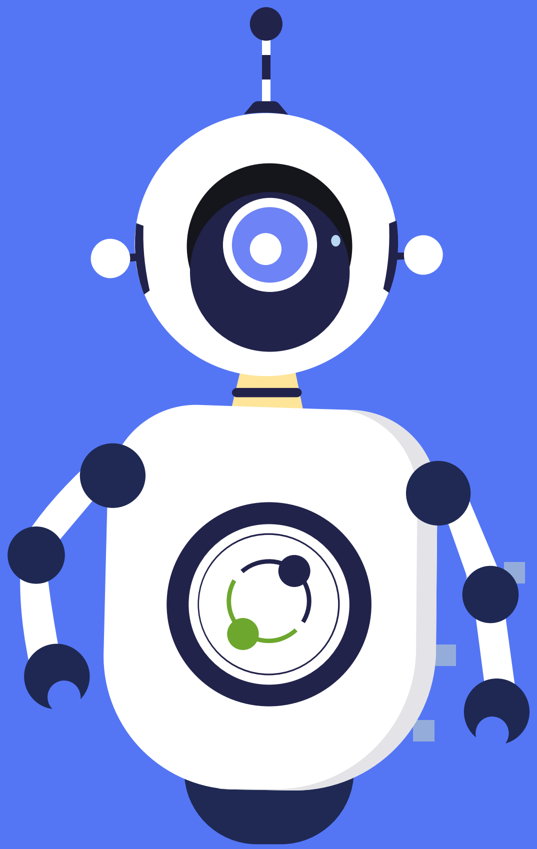
Predictive Analytics in Football
- Predicting Top Goal Scorers -



OR-568 (Prof Vadim Sokolov)

Chanyoung Park

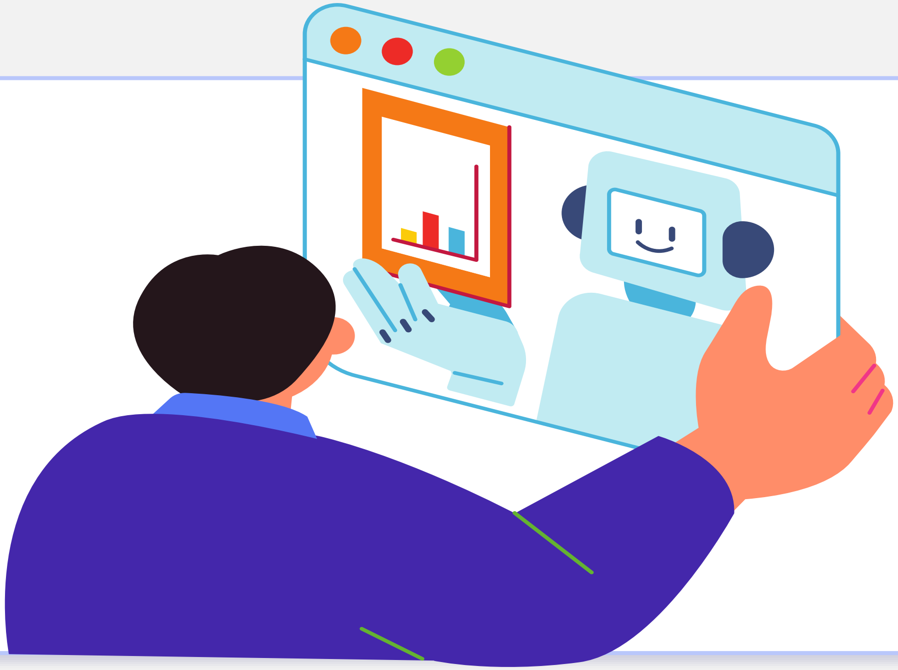
Hussain Alsharif

Oluwasegun Adegoke

Introduction

- Predictive analytics, a branch of advanced analytics, leverages data to make predictions about future events. In the world of sports, this translates to using player statistics, game results, and other relevant metrics to forecast future performances and outcomes.

- In football, where player performance can be quantified in numerous ways, predictive analytics becomes a game-changer. It not only enhances our understanding of the game but also helps in strategic planning, player assessments, and predicting future stars.



Our Goal

Our objective is to investigate whether the performance of football players in the first half of the season can be leveraged to predict the highest goal scorer by the end of the season.

Reason of selecting this topic

- We are avid fans of the Premier League, deeply engaged with its teams and players

- The Premier League is a common thread that unites all our group members.

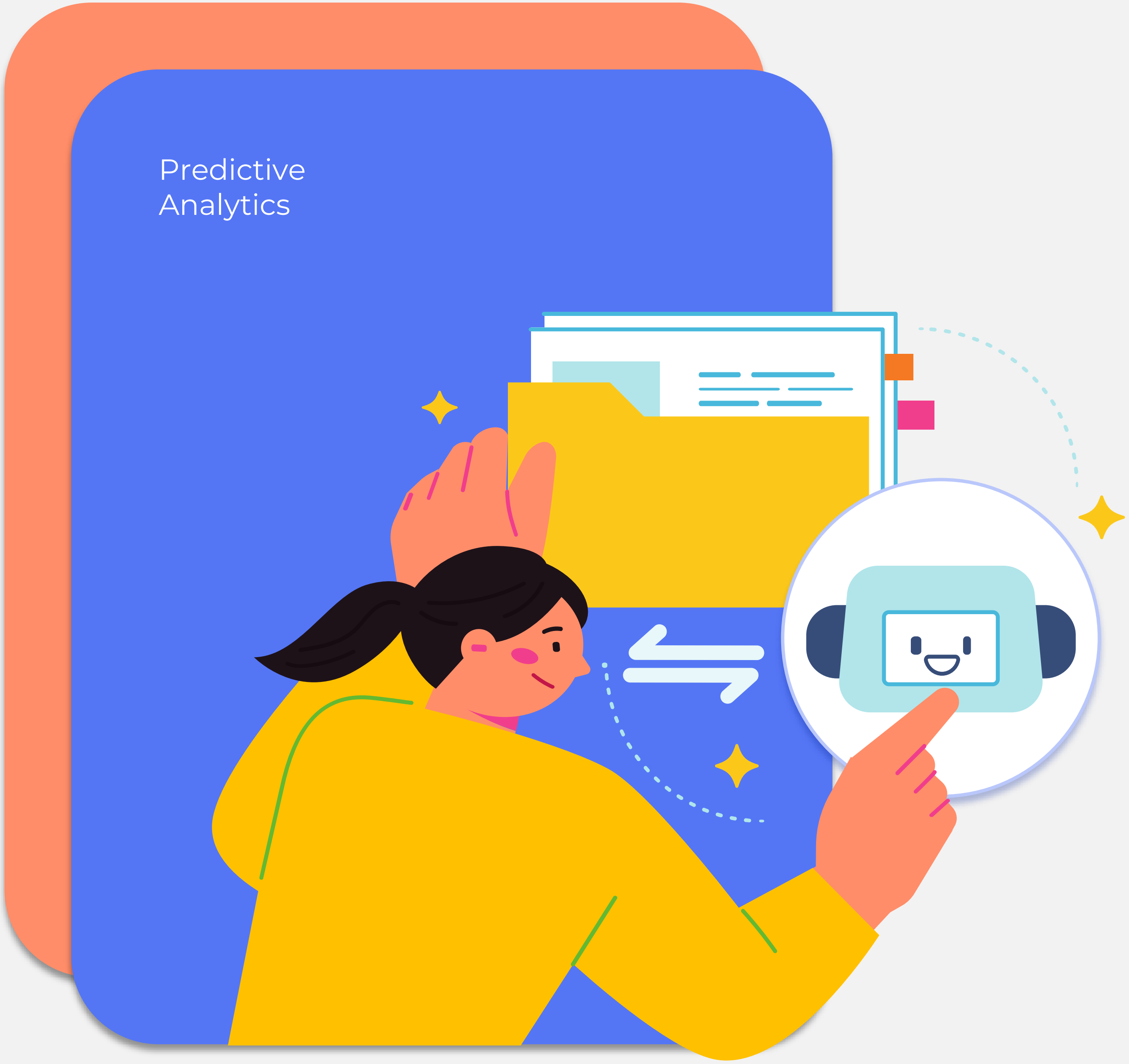
- We believe the Premier League stands out as the premier football league globally, famous for its competitive spirit and talented players.

Expectation of project

-Our study aims to show predictive analytics' usefulness in football, illustrating data's role in sports management and betting decisions.

-We anticipate challenges, especially with sports' inherent variability, which may result in low accuracy bcos of unseen variable like injuries

-We aim to offer practical insights and advice from our results, benefiting teams, coaches, and analysts.



About Data

Dataset was scraped from kickest

- We have carefully selected a dataset that aligns with our research question and can provide us with the necessary information to predict the top goal scorer based on mid-season data.
- We combined 50% of the 2019/2020, 2020/2021 and 2021/2022 season with their goals at the end of the season resulting 1442 observations and 26 features.
- It includes detailed metrics such as goals, assists, shots, passes, and other relevant in-game statistics for each player.
- This robust dataset serves as the backbone for our predictive models, offering a nuanced view of player performances in the league.

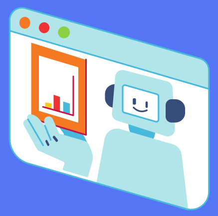
Dataset

	Player	Pos	Team	PTS	CR	Plus	Apps	Starter	Mins	Goals_x	Shots	On Tar. Shots	Pen Goals	Successful Dribbles	Ast	Acc Pass	Key Pass	Fouls	Was Fouled	YC	RC	Rec Ball	Tackles	Clean Sheets	Saves	Goals_y	top_10_scorer
0	B. Chilwell	Defender	CHE	35.38	15.4	0.9	6	6	540	0.5	2.17	1.0	0.0	0.5	0.17	37.17	1.67	1.67	2.5	0.0	0.0	1.17	1.33	0.67	0.0	3	No
1	T. Alexander-Arnold	Defender	LIV	32.66	20.2	3.5	17	17	1504	0.12	1.65	0.59	0.0	0.47	0.47	53.29	3.12	0.29	0.29	0.06	0.0	2.29	1.29	0.59	0.0	2	No
2	Joao Cancelo	Defender	MCI	31.33	21.9	4.3	18	18	1608	0.06	2.22	0.83	0.0	1.22	0.22	72.17	1.11	0.89	0.44	0.28	0.0	3.28	2.0	0.56	0.0	1	No
3	M. Sarr	Defender	CHE	30.4	5.1	-0.4	1	1	90	0.0	0.0	0.0	0.0	0.0	0.0	57.0	0.0	0.0	0.0	0.0	0.0	7.0	3.0	1.0	0.0	0	No
4	Mohamed Salah	Attacker	LIV	30.09	19.3	-0.5	19	19	1693	0.89	4.05	1.95	0.21	1.63	0.53	27.79	2.0	0.32	0.74	0.05	0.0	0.42	0.47	0.58	0.0	23	Yes
5	L. Diaz	Midfielder	LIV	26.7	15.5	0.0	1	1	85	0.0	4.0	2.0	0.0	6.0	0.0	43.0	2.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	4	No
6	A. Robertson	Defender	LIV	26.49	19.4	1.5	15	15	1307	0.07	0.4	0.27	0.0	0.27	0.4	53.67	2.0	0.8	0.27	0.27	0.07	1.47	1.07	0.47	0.0	3	No
7	Aymeric Laporte	Defender	MCI	26.43	19.0	2.3	15	15	1273	0.13	1.47	0.33	0.0	0.2	0.0	82.2	0.27	0.47	0.27	0.33	0.07	1.27	1.07	0.6	0.0	4	No
8	V. van Dijk	Defender	LIV	26.38	18.2	0.9	17	17	1530	0.12	1.06	0.29	0.0	0.18	0.06	71.0	0.35	0.29	0.18	0.12	0.0	1.29	0.47	0.65	0.0	3	No
9	R. James	Defender	CHE	26.26	15.6	2.7	15	13	1088	0.27	1.4	0.47	0.0	1.13	0.33	42.33	2.0	0.73	1.07	0.27	0.07	1.8	1.27	0.33	0.0	5	No
10	O. Zinchenko	Defender	MCI	26.23	14.1	0.5	7	5	506	0.0	0.71	0.14	0.0	0.43	0.14	57.14	0.57	0.71	0.29	0.0	0.0	2.57	1.43	0.43	0.0	0	No
11	J. Matip	Defender	LIV	25.42	15.5	2.7	15	15	1350	0.07	1.13	0.27	0.0	0.33	0.0	67.73	0.27	0.4	0.4	0.0	0.0	2.2	1.33	0.53	0.0	3	No
12	Bernardo Silva	Midfielder	MCI	25.34	16.0	2.3	18	18	1524	0.39	1.28	0.78	0.0	1.67	0.06	49.61	1.61	0.83	1.06	0.22	0.0	1.39	1.72	0.5	0.0	8	No
13	W. Boly	Defender	WOL	25.0	7.9	-1.8	1	1	90	0.0	0.0	0.0	0.0	0.0	0.0	74.0	0.0	1.0	0.0	0.0	0.0	2.0	1.0	1.0	0.0	0	No
14	J. Stones	Defender	MCI	24.9	13.5	-0.1	6	4	411	0.17	0.5	0.33	0.0	0.17	0.0	48.17	0.0	0.17	0.67	0.0	0.0	1.17	0.5	0.67	0.0	1	No
15	Ruben Dias	Defender	MCI	24.49	18.1	1.0	18	17	1503	0.11	0.5	0.11	0.0	0.06	0.11	72.72	0.56	1.06	0.22	0.22	0.0	1.67	1.17	0.5	0.0	2	No
16	C. Kelleher	Goalkeeper	LIV	24.3	9.9	-1.5	1	1	90	0.0	0.0	0.0	0.0	0.0	0.0	36.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	0	No
17	C. Gallagher	Midfielder	CRY	23.97	10.2	5.7	18	18	1574	0.39	1.94	0.83	0.0	1.28	0.17	27.94	1.72	1.78	1.78	0.22	0.0	2.28	2.67	0.22	0.0	8	No
18	Rodri	Midfielder	MCI	23.92	18.5	2.0	16	16	1384	0.13	1.0	0.31	0.0	0.56	0.06	76.25	1.0	1.5	0.56	0.25	0.0	2.0	2.44	0.56	0.0	7	No
19	K. Walker	Defender	MCI	23.88	15.6	1.5	12	12	1035	0.0	0.42	0.08	0.0	0.42	0.08	69.5	0.5	0.25	0.75	0.08	0.0	0.75	0.67	0.5	0.0	0	No
20	Thiago Silva	Defender	CHE	23.51	16.8	1.8	16	13	1283	0.13	0.63	0.25	0.0	0.19	0.0	69.19	0.25	0.25	0.31	0.06	0.0	2.19	1.13	0.38	0.0	3	No
21	E. Smith Rowe	Midfielder	ARS	23.42	11.0	4.0	17	13	1186	0.47	1.47	0.94	0.0	1.35	0.12	23.88	1.18	0.18	0.71	0.0	0.0	0.76	0.41	0.29	0.0	10	No
22	M. Cornet	Defender	BRN	23.1	10.3	1.6	12	11	834	0.5	1.67	1.08	0.0	0.67	0.08	9.83	0.75	0.58	0.58	0.17	0.0	0.5	0.33	0.17	0.0	9	No
23	Jonny Castro	Defender	WOL	23.1	8.9	-1.8	1	1	82	0.0	0.0	0.0	0.0	0.0	0.0	42.0	1.0	1.0	3.0	0.0	0.0	1.0	0.0	1.0	0.0	2	No
24	P. Hojbjerg	Midfielder	TOT	22.71	14.2	2.1	18	18	1608	0.11	0.94	0.44	0.0	1.39	0.06	62.61	0.56	0.89	0.89	0.06	0.0	3.06	2.56	0.44	0.0	2	No
25	A. Rudiger	Defender	CHE	22.62	16.8	1.7	18	18	1620	0.11	1.39	0.33	0.0	0.17	0.0	61.61	0.56	1.06	0.33	0.17	0.0	1.89	1.28	0.44	0.0	3	No
26	M. Mount	Midfielder	CHE	22.53	16.4	0.6	16	12	1066	0.44	2.06	1.06	0.06	0.5	0.25	25.38	1.44	0.63	0.56	0.06	0.0	1.06	1.25	0.31	0.0	11	No
27	M. Kovacic	Midfielder	CHE	22.46	13.1	0.6	11	8	757	0.09	1.09	0.27	0.0	1.82	0.45	45.82	1.27	1.0	0.82	0.09	0.0	1.73	2.55	0.55	0.0	2	No
28	P. Foden	Midfielder	MCI	22.42	16.6	-0.4	12	9	831	0.33	2.08	1.0	0.0	0.83	0.25	31.08	1.75	0.5	0.58	0.0	0.0	0.5	0.33	0.33	0.0	9	No
29	Y. Tielemans	Midfielder	LEI	22.41	14.5	0.8	15	15	1318	0.33	2.13	0.67	0.07	0.73	0.13	48.6	1.6	1.07	1.2	0.07	0.0	1.87	1.73	0.13	0.0	6	No
30	C. Jones	Midfielder	LIV	22.03	8.1	-0.3	6	4	399	0.17	1.67	0.67	0.0	2.0	0.17	43.67	0.67	0.83	1.67	0.0	0.0	0.83	0.83	0.33	0.0	1	No

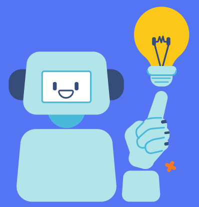
Core Methods

With predictive analysis, it is advised to use different approaches. Therefore, different models were applied in this project:

Linear Regression, Random Forest, Boosting, and Deep Learning. Each offers a unique approach to predictive analysis, allowing us to compare and contrast their predictive capabilities.

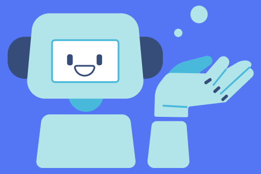


1. Linear Model



2. Random Forest

3. Boosting model



4. Deep Learning

Linear Model

Model 1: Kitchen sink

- R-Squared: 0.8062 (80.62% of the variance explained)

-RMSE: 1.5599 (Indicates the typical prediction error)

-# P-value predictors < 0.1 = Ten

-MAE : 1.095538

-Intercept: -1.450

Model 2: Interactions

-R-Squared: 0.9342 *(93.42% of the variance explained)*

-RMSE: 1.471316 *(Lowest error, indicating highest prediction accuracy)*

-# P-value predictors < 0.1 = 46

-MAE: 0.9827922

- Intercept: –0.1958

Model 3: P-value Hack

-R-Squared: 0.9336 *(93.36% of the variance explained)*

-RMSE: 1.475439 *(Reduced error compared to Model 1)*

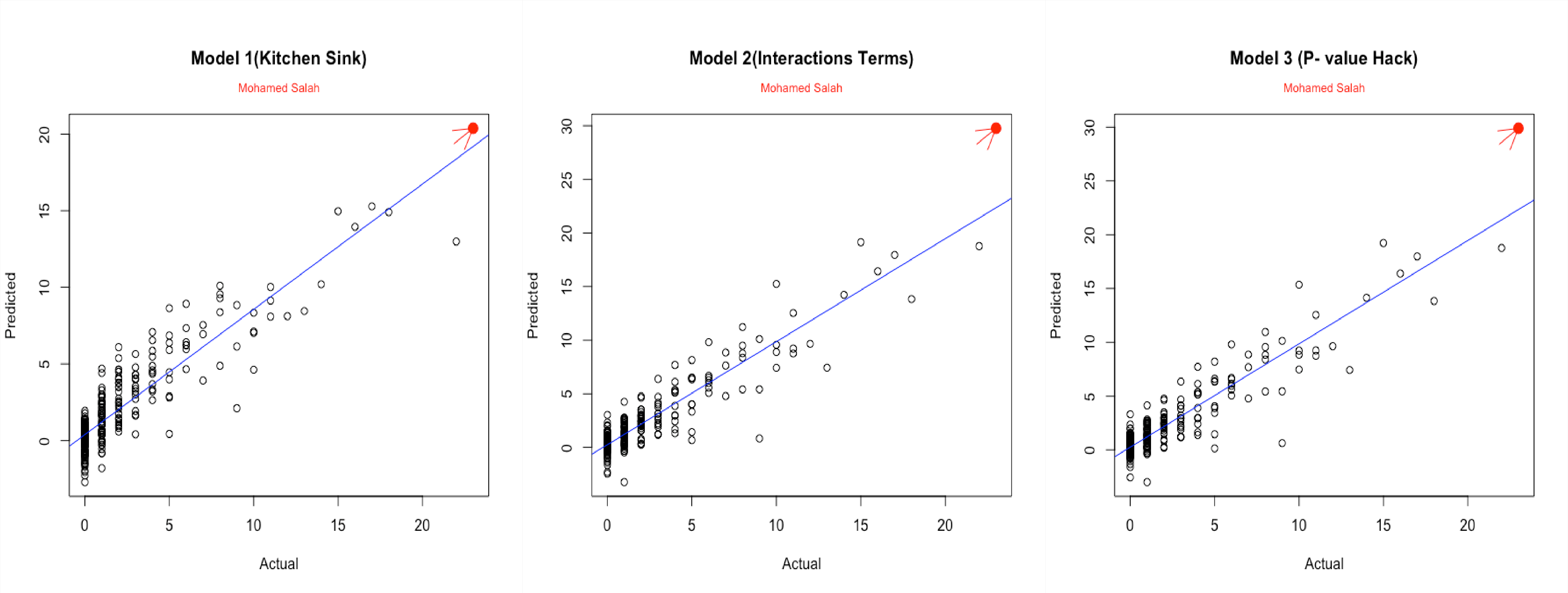
- Used only the 46 significant predictors in model 2

-MAE: 0.9757574

- Intercept: -0.06101

- The second model showed a Prediction Accuracy with an RMSE value of 1.47 on the test dataset, indicating high predictive accuracy."
- It successfully predicted Mohamed Salah as the top scorer, aligning with the actual outcome.
- We tried standardize the dataset but had no effect on the models

Linear Model



Random Forest Model

Performance Metrics:

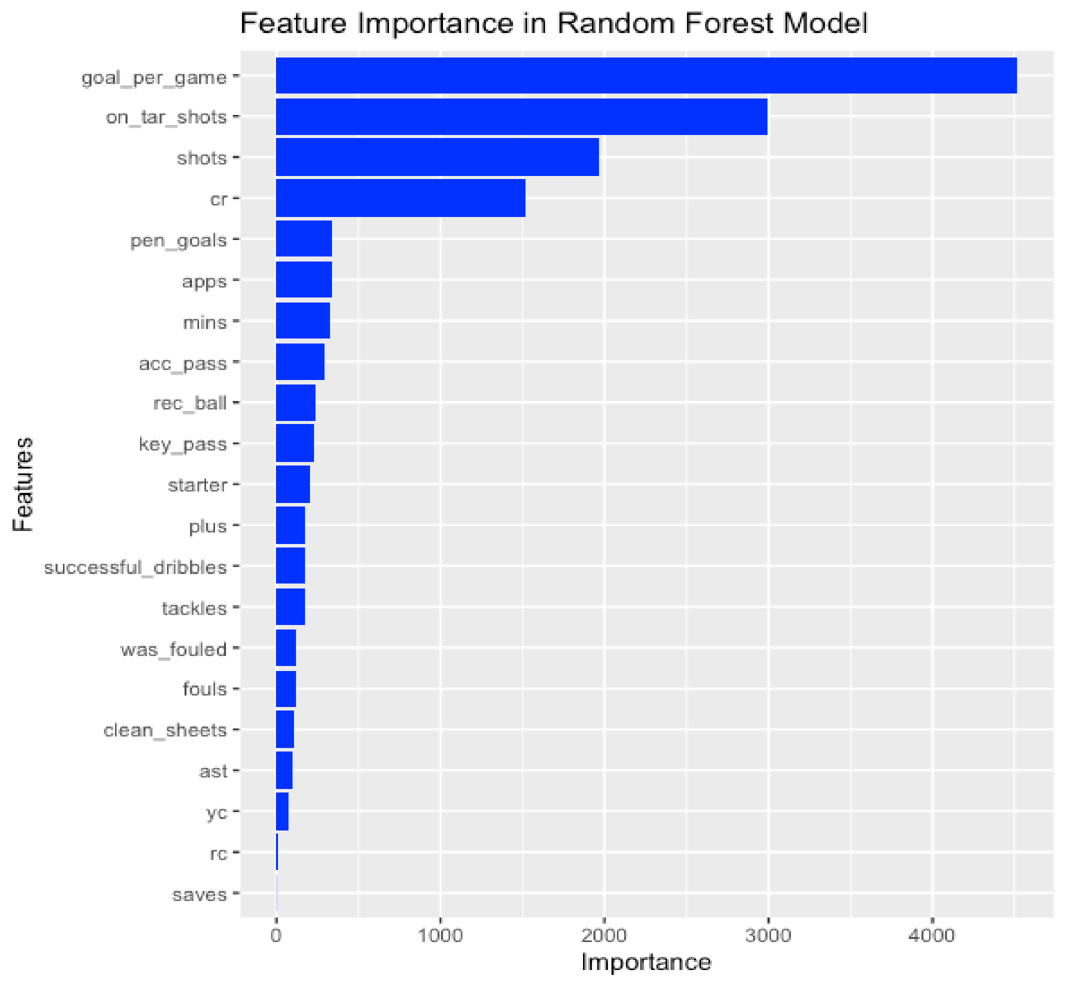
- Accuracy: 48.61%
- Root Mean Squared Error (RMSE): 1.438381
- Mean Absolute Error (MAE): 0.8627999

Confusion Matrix Insights:

- High sensitivity in correctly identifying non-scorers (Class 0).
- Specificity indicates good true negative rate across classes.
- Balanced accuracy varied, reflecting class imbalances.

Model Highlights:

- Successfully predicted Mohamed Salah as the top goal scorer.
- Demonstrates robust predictive capability and is better than Boosting



Boosting Model

- **Model Type**: Gradient Boosting Machine (GBM)

Key Model Parameters:

- Number of Trees: 900
- Interaction Depth: 5
- Learning Rate (Shrinkage): 0.01
- Cross-Validation Folds: 10
- Minimum Observations per Node: 8

Performance Metrics:

- Model Accuracy: 55.75%
- Kappa Score: 0.4007 (Indicates Moderate Predictive Power)
- Mean Absolute Error (MAE): 0.795233
- Root Mean Squared Error: 1.292113

Feature Importance: Top Influential Features: 'goal_per_game', 'shots', 'appearances'

Model Highlights:

- Successfully predicted Mohamed Salah as the top goal scorer.
- Demonstrates the model's effectiveness in identifying key players' performance.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  22  23
        0  112  14   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        1   20  23   8   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        2    0  15  11   5   1   2   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
        3    0   2   7   5   4   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        4    0   0   5   2   4   1   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
        5    0   1   0   1   1   0   2   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0
        6    0   0   0   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        7    0   0   0   0   2   4   1   0   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0
        8    0   0   0   0   0   0   1   1   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0
        9    0   0   0   0   0   0   1   1   2   0   0   1   0   1   0   0   0   0   0   0   0   0   0
        10   0   0   0   0   0   0   0   0   0   1   0   1   1   0   0   0   0   0   0   0   0   0   0
        11   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0
        12   0   0   0   0   0   0   0   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0   0
        13   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        14   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        15   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
        16   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
        17   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
        18   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0
        19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        20   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        22   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        23   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.5575
                 95% CI : (0.4979, 0.6158)
    No Information Rate : 0.4599
    P-Value [Acc > NIR] : 0.0005751

                  Kappa : 0.4007
```

Deep Learning

Model Architecture:

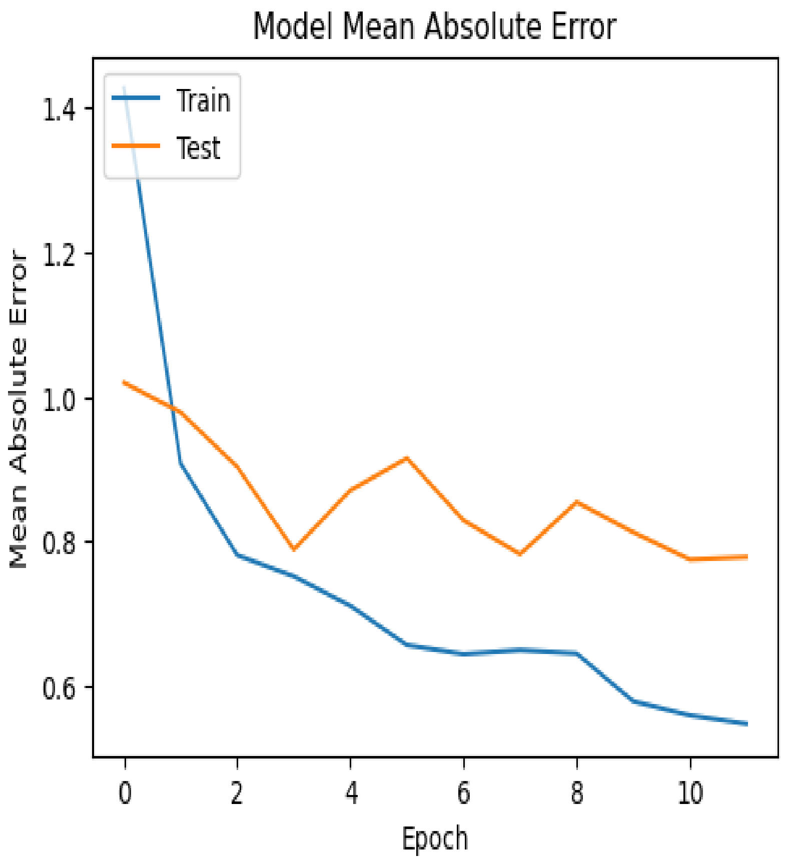
- Type: Sequential Neural Network
- Layers:
  - Dense Layer: 256 neurons, activation='relu'
  - Dense Layer: 128 neurons, activation='relu'
  - Dense Layer: 64 neurons, activation='relu'
  - Dense Layer: 32 neurons, activation='relu'
  - Output Layer: 1 neuron
- Optimizer: 'adam'
- Loss Function: 'mean_squared_error'
- Performance Metrics: 'mean_absolute_error'

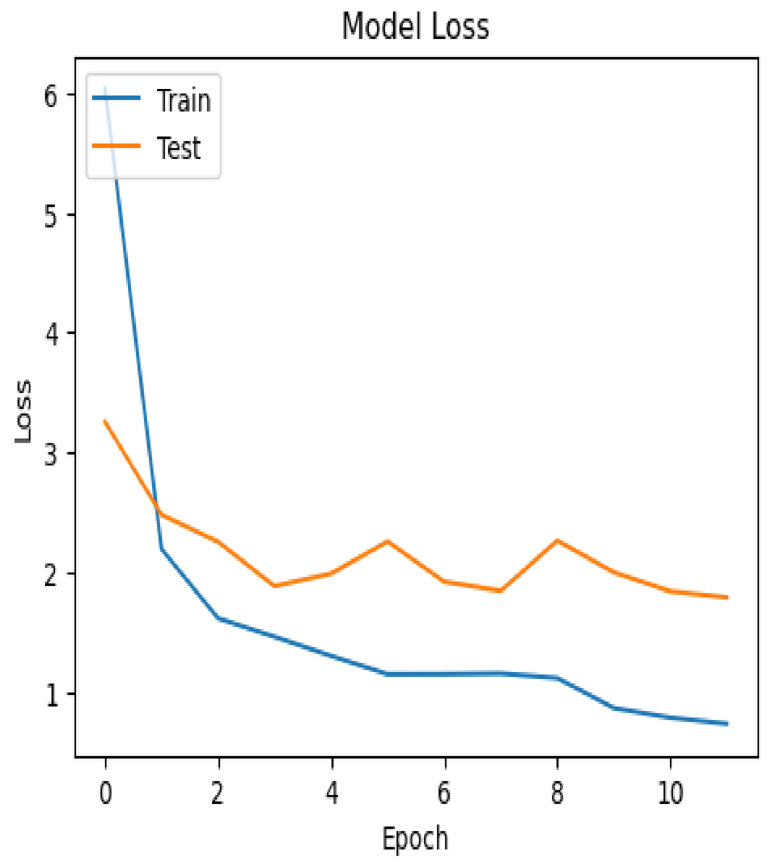
Training and Validation:

- Epochs: 12

Model Performance:

- Test Loss: 2.0596
- Test Mean Absolute Error: 0.7455





Key Insights:

- The model successfully predicted Mohamed Salah as the top goal scorer.
- Prediction closely aligns with actual goal scores, indicating high model accuracy.

Models Comparison

Model	RMSE	MAE	Accuracy
Linear Model	1.471316	0.9827922	43.32%
Boosting model	1.292113	0.792431	55.75%
Random Forest	1.4383	0.8627	48.61%
Deep Learning	1.4351	0.7455	49.13%

Applications

- Serie A league

Linear Performance Metrics:

- Accuracy: 49.89%
- Root Mean Squared Error (RMSE): 1.3868

Boosting Performance Metrics:

- Accuracy: 54.3%
- Root Mean Squared Error (RMSE): 1.1990

Random Forest Performance Metrics:

- Accuracy: 48.54%
- Root Mean Squared Error (RMSE): 1.3082

Deep Learning Performance Metrics:

- Accuracy: 43.56%
- Root Mean Squared Error (RMSE): 1.2753

Best Model: Boosting Model

Boosting Model appears to be the best model for this prediction task. It has been highlighted for its superior accuracy and lower error metrics compared to the other models.

Top Scorers Across All Models

Linear Model	Boosting Model	Forest Model	Deep Learning	Actual
V. Osimhen	V. Osimhen	V. Osimhen	V. Osimhen	V. Osimhen
M. Lautaro	M. Lautaro	M. Lautaro	M. Lautaro	M. Lautaro
R. Leao	K. Kvaratskhelia	K. Kvaratskhelia	D. Vlahovic	R. Leao
D. Vlahovic	R. Leao	R. Leao	K. Kvaratskhelia	A. Lookman
D. Berardi	C. Immobile	D. Vlahovic	A. Lookman	Bala Nzola

Limitations

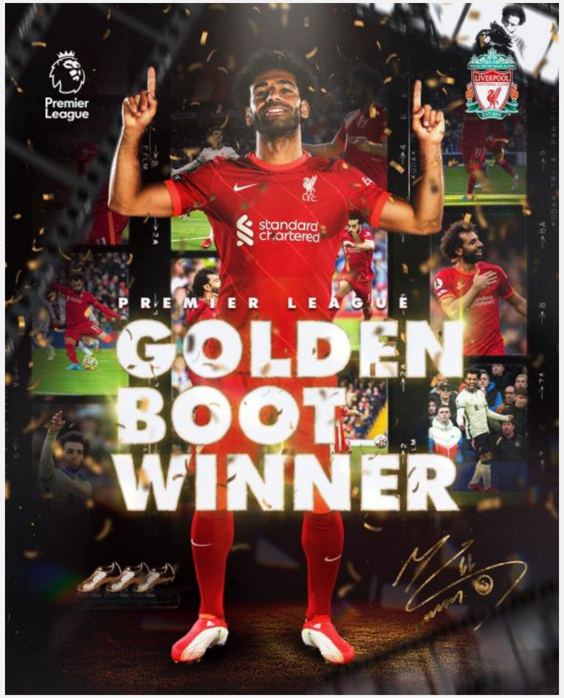
- **Data Constraints:** Limited to specific seasons and may not account for all variables affecting a player's performance (e.g., injuries, transfers).
- **Model Bias:** Potential biases in the models due to the data used. Need for more diverse datasets to improve model robustness.
- **Dynamic Nature of Football:** The unpredictable nature of sports, including player form fluctuations and tactical changes, can impact model accuracy.

Future Research

- **Incorporating More Data:** Expand dataset to include more seasons, leagues, and player-specific data like fitness levels or psychological factors.
- **Player Development Focus:** Shift from predicting top goal scorers to identifying potential star players based on early-career performance data.
- **Interdisciplinary Approaches:** Collaborate with sports scientists and psychologists to integrate physical and mental health metrics into predictive models.

Conclusion

- The primary objective of our project was to utilize predictive analytics to forecast the season's top goal scorer from mid-season performance data.
- Our comprehensive analysis determined that the Boosting Model outperformed other models with the highest accuracy, validating our hypothesis.
- The success of the Boosting Model highlights the potential of machine learning in sports analytics and its impact on strategic decision-making.
- We advocate for a diverse modeling approach, as it's critical for robust and resilient predictions in dynamic environments like football.
- This project underscores the importance of data-driven insights in football strategies and enhancing the understanding of player performance.



1

Best RMSE

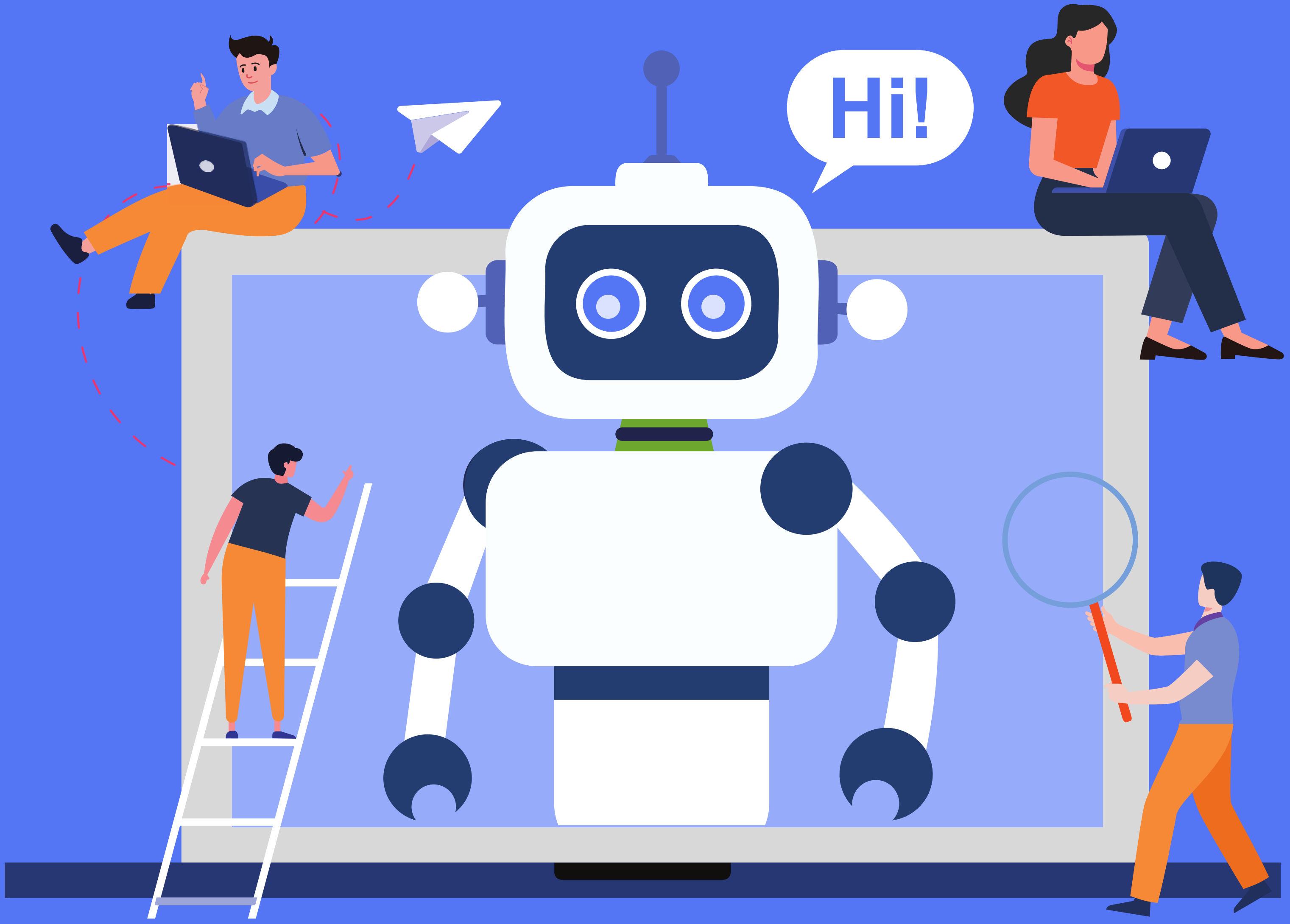
1.29

2

Best Accuracy

55.75%

Q&A



Thanks