ScoreSight

AI-Driven Prediction for EPL Points Tally & Top Scorers

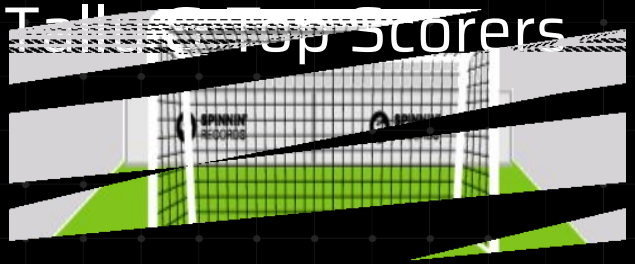
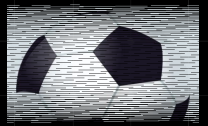


Table of Contents



Project Overview

Objective

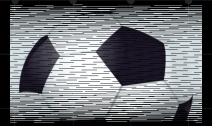
To build ML models to predict English Premier League (EPL) Season Outcomes.

Core Tasks

- ➔ Predict the Match Winner(Home Team/Not Home Team)
- ➔ Predict League Winner
- ➔ Predict Total Points
- ➔ Predict Top Scorer(No. of Goals and Assists)

Use Case

Helping users understand team performance and player's performance based on historical data of the past 25 years of EPL seasons.



Data set Overview & Key Insights

Sources: EPL Score Prediction and Player Performance datasets.

Key Features From Data set:

1. Team 1
2. Team 2
3. Home/Away Stadium
4. Player History
5. Previous Club Performances

Why this Data set?

It contains the "Historic Data" required to train regression models for future outcomes.

Methodology

Data Preprocessing

Gather raw data from various sources (databases, APIs) and sanitize it by handling missing values, removing duplicates, and addressing outliers. Perform Exploratory Data Analysis (EDA) to visualize distributions and ensure the dataset is clean and structured for further processing.

Feature Extraction

Transform raw variables into machine-readable numerical formats using techniques like One-Hot or Label encoding for categorical data. Apply feature scaling (Standardization or Normalization) and dimensionality reduction (like PCA) to create meaningful inputs that optimize algorithm performance.

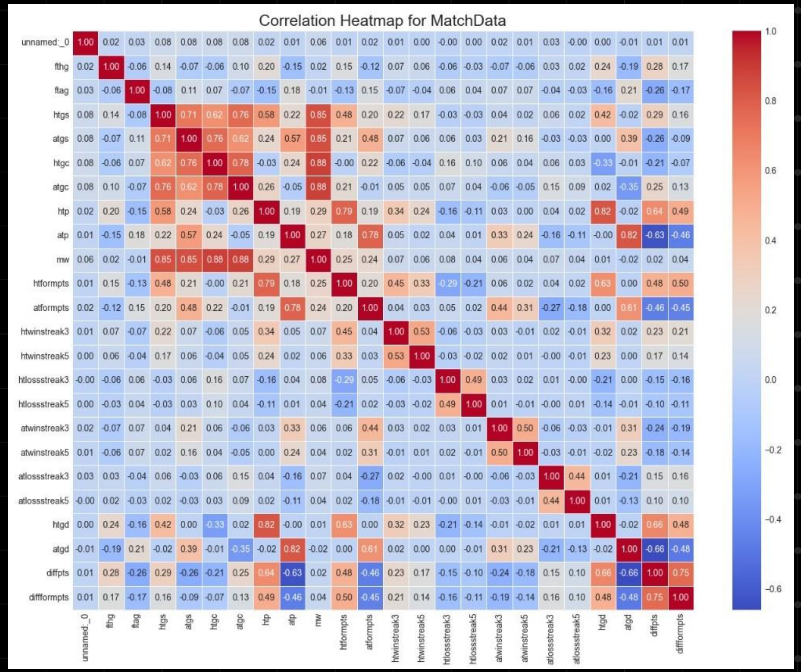
Model Architecture Extraction

Select the appropriate algorithm structure based on the problem type, such as Linear Regression for continuous data or SVM/Random Forest for classification. Define the underlying model framework that best fits the complexity of the engineered features and intended predictions.

Training And Evaluation

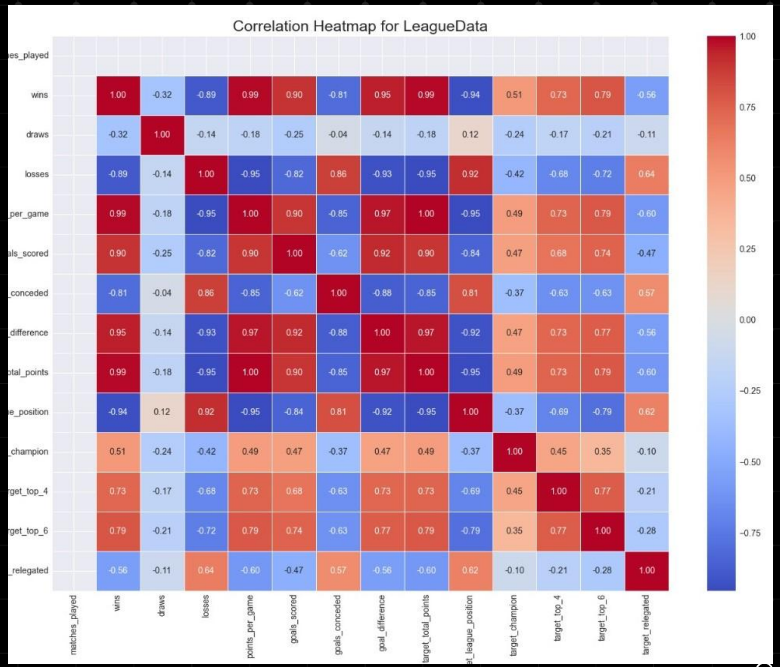
Split data into training, validation, and testing sets to teach the model patterns while tuning hyperparameters to prevent overfitting. Assess the final model's generalization capabilities on unseen data using rigorous metrics like Accuracy, F1-score, or RMSE.

Exploratory Data Analysis



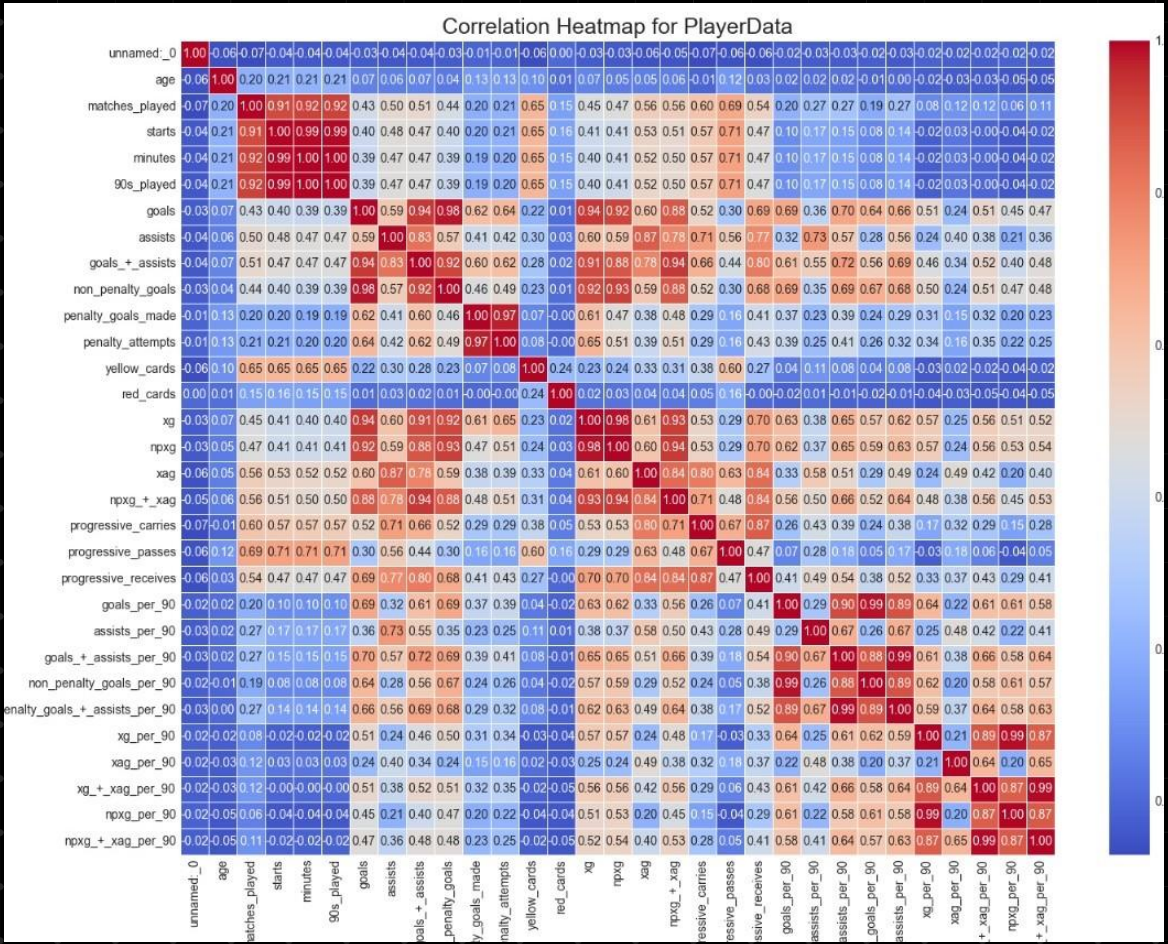
Correlation Heatmap for Match data

Variables related to point differentials (diffpts) and form points show the strongest correlations, making them potentially the most significant predictors in this dataset.



Correlation Heatmap for League data

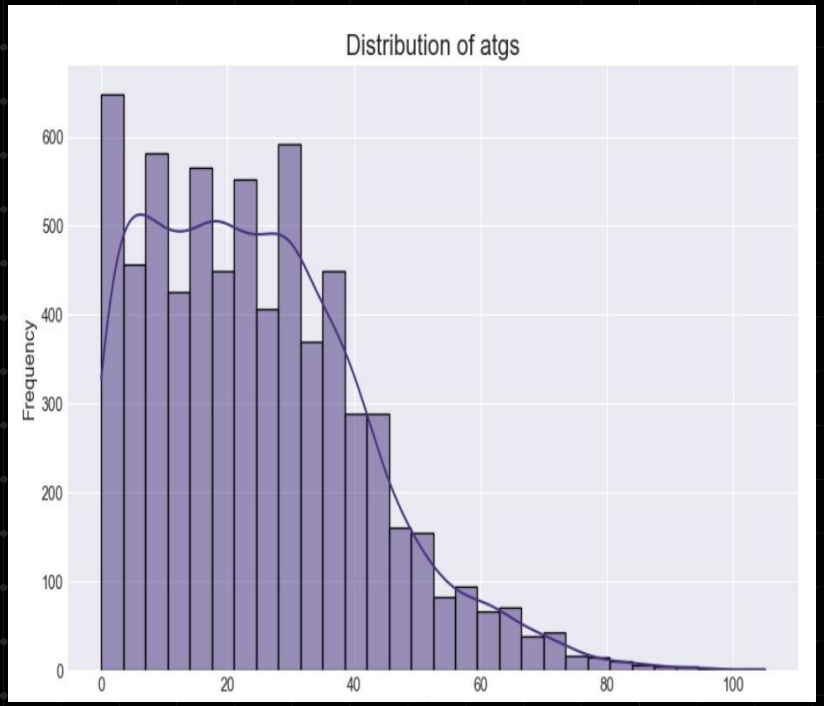
Goal Difference (0.97) proves to be a stronger predictor of Total Points than Goals Scored (0.90) alone, highlighting the critical role of defense.



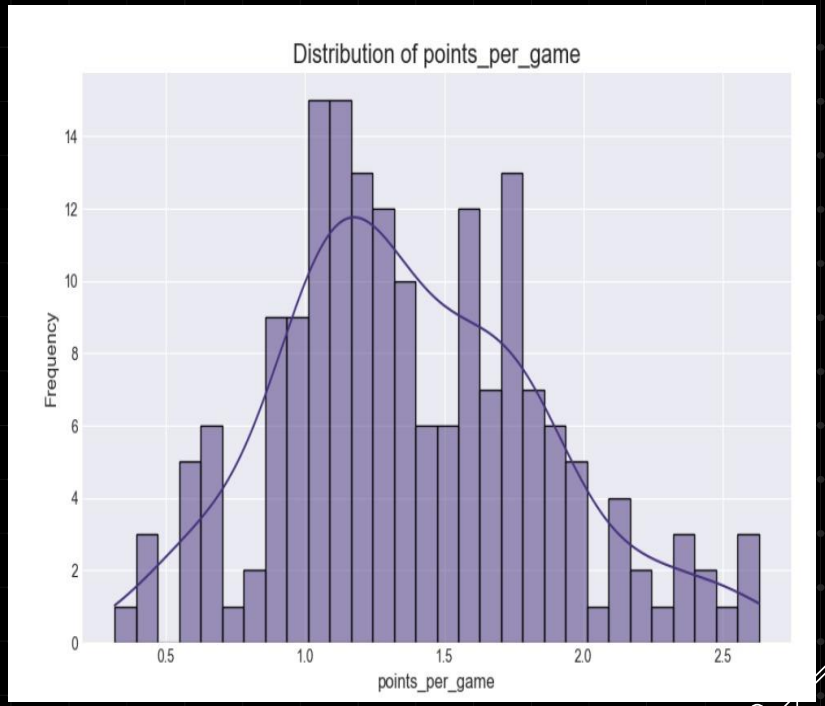
Correlation Heatmap for Player data

The strong positive correlation between Expected Goals (xg) and actual goals confirms that xg is a highly reliable metric for predicting player scoring performance in this dataset.

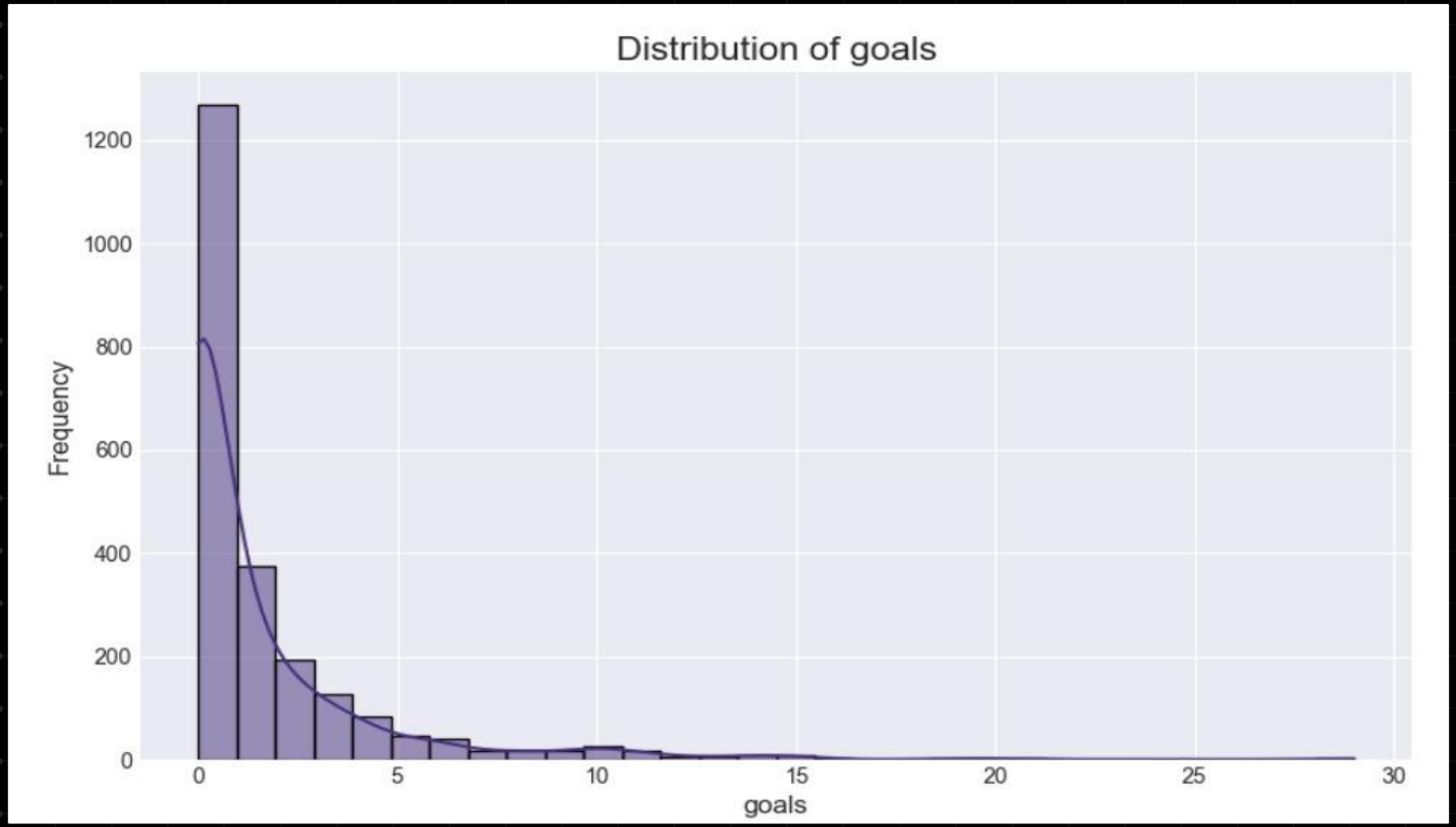
Visualisation



The distribution of atgs is strongly right-skewed, indicating that while most values cluster in the 0–40 range, values exceeding 60 represent rare, high-performing outliers.



The broad spread from roughly 0.3 to 2.6 PPG visualizes the significant performance gap between the league's struggling teams and its dominant champions.



The extreme right-skewed distribution highlights that goal-scoring follows a 'power law,' where a tiny elite fraction of players contribute the vast majority of the league's goals.

Data Preprocessing

1. **Data Collection**

   Gather raw datasets from diverse sources such as internal databases, external APIs, web scraping, or flat files to build a comprehensive foundation for the model.

2. **Data Cleaning**

   Ensure data integrity by identifying and addressing missing values (NaNs) through removal or imputation, and eliminating duplicate records to guarantee unique, valid inputs.

3. **Exploratory Data Analysis (EDA)**

   Visualize and summarize the dataset's key characteristics to understand underlying distributions, detect relationships between variables, and spot potential outliers before modeling begins.

--- Match Winner Data (Outlier Check) ---

	HTGS	ATGS	HTGC	ATGC	HTP	ATP	HTGD	ATGD	DiffPts
count	6639.000000	6639.000000	6639.000000	6639.000000	6639.000000	6639.000000	6639.000000	6639.000000	6639.000000
mean	25.152884	25.251996	25.234373	25.080584	1.243961	1.261424	-0.008929	0.014753	-0.017464
std	16.899263	16.853352	16.083548	16.026097	0.495933	0.487906	0.702314	0.702252	0.676500
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-3.000000	-3.333333	-2.363636
25%	12.000000	12.000000	12.000000	12.000000	0.913043	0.933333	-0.500000	-0.480000	-0.470588
50%	23.000000	23.000000	24.000000	24.000000	1.190476	1.200000	-0.100000	-0.076923	0.000000
75%	36.000000	36.000000	37.000000	36.000000	1.571429	1.573593	0.416667	0.444444	0.440588
max	102.000000	105.000000	85.000000	82.000000	2.736842	2.761905	4.000000	3.500000	2.285714

The dataset contains 6,639 entries with goal scoring metrics ranging significantly (0 to 105), indicating a wide spread of performance data used for outlier detection.

--- Goals and Assists Data (Outlier Check) ---

	Age	Matches Played	Starts	Minutes	90s Played	Goals	Assists	Goals Per 90	Assists Per 90
count	2057.000000	2057.000000	2057.000000	2057.000000	2057.000000	2057.000000	2057.000000	2057.000000	2057.000000
mean	25.183763	19.060282	14.081186	1257.741371	13.974088	1.482742	1.038405	0.095571	0.062178
std	4.482811	11.592936	11.506500	986.497676	10.961787	2.959654	1.779144	0.184147	0.104574
min	15.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	22.000000	9.000000	3.000000	349.000000	3.900000	0.000000	0.000000	0.000000	0.000000
50%	25.000000	19.000000	12.000000	1085.000000	12.100000	0.000000	0.000000	0.000000	0.000000
75%	28.000000	30.000000	24.000000	2109.000000	23.400000	2.000000	1.000000	0.120000	0.090000
max	41.000000	38.000000	38.000000	3420.000000	38.000000	29.000000	18.000000	2.430000	1.010000

The data reveals massive performance gaps, where the top scorer (29 goals) and top assister (18) sit far above the 75th percentile (2 goals, 1 assist).

```
--- League Data (Outlier Check) ---
```

	wins	draws	losses	goals_scored	goals_conceded	goal_difference	target_total_points	target_champion
count	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000
mean	14.544444	8.911111	14.544444	53.833333	53.833333	0.000000	52.544444	0.050000
std	6.297399	2.874058	6.018892	17.635509	13.931396	28.486006	18.168769	0.218553
min	2.000000	2.000000	1.000000	20.000000	22.000000	-69.000000	12.000000	0.000000
25%	10.000000	7.000000	10.750000	40.000000	44.750000	-19.250000	40.000000	0.000000
50%	13.000000	9.000000	15.000000	51.000000	54.000000	-2.500000	50.000000	0.000000
75%	19.000000	11.000000	18.000000	65.250000	63.000000	15.250000	66.000000	0.000000
max	32.000000	15.000000	30.000000	106.000000	104.000000	79.000000	100.000000	1.000000

The data highlights extreme polarization in team performance, with goal differences swinging drastically from -69 (worst defense) to +79 (best offense).

Feature Extraction

Handle Outliers

Identify extreme values that could skew the model's performance and address them by either capping them (winsorization) or removing the data points entirely, depending on the specific context.

Encode Categorical Variables

Convert non-numeric text data into machine-readable numbers using One-Hot Encoding for nominal categories (creating binary columns) or Label Encoding for ordinal data to ensure the model can process the inputs.

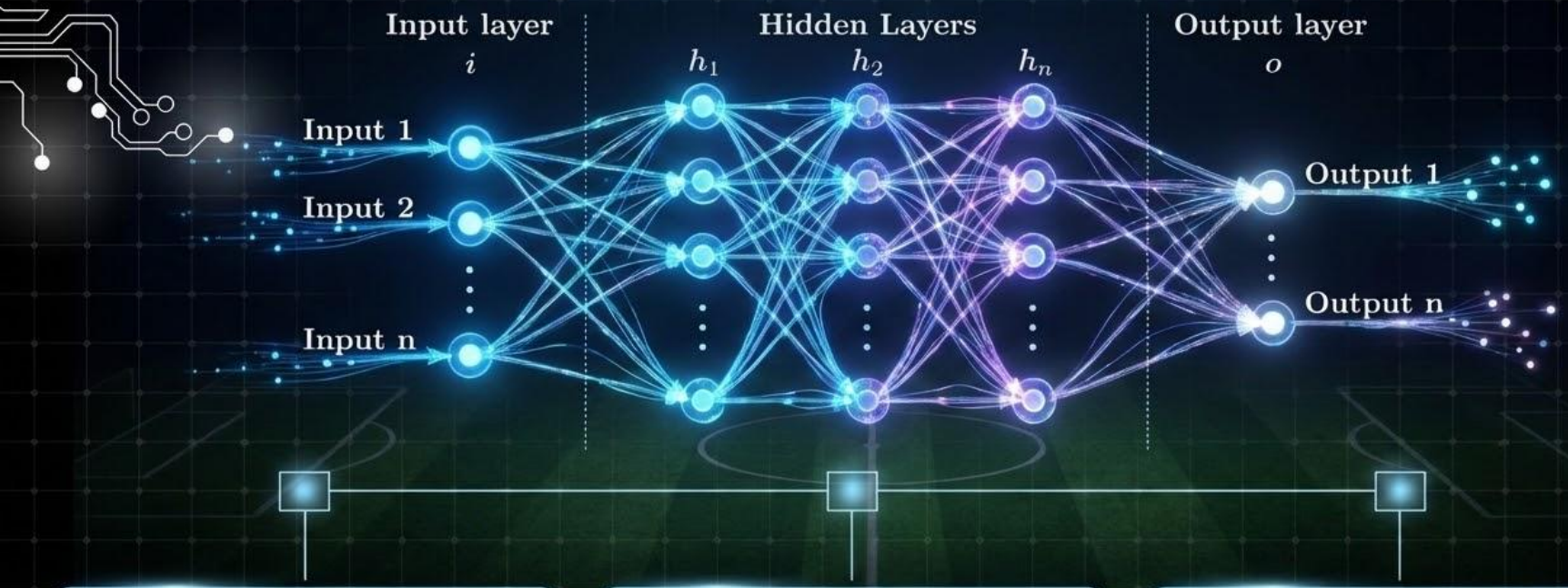
Feature Scaling

Standardize the range of independent variables using techniques like Z-score standardization or Min-Max normalization, which is critical for algorithms sensitive to the magnitude of values like SVMs and k-NN.

Dimensionality Reduction

Apply techniques like Principal Component Analysis (PCA) to reduce the number of input features, helping to simplify complex datasets with many correlated variables and improve computational efficiency.

Model Architecture



Input Layer

Acts as the entry point where raw data enters the network, with each node representing a specific feature of the dataset. It performs no computation but simply passes these values forward to the subsequent hidden layers.

Hidden Layer

The core processing units where neurons apply weights, biases, and activation functions (like ReLU) to introduce non-linearity. This allows the network to "learn" by detecting complex patterns and relationships within the data.

Output Layer

Produces the final prediction or classification, with a structure tailored to the specific task (e.g., single node for binary, multiple for multi-class). It applies a final activation function to convert processed signals into probabilities, labels, or continuous values.

Training And Evaluation

Total Points

MODEL PERFORMANCE RANKINGS

Model Rankings by R^2 Score:

Ridge (Points)	R^2: 0.9375	MAE: 3.7045	RMSE: 4.6429
RandomForest (Points)	R^2: 0.9313	MAE: 3.9637	RMSE: 4.8674
XGBoost (Points)	R^2: 0.9222	MAE: 4.2288	RMSE: 5.1801

PERFORMANCE METRICS ANALYSIS

✓ **BEST MODEL: Ridge (Points)**

- R^2 Score: 0.9375
- MAE: 3.7045 points
- RMSE: 4.6429 points

Performance Interpretation:

- ✓ Excellent R^2 (0.9375) - Model explains 93.7% of variance
- ✓ Average prediction error (MAE): ±3.70 points
- ✓ Root mean squared error (RMSE): 4.64 points

Goals & Assists

```
--------------------------------------------------------------------------------
FINAL PREDICTION PERFORMANCE SUMMARY
--------------------------------------------------------------------------------
Target: GOALS
  Selected Model: XGBoost
  Test R2:   0.9621
  Test MAE:  0.2202
  Test RMSE: 0.5696
  Best Parameters: {'model': 'XGBoost', 'xgb_n_estimators': 155, 'xgb_learning_rate': 0.8807730502835389, 'xgb_max_depth':
5, 'xgb_subsample': 0.853041121958712 8, 'xgb_colsample_bytree': 0.667923042626643 3}
-------------------------------------
Target: ASSISTS
  Selected Model: HistGradientBoosting
  Test R2:   0.9675
  Test MAE:  0.1436
  Test RMSE: 0.3232
  Best Parameters: {'model': 'HistGradientBoosting', 'hgb_learning_rate': 0.202454310510827, 'hgb_max_iter': 277, 'hgb_max_
depth': 5}
-------------------------------------

Summary Table:
 Target       Selected Model  Test R2  Test MAE  Test RMSE
  GOALS              XGBoost   0.9621    0.2202     0.5696
ASSISTS HistGradientBoosting   0.9675    0.1436     0.3232
--------------------------------------------------------------------------------
```

```
================================================================================
MODEL: LogisticRegression
================================================================================
Overall Accuracy: 65.0585%

Classification Report:
--------------------------------------------------------------
              precision    recall  f1-score   support

           H     0.6221    0.6299    0.6260       635
          NH     0.6759    0.6685    0.6722       733

    accuracy                         0.6506      1368
   macro avg     0.6490    0.6492    0.6491      1368
weighted avg     0.6509    0.6506    0.6507      1368

--------------------------------------------------------------

################################################################################

================================================================================
MODEL: RandomForest
================================================================================
Overall Accuracy: 63.1579%
...
     NeuralNetwork   64.91%   0.6467      0.6488
  StackingEnsemble   64.47%   0.6392      0.6424
      RandomForest   63.16%   0.6262      0.6294
  GradientBoosting   63.08%   0.6247      0.6281
```

Match Winner

```
================================================================================
MODEL EVALUATION
================================================================================
Evaluating models...

🏆 Best Model: RandomForest (F1 Score: 0.9577)
----------------------------------------
--- RandomForest Evaluation ---
Accuracy: 0.9500
Precision (Weighted): 0.9750
Recall (Weighted): 0.9500
F1 Score (Weighted): 0.9577
ROC AUC: 0.9766

Classification Report:

              precision    recall  f1-score   support

Not Champion       1.00      0.95      0.97        57
    Champion       0.50      1.00      0.67         3

    accuracy                           0.95        60
   macro avg       0.75      0.97      0.82        60
weighted avg       0.97      0.95      0.96        60
```

League Winner

Results

- ## League Winner :

  Best Model: LightGBM achieved the highest performance.

  Accuracy: Reached a 97.2% Test Accuracy, correctly identifying champions in almost all test cases.

- ## Match Winner :

  Best Model: XGBoost outperformed Random Forest and Gradient Boosting.

- ## Top Scorer :

  Winning Model: Random Forest Regressor.

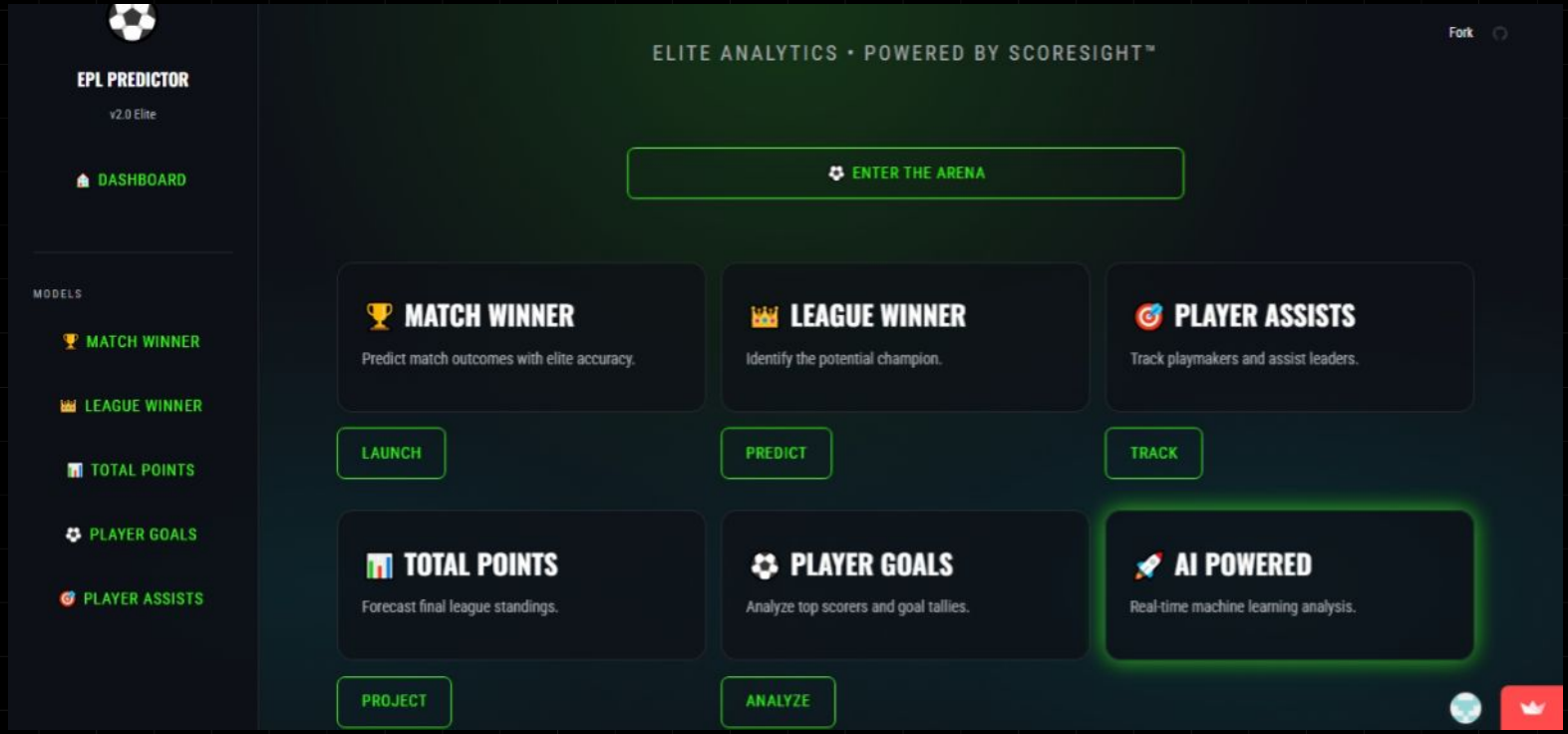
  Precision: Achieved a low Mean Absolute Error (MAE) of 0.24 goals, meaning predictions were incredibly precise (typically off by less than a quarter of a goal).

- ## Total Points :

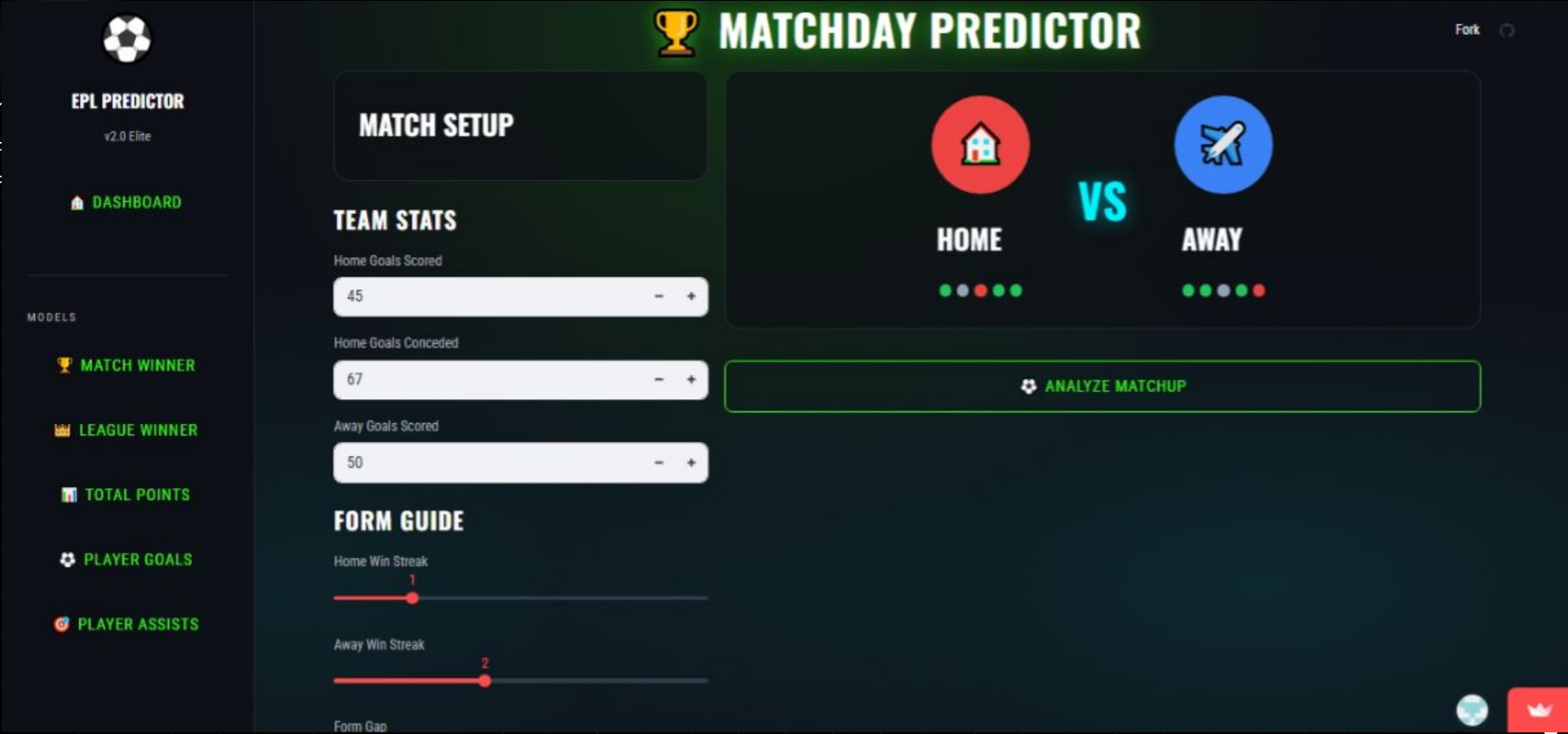
  Winning Model: Linear Regression.

  Fit Quality: Scored a high R-Squared (R^2) of 0.86, confirming that a team's Goal Difference is a very strong linear predictor of their final points total.

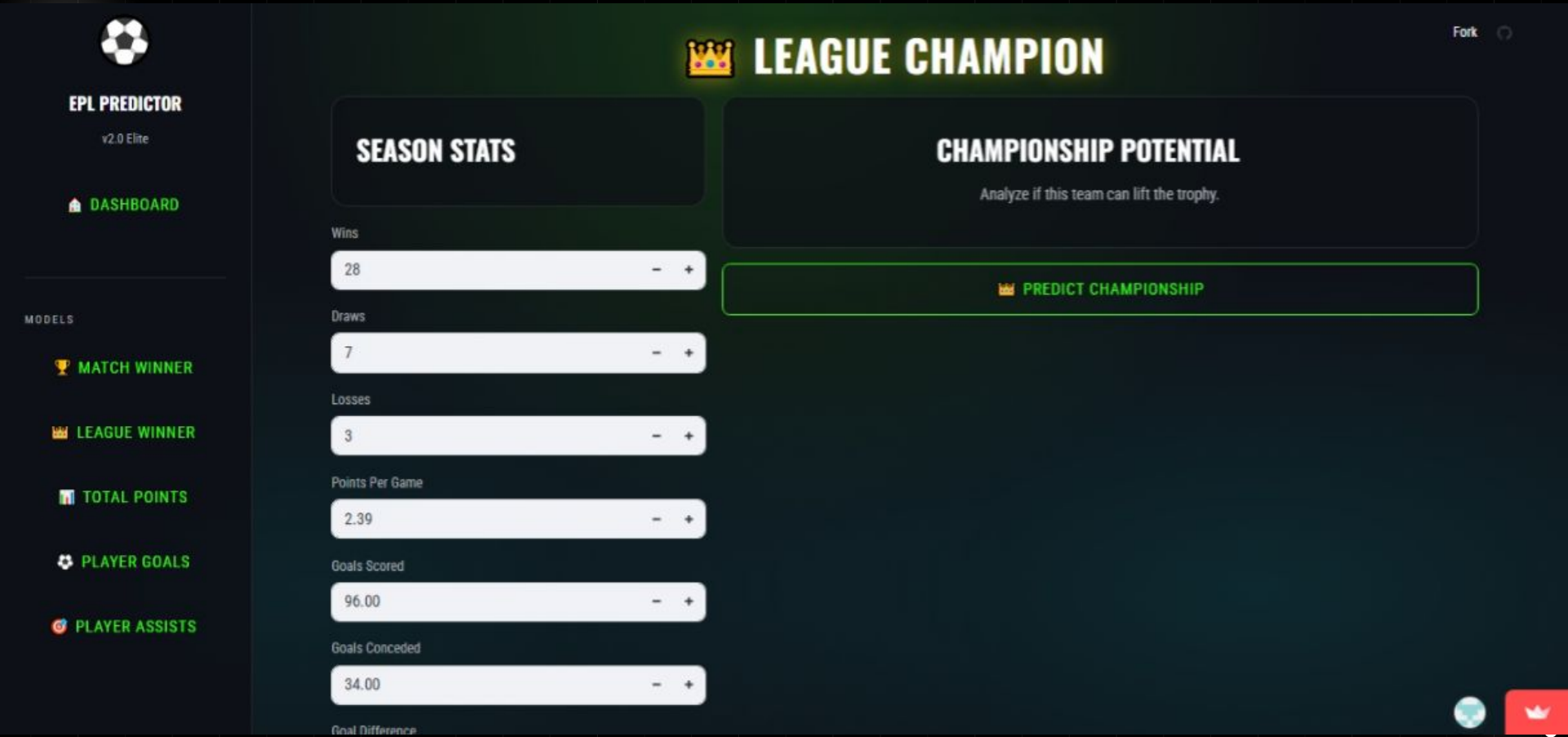
User Interface



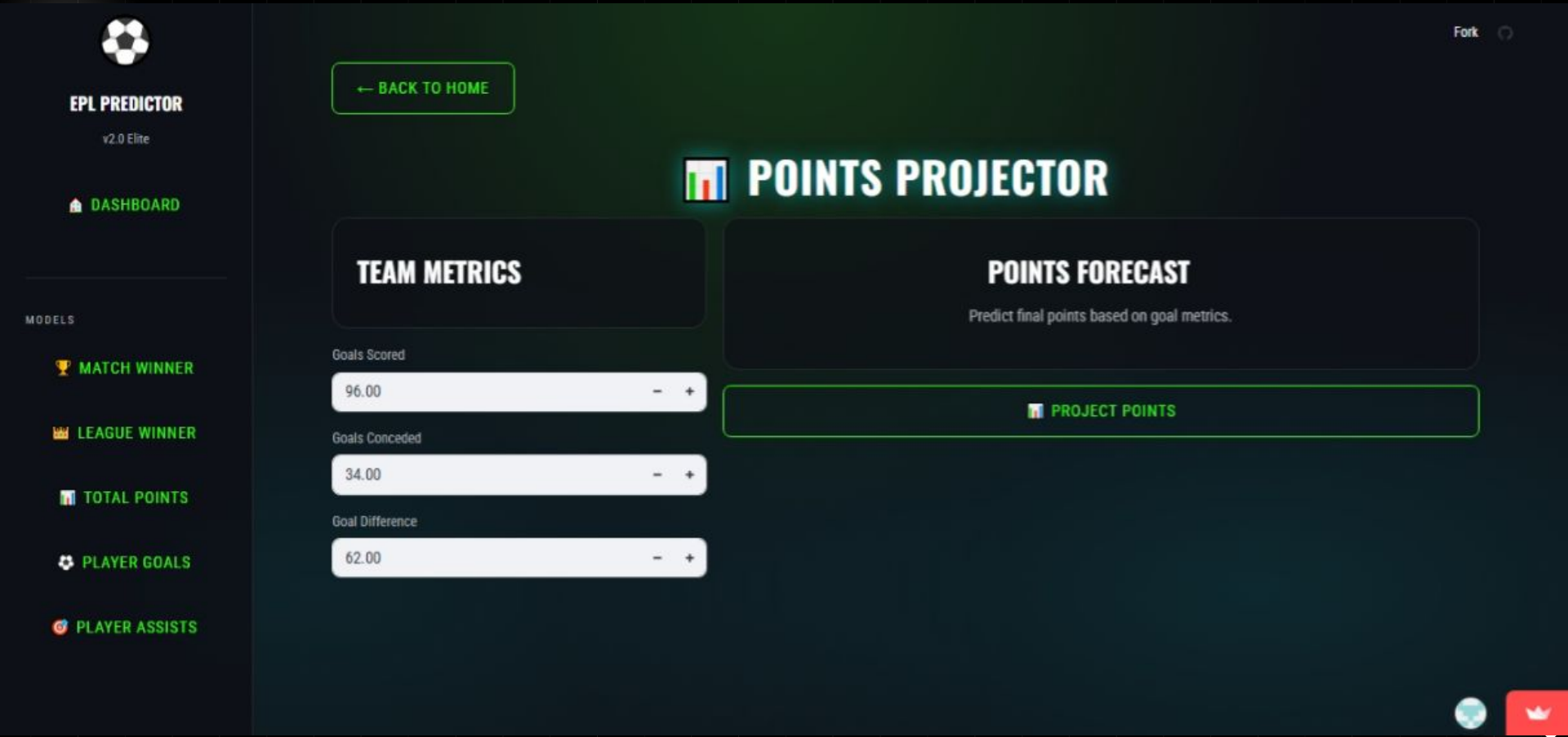
The central dashboard provides seamless access to all predictive modules, encompassing “Match Winner”, ”League Winner”, “Total Points”, “Player Goals”, ”Player Assists”



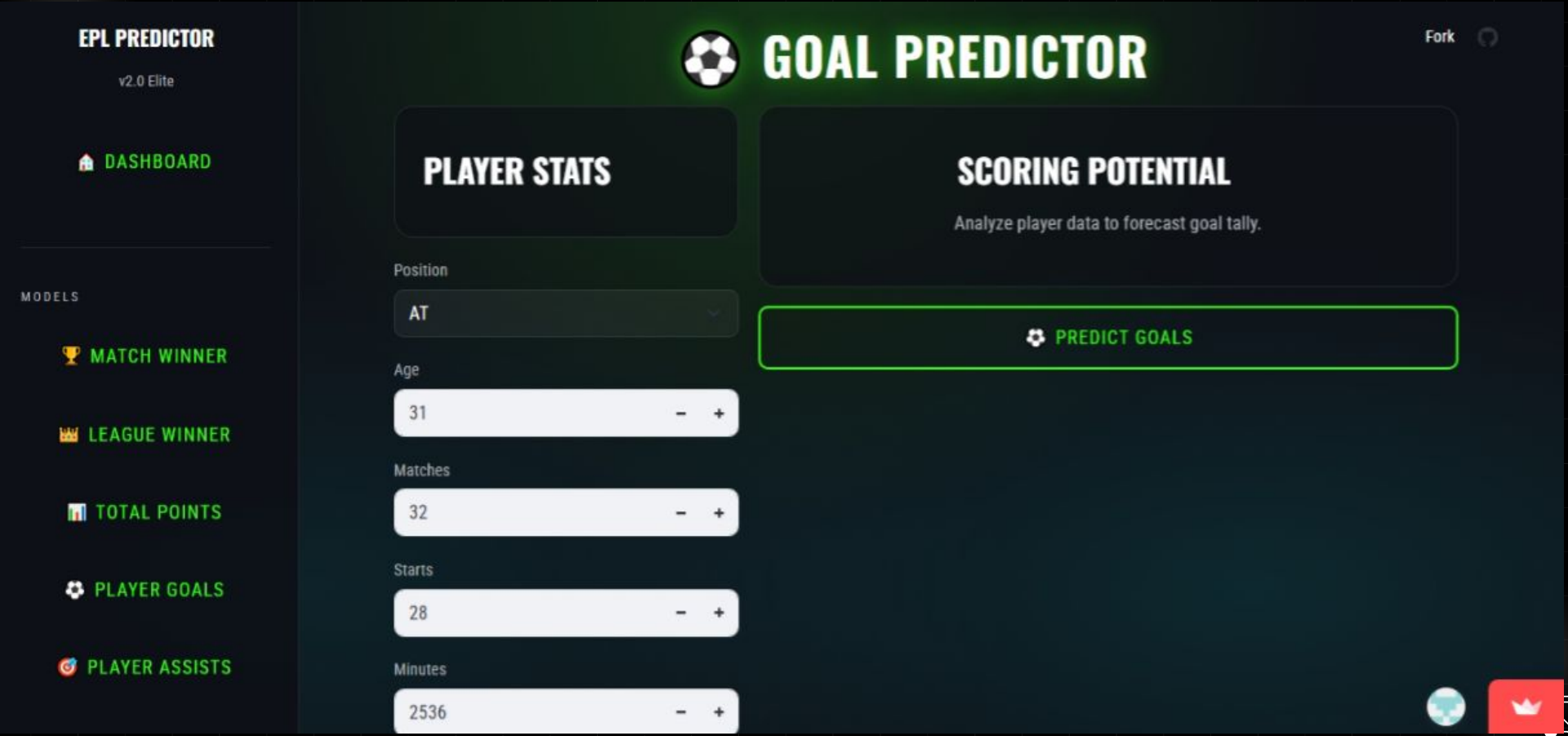
The “Match Predictor” compares the relative performance of both teams to estimate the likelihood of a home win,away win, or draw with quantified confidence levels.



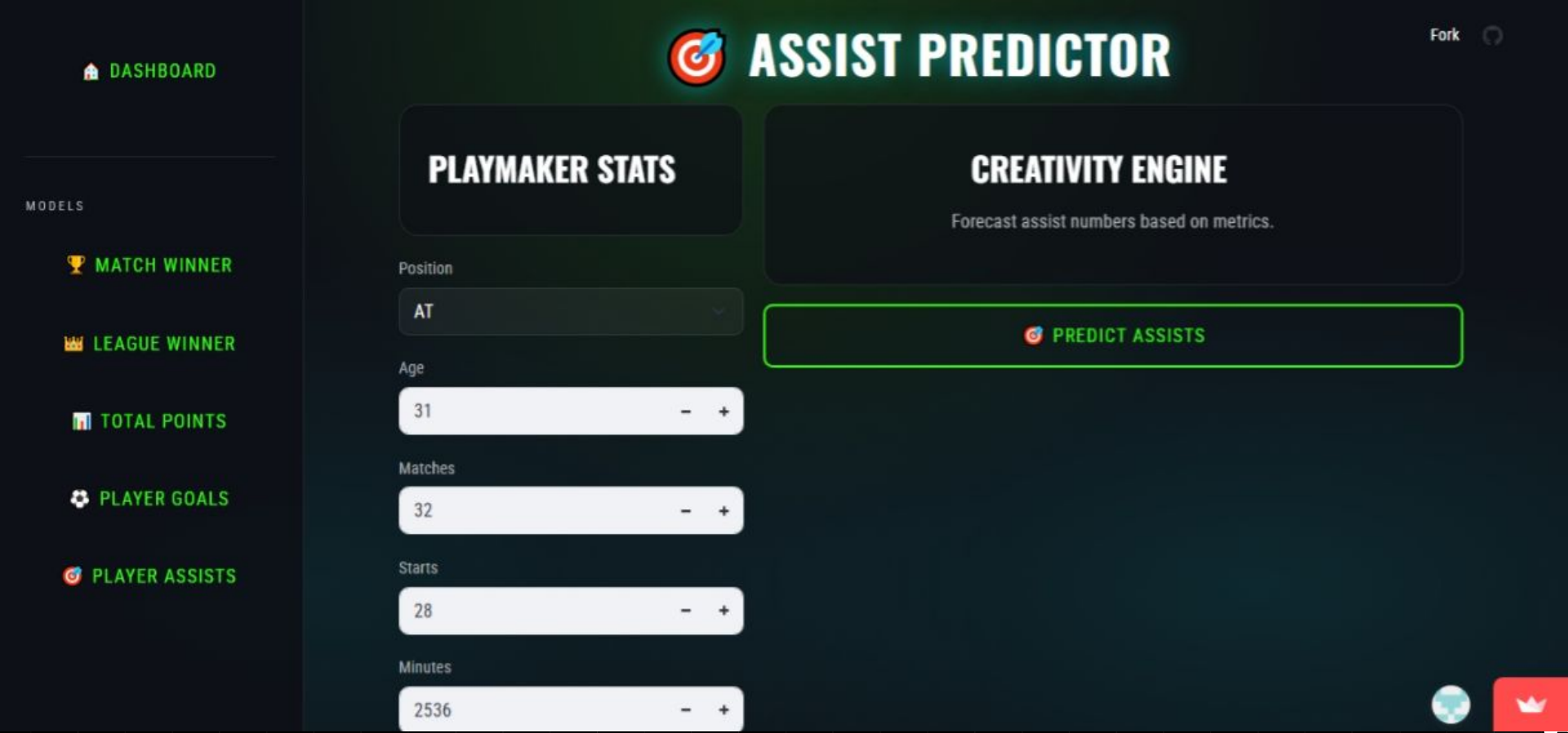
"League Winner" predicts whether a team will win "League" or not. It mainly depends on the factors like: Wins, Losses, Goals Scored and Goals Conceded by the team.



"Total Points " is used to predict the total points a team can score. For computation it requires Goals Scored, Goals Conceded and Goal Difference as an input.



The system evaluates team statistics such as player's position, Expected Goals (xG) Metrics, Discipline and Per 90 Stats to predict "Goals Prediction".



"Assist Predictor" predicts the Player Assist.

Challenges

- Data Formatting:

  Raw dataset not being in the exact format required for training.

- New Players:

  Predicting performance for players playing for the first time (no history in current club).

- Fine-tuning:

  Adjusting hyperparameters to minimize loss.

Future Scope

- Real-time Inference:

  Testing with live test data.

- Model Expansion:

  Trying different architectures or features to improve accuracy and personalizing user experience.

Conclusion

Successfully implemented advanced regression and classification models to generate accurate predictions of English Premier League match outcomes, incorporating key performance variables and historical trends.

Designed and developed a comprehensive player performance evaluation system that analyzes multi-season statistical data to forecast individual output and support data-driven decision-making.

THANK YOU!