implementing Predictions Into Betting Strategy

I established that the best markets to bet on for my needs is the following markets:

Home Goals – How many goals home team will score

Away Goals – How many goals away team will score

Total Goals – Total goals per game

Over/Under – Over or under the amount of goals set by the bookies.

Making Predictions and Placing Bets (fictional)

I am strictly only betting on the premier league for now. Therefore, I have 10 games per week to predict on.

HOME Team		AWAY Team	HOME Team	AWAY Team	MY PREDICTION
			Actual Scores		
Southampton	vs	Manchester United	0	3	
Brighton	vs	Ipswich	0	0	
Fulham	vs	West Ham	1	1	1 1
Liverpool	vs	Nottingham Forest	0	1	3 1
Crystal Palace	vs	Leicester City	2	2	
Manchester City	vs	Brentford	2	1	
Aston Villa	vs	Everton	3	2	
Bournemouth	vs	Chelsea	0	2	0 2
Tottenham	vs	Arsenal	0	1	

The diagram above shows the the betting spreadsheet I created in excel to keep track of progress.

HOME TEAM	AWAY TEAM	SCORE	MY PREDICTION	BET/Odds	STAKE	OUTCOME	RETURN
Fulham	West Ham	1-1	1-1	Under 2.5/1.85	2	WIN	5.7
Liverpool	Nottingham Forest	0-1	3-1	Over 3.5/1.6	2	LOSS	-2
Bournemouth	Chelsea	0-2	0-2	Under 2.5/1.8	3	WIN	8.4
Wolves	Newcastle	1-2	1-2	Under 3.5/1.8	3	WIN	8.4
Leicester City	Everton	1-1	1-1	Under 2.5/2.2	2.5	WIN	8
Fulham	Newcastle	3-1	2-1	Under 2.5/1.4	2.5	LOSS	-2.5
Aston Villa	Wolves	3-1	3-1	Over 3.5/2.75	5	WIN	18.75
Everton	Crystal Palace	2-1	2-2	Over 3.5/2.75	2.5	LOSS	-2.5
Brentford	West Ham	1-1	2-1	Under 3.5/1.4	5	LOSS	12
Wolves	Liverpool	3-1	1-2	Under 3.5	2.5	LOSS	2.5
Leicester	Bournemouth	1-0	0-1	Under 1.5/2.6	5	WIN	18
Everton	Newcastle	0-0	0-1	Under 1.5/2.8	5	WIN	19
Ipswich	Everton	0-2	0-1	Under 1.5	5	LOSS	-5
Southampton	Leicester	2-3	2-1	Under 3.5/1.9	5	LOSS	-5
Liverpool	Brighton	2-1	3-1	Under 3.5/2	5	WIN	15
Wolves	Crystal Palace	1-1	1-1	Under 2.5/2.5	5	LOSS	-5

CONCLUSIONS

Answers to Objectives are presented below:

After extensive testing, I identified a market that would best suit the models I produced. I am satisfied that I have built the best models I could've given my current experience and knowledge. That being said, there's always room for improvement and I would like to improve from the 76% accuracy score.

During the testing of the model and strategy building, I noticed some flaws. As I need to have a result feature in the training data, it can lead to some incorrect predictions. However, there is some leeway there as I can tailor the prediction to various bets so the prediction may be off by a goal but that can still fall within a winning bet. This also coincides with the r^2 score of 0.37 for the random forest model.

Overall, I am 38.75 euro in profit (fictionally). I have placed 16 bets. Of which 9 have been winning bets. My win percentage is 56.25%. For now, at least, I have answered the question that data can be used to gain an advantage in sports betting

REFERENCES

Online Sports Betting - Global | Statista Market Forecast [online] [2024] Statista, available: [accessed 28 Mar 2024]

Khan, MASH, (2021) Cricket Based Betting and University Students: A New Income Source or a Curse for University Students in Bangladesh?, preprint, Sociology, available: [accessed 28 Mar 2024]

TryPad (2018), 'HOW TO CALCULATE BOOKIE EDGE' available: [accessed 21 Mar 2024]

Chris Banban-a (2024) 'Sports Betting's Rise Is a Cash Cow. Are States Doing Enough to Curb Gambling Addiction?' [online] CSA TODAY, available: [accessed 22 Mar 2024]

Brent, H. (2024), A Complete Guide to Irish Gambling Laws as a Tourist in 2024 [online], The Irish Post, available: [accessed 25 Mar 2024]

Health Research Board (August, 2023) - Gambling, factsheet.pdf - available: [accessed 26 Mar 2024]

Saylin, Arlens (2024) - Understanding Value Betting in Football | Goal.Com South Africa [online] (2024) available: [accessed 26 Mar 2024]

Yale Medicine (2024) - Gambling Disorder - > Fact Sheets > Yale Medicine [online] available: [accessed 27 May 2024].

Analytics Vidhya, (2024) - One Hot Encoding Using Categorical Data | Analytics available: [accessed 14 May 2024].

R, S C, (2023) 'Understand Random Forest Algorithms With Examples (Updated 2024)', Analytics Vidhya, available: [accessed 14 Mar 2024]