Applying AI to Improve Performance in Soccer

Fernando Perna
Gia Chiem
An Pham
Nicholas Sarbinowski

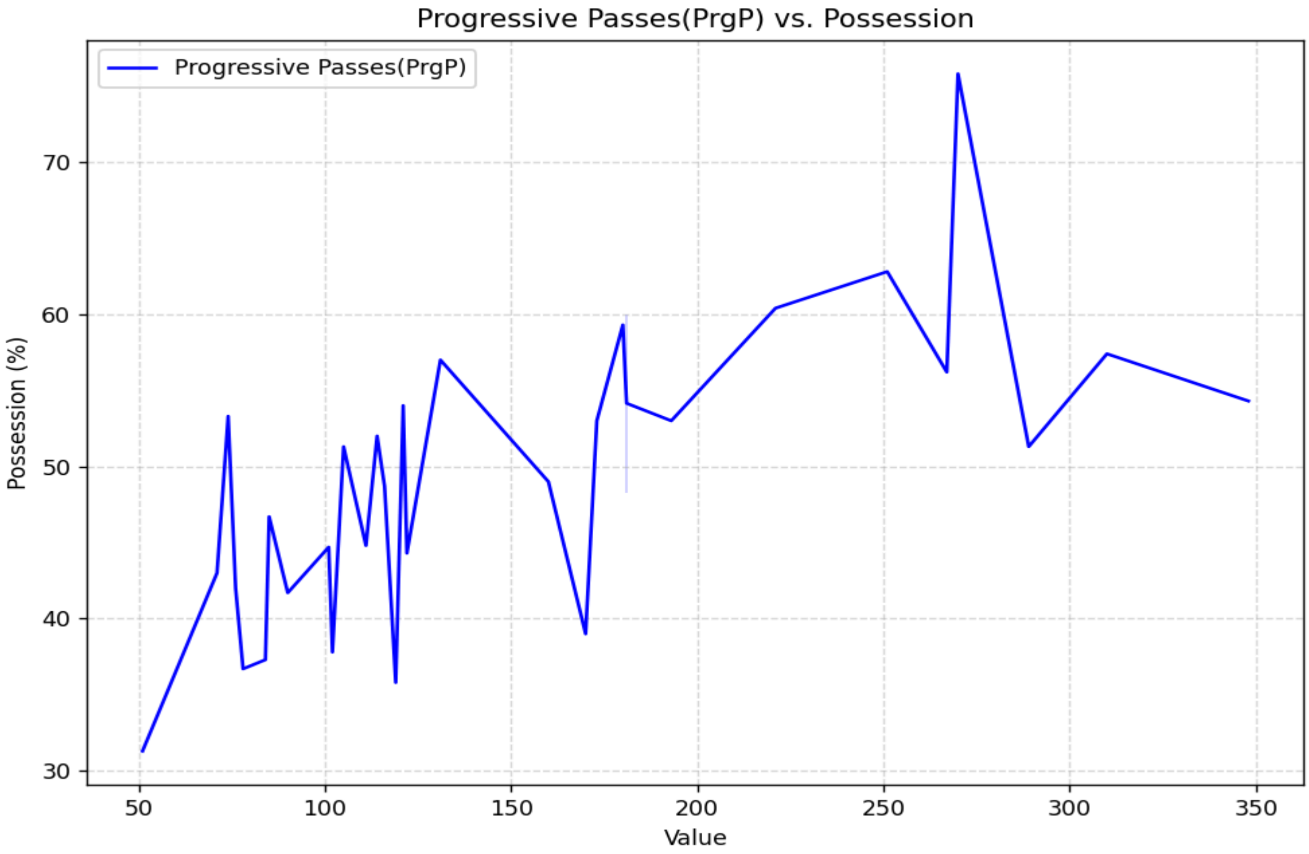
1. Framing the problem: The focus of our project lies within the realm of Supervised Learning, where our dataset comprises labeled training examples. Our objective revolves around predicting possession, thus presenting us with a regression task. Given the limited size of our dataset and the absence of a continuous data flow, we opted for batch learning techniques.

Big Picture: The business objective is to predict the possession each country should have during soccer matches based on various factors. The company could expect to leverage this model to gain insights into team performance and strategize accordingly. By accurately predicting possession, the company could enhance decision-making processes for teams, coaches, and analysts, ultimately improving overall performance on the field. This model's benefits include optimizing gameplay strategies, identifying areas for improvement, and potentially gaining a competitive edge in soccer tournaments.

Description of the Dataset: This dataset encompasses comprehensive statistics of all 32 national teams participating in the 2022 World Cup, covering both the group and knockout stages. It comprises a wide array of data, including fundamental metrics such as player counts, average ages, match tallies, total goals, assists, and more. Moreover, it incorporates advanced metrics like expected goals, progressive, passes, goals, and assists per 90 minutes. Essentially, this dataset offers insights not only into teams' overall performance but also sheds light on the strategies contributing to their placements throughout the tournament.

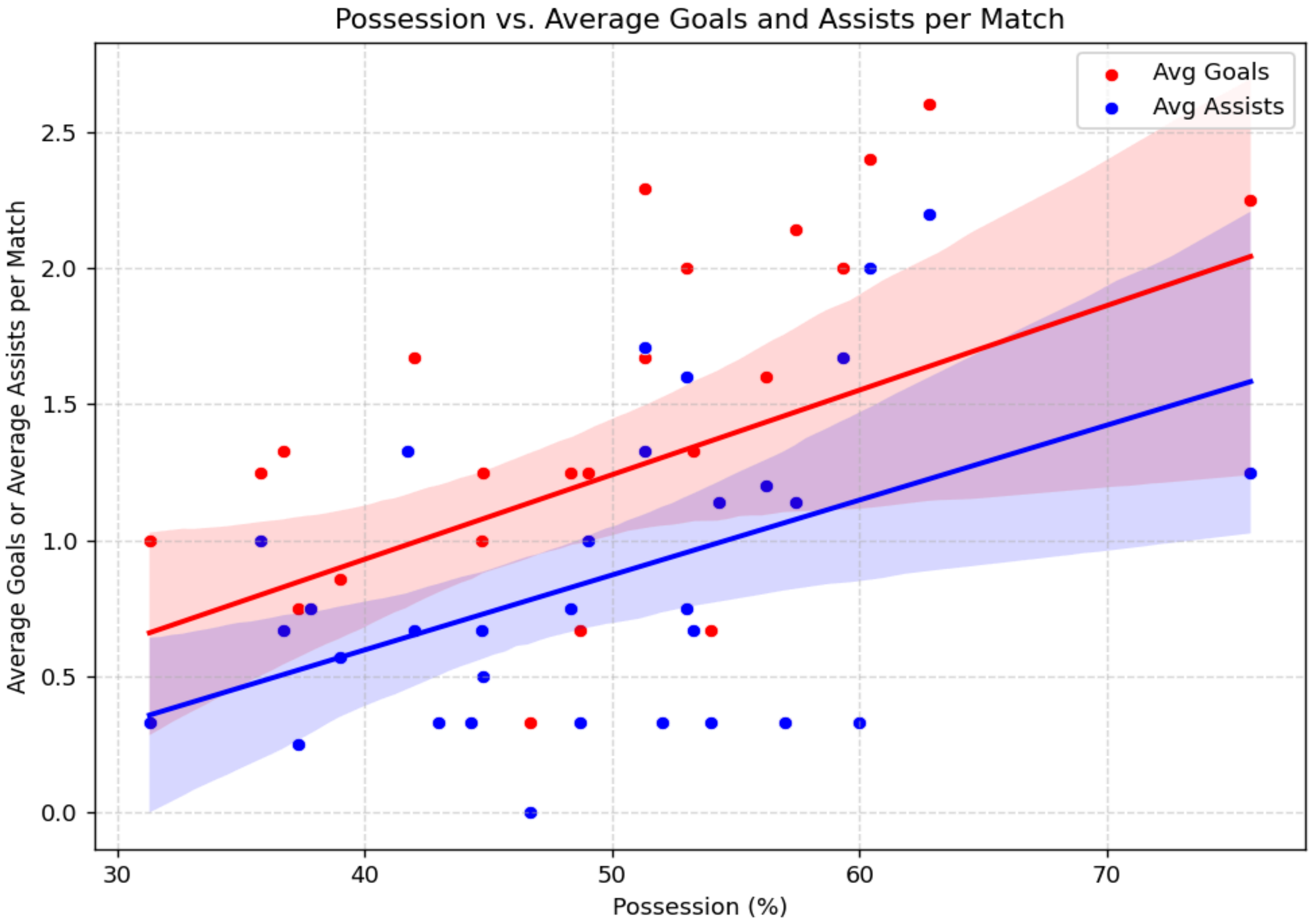
2. Taking a closer look at the data by implementing EDA graphs:

❖ Possession and Progressive Passes



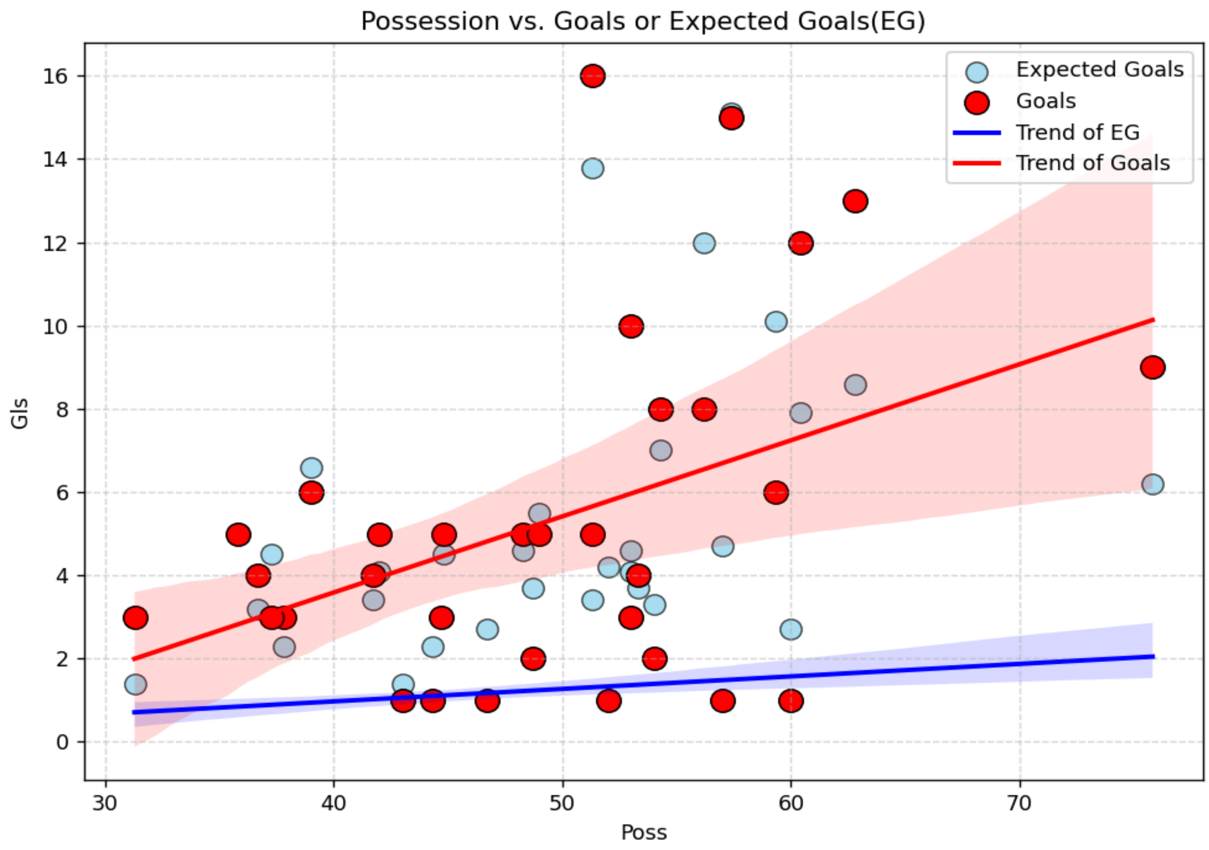
- Completed passes that move the ball towards the opponent's goal line at least 10 yards from its furthest point in the last six passes, or any completed pass into the penalty area. Excludes passes from the defending 40% of the pitch
- Teams with high possession percentages generally have more opportunities to attempt progressive passes.
- When possession is high (e.g: 60% or above) the number of progressive passes tends to increase, as seen in the graph where possession percentages are 60 percentage points the corresponding number of progressive passes are relatively high (eg: 180, 251, 270)

- ❖ Possession vs. Average Goals and Average Assists per Match



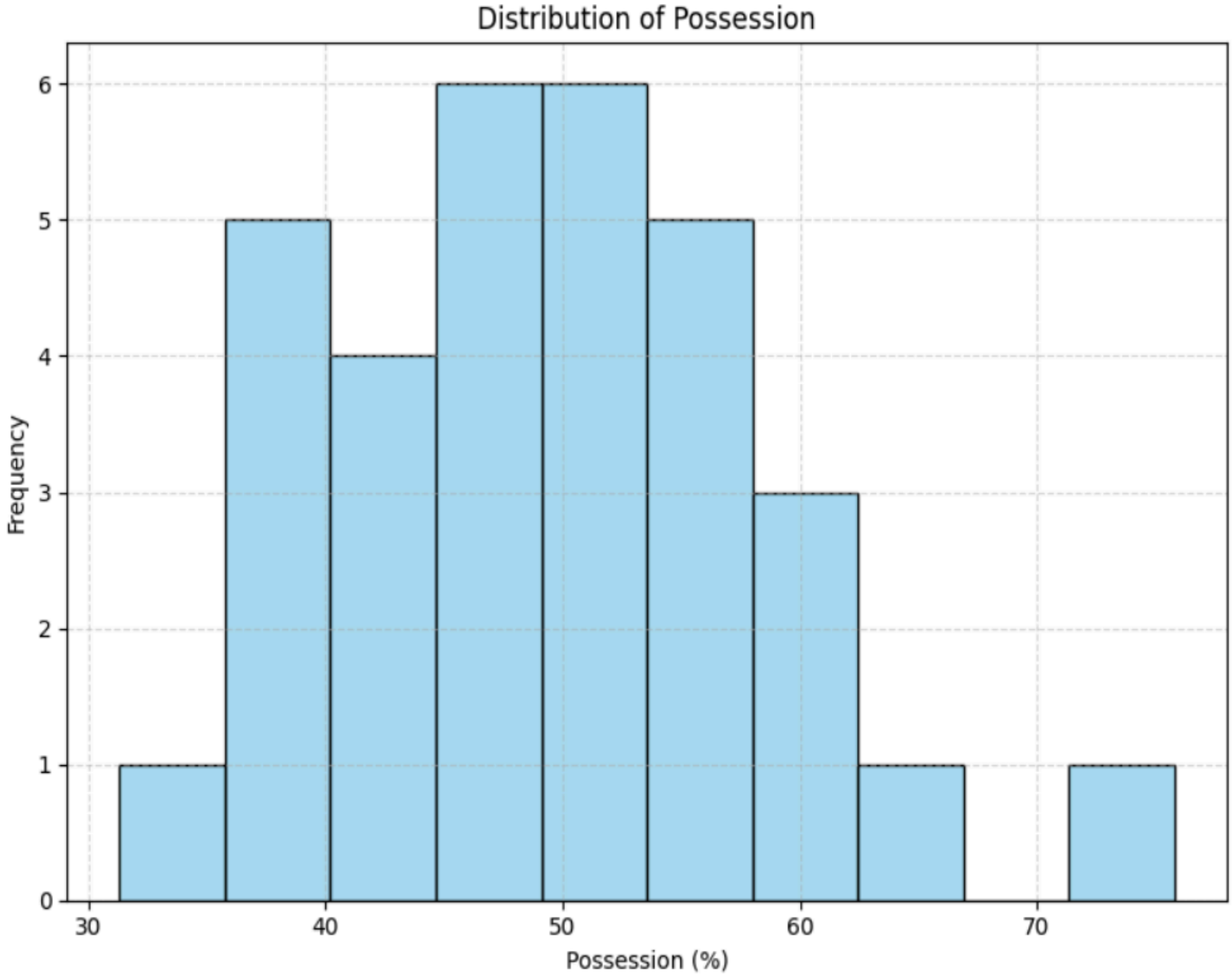
- The graph can be divided into three groups:
  - Low percentage of possession (between 30% to 45%) group and wins average number of goals and assists per match
  - High percentage of possession group and having less assists (under 0.5 assist per match)
  - High percentage of possession (between 45% to 60%) group tends to won high number of goals and assist (over 1.25 goals and over 1 assist per match)

- Possession Vs. Goal(Gls) and Expected Goal(Eg)



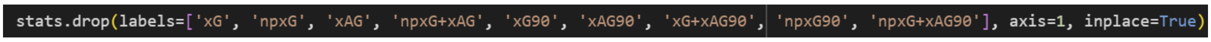
- Possession (between 45% to 55%) accounts for the most number of expected goals ranging from 2 to 6 goals for the tournament
- Teams has the most goals having possession in 50% to 60% which can imply this range of possession is ideal
- An exception for one over 75% of possession is in the sixth of the most scored team

- Distribution of Possession (Poss)



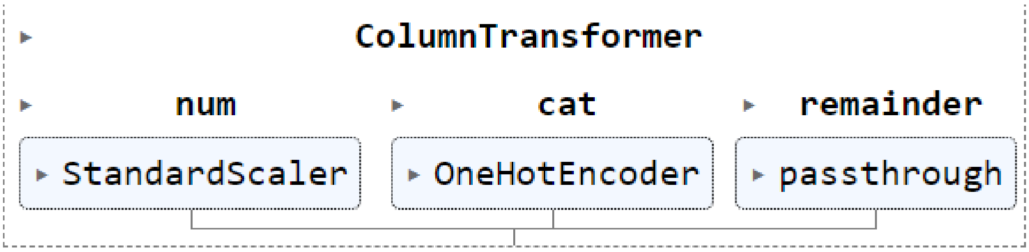
- Possession can be measured as a percentage of the total time each team spends with the ball.
- Possession ranging from 45 to 55 percentages accounted for over one third of the total number of teams which indicates that has the opportunity to create scoring chances while denying the opposing team the opportunity to do so.

3. **Data cleaning and preprocessing:** During the data cleaning and processing phase, we discovered that our dataset was devoid of duplicate rows, columns, or missing values, obviating the need for any removal or imputation operations. Nevertheless, we opted to discard several columns from the dataset as they were deemed irrelevant for predicting each country's possession, being predictive in nature themselves.

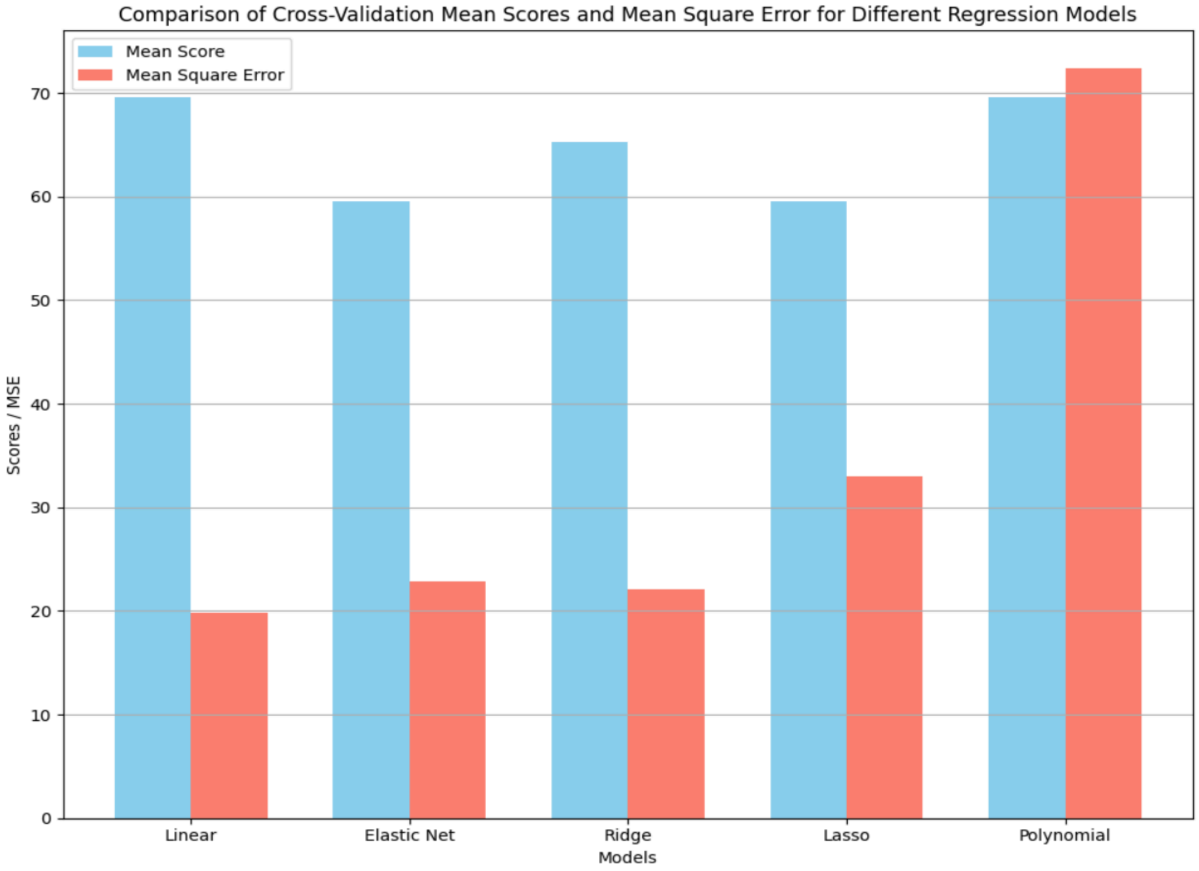


```
stats.drop(labels=['xG', 'npxG', 'xAG', 'npxG+xAG', 'xG90', 'xAG90', 'xG+xAG90', 'npxG90', 'npxG+xAG90'], axis=1, inplace=True)
```

Subsequently, we devised a preprocessing pipeline to encode categorical data utilizing OneHotEncoder and scale numerical data using StandardScaler. This pipeline was then applied to the dataset, following which the column headers were reinstated.

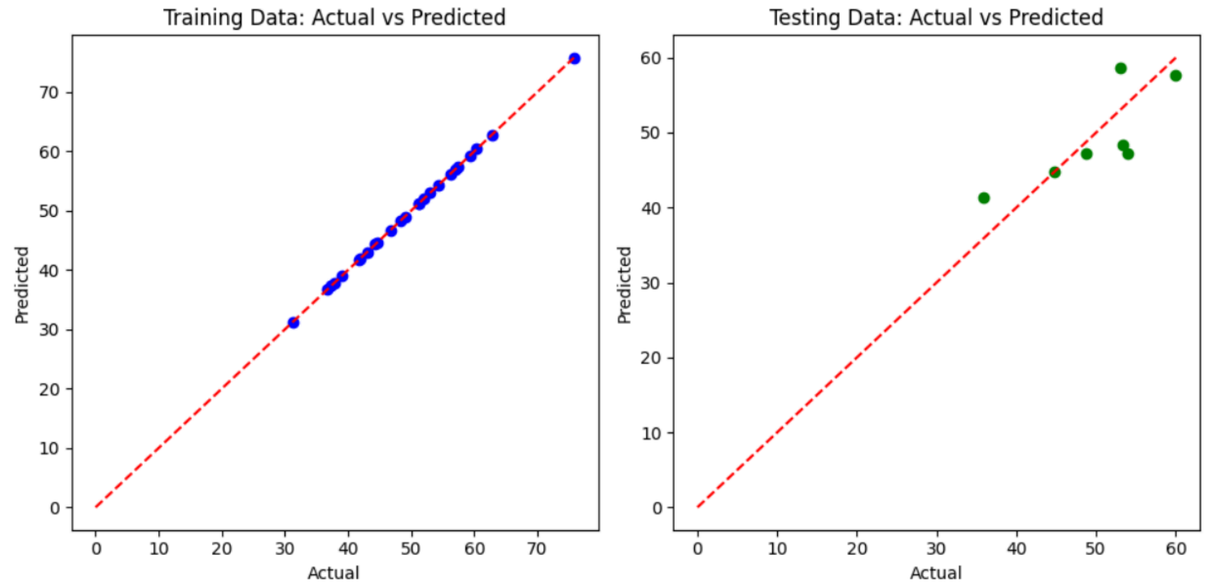


4. **Train and evaluate three machine learning models. Discuss and analyze findings and compare the results of the three models:**
We have conducted testing and training using a variety of models. Specifically, we evaluated our dataset with five regression models: linear, ridge, lasso, elastic net, and polynomial. The rationale behind employing multiple models stems from the limited size of our dataset and the relatively weak correlation between our primary training variable, possession, and other features. Upon analysis, we observed that the linear, ridge, and elastic net models exhibited superior performance compared to the others, with the linear model demonstrating the lowest error rate. These three models, especially the linear regression model, displayed the lowest mean square error values, which means the regression line is closest to all the data points and achieved the highest cross-validation, which provides the highest predictive accuracy. This suggests that they generalize well to unseen data and are robust across different samples. As a result, we can see that the linear model provides the best algorithm.

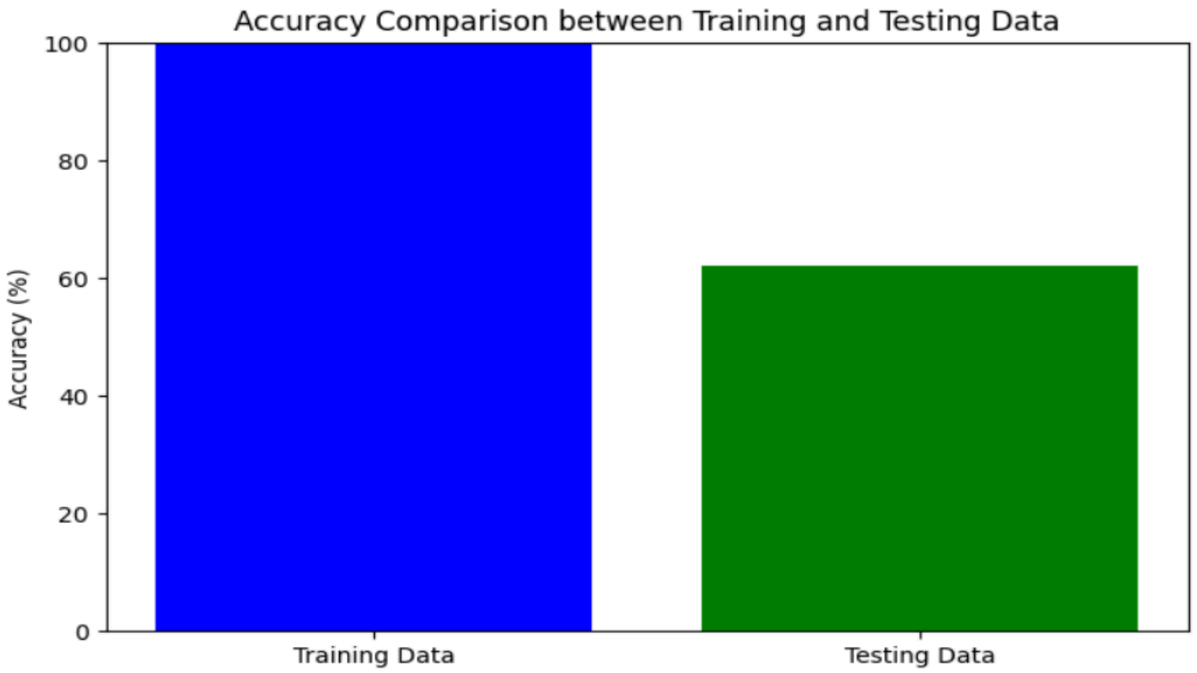


- The graph shows that although Linear algorithm has the same Mean Score as Polynomial, it has the significantly small Mean Square Error. As a result, Linear Model is considered as the best performance algorithm.

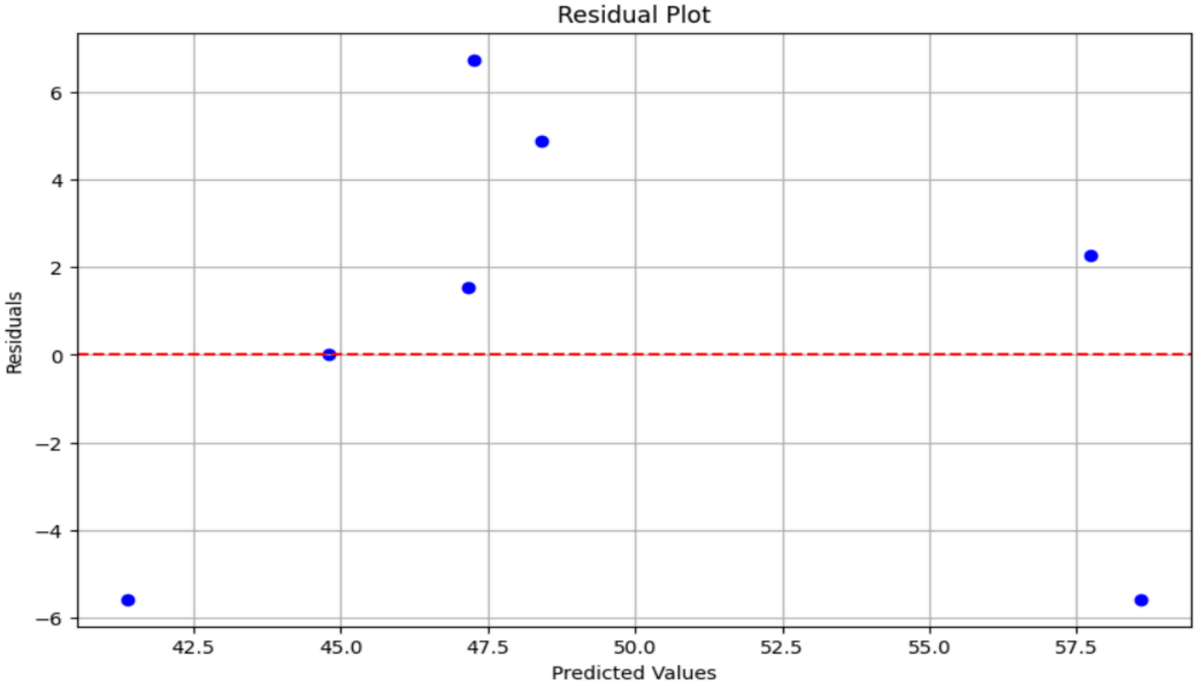
5. **Graphs for the best performing algorithm (Linear Regression).**



- Training Data: all of the dots appear on the regression line which mean the training data is accurate and consistent
- Testing Data: Only one dot appears on the regression line and some dots partially closed to the line which indicates the model is less accurate and there may be some room for error.



- The graph shows accuracy between training and testing data: the training data is 100% accurate and testing data has 60% accuracy which means it is necessary for improvements .



- The red dashed line at y=0 indicates where the residuals would be if the predictions were perfect.
- The average error of the Linear Regression model is 3.8 which indicates there are slight differences in prediction and actual value. However, the results can be considered as acceptable

6. **Limitations Faced:**

Large Amount of Predicted Statistics: The dataset contained a large amount of pre-calculated statistics, constraining our ability to identify a suitable target variable for our prediction model.

Selection of Optimal Training Model: The process of determining the most suitable machine learning algorithms for our prediction task presented a significant challenge. We encountered the difficulty of selecting from a range of models, each with its strengths and weaknesses.

Sparse Correlation with Possession: Despite extensive data exploration, we found a scarcity of features exhibiting strong correlations with possession. This lack of strong correlations hindered our ability to build a predictive model with satisfactory accuracy. Consequently, during testing, we observed higher mean standard errors, highlighting the challenge of predicting possession accurately with the available data.

7. **Next steps:**

After completing the goal of predicting possession for each team and comparing it with the actual values, three potential next steps and future plans could be considered to further enhance the project:

Real-Time Prediction: Develop a real-time prediction system that provides live updates on possession probabilities during matches. This could be valuable for coaches, analysts, and fans alike, offering actionable insights and enhancing the viewing experience.

Cross-Sport Analysis: Apply the techniques and methodologies developed in this project to other sports or competitions beyond soccer. Exploring possession dynamics in different contexts could yield valuable insights and contribute to a broader understanding of game strategies and performance metrics.

Collaboration with Stakeholders: Collaborate with soccer teams, coaches, analysts, or sports organizations to integrate the possession prediction model into their decision-making processes. This could involve customizing the model to specific team strategies or player profiles, ultimately contributing to improved performance on the field.

Overall, the next steps would revolve around continuous iteration, exploration of new features and techniques, and collaboration with stakeholders to leverage the predictive insights for practical applications in soccer analytics.

8. **Appendix 1:**
Github link: https://github.com/chiemgiabaost/Project_group_12
Dataset Source: https://www.kaggle.com/datasets/greysonmb/2022-world-cup-stats

9. **Appendix 2:** Source code can be found in the following file:
https://github.com/chiemgiabaost/Project_group_12/blob/main/worldcup_src.py