AI ScoreSight: Predicting EPL League Winner & Top Assists

A Machine-Learning-Based Football Performance Prediction Project

This presentation summarizes the methodology and results of **AI ScoreSight**, a project leveraging machine learning to forecast outcomes in the English Premier League (EPL).

Presented By: Srikala Kanduri

Project Overview

AI ScoreSight

1

Dual Predictive Focus

Combining two distinct models—a classifier for team success and a regressor for individual player performance.

2

EPL Focus

Leveraging extensive historical and player data from the English Premier League (EPL) spanning over three decades.

3

Methodology

Utilizing robust ML pipelines, feature engineering, and hyperparameter tuning to achieve high fidelity predictions.

Problem Statements and Motivation

Our project addresses two critical challenges in football analytics, driven by the need for advanced tactical forecasting and enhanced fan engagement.

Predicting the EPL Champion

Classifying the winner of the English Premier League using historical match-level data from 1993 to 2024. This model supports strategic long-term planning.

Forecasting Top Assists

Regressing a player's expected total assists for the upcoming season based on their prior-season performance metrics.

Motivation: Driving Value in Football

- Informs scouting and player acquisition decisions.
- Enhances tactical analysis for coaches and analysts.
- Creates data-driven content for media and fan engagement platforms.

Core Objectives: Building Robust ML Models

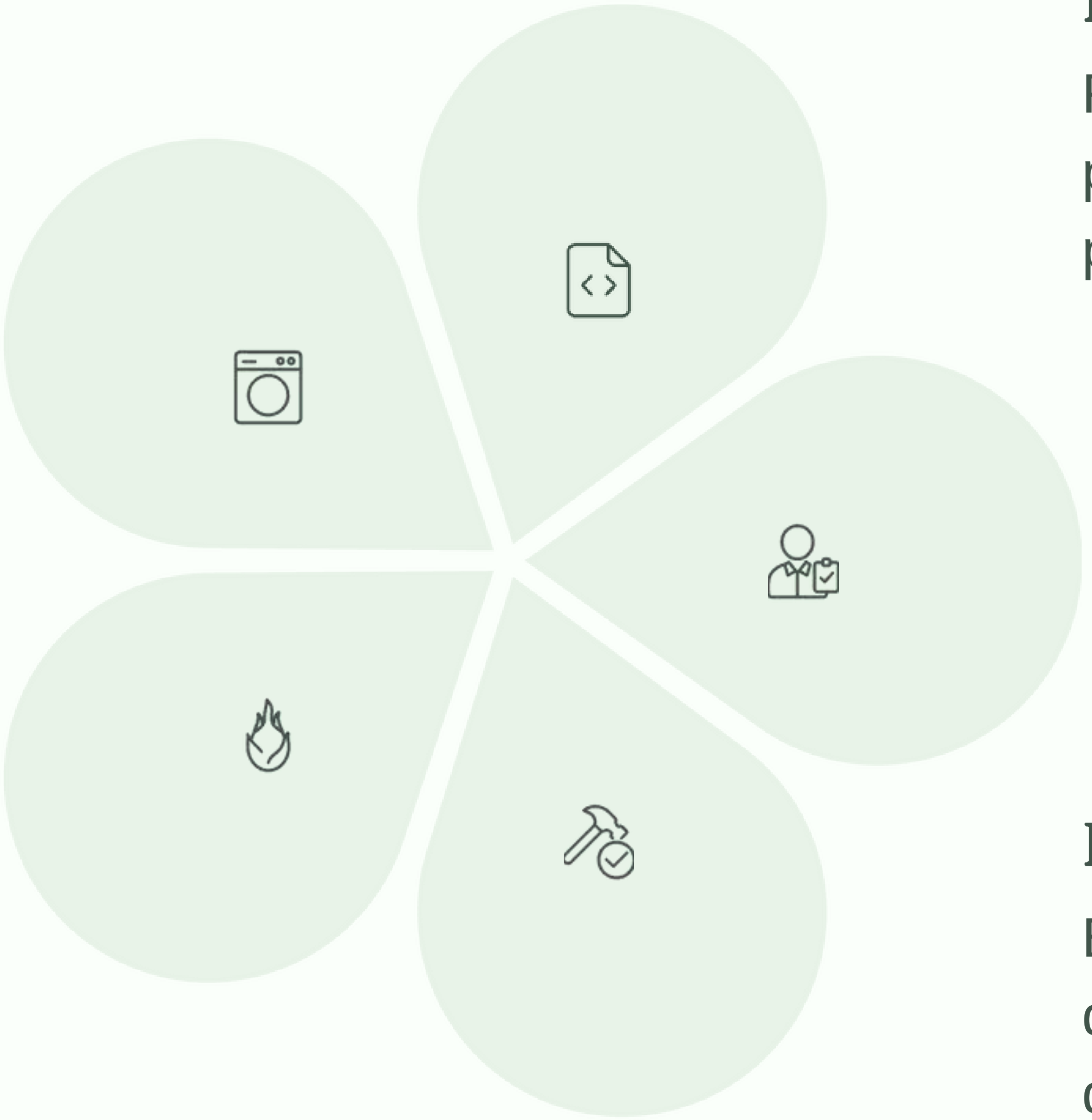
Our primary goal was to construct reliable, explainable machine learning systems capable of delivering measurable accuracy using verified EPL data.

Model Building

Develop specialized classification and regression models to forecast team and player outcomes.

Explainability

Prioritize model interpretability to understand the drivers behind the predictions.



Data Processing

Rigorous feature engineering and data preparation to transform raw statistics into predictive features.

Evaluation

Conduct thorough evaluation using industry-standard metrics to ensure model validity and reliability.

Deployment Readiness

Build models structured for potential deployment into real-world analytics dashboards or tools.

The methodology emphasizes clarity, replicability, and practical application within the sports domain.

Datasets and Feature Focus

Two distinct datasets were utilized, each tailored to its specific predictive task, ensuring relevant features for classification and regression.

Use Case	Dataset	Key Attributes	Target Variable
League Winner	pl-tables-1993-2024.csv	Played, Won, Drawn, Lost, GF, GA, GD, Points, Team	isChampion (Binary)
Top Assists	topassist.csv	Age, Position, Minutes Played, xA, Key Passes, Club xG	Total Assists (Numeric)

Visualizing Data Relationships

Visual exploratory analysis included correlation heatmaps to assess feature interdependence and scatter plots (e.g., Goal Difference vs. Final Position) to guide feature engineering.

Data Granularity

The League Winner model operates on team-season aggregates, while the Top Assists model relies on detailed player metrics like Expected Assists (xA).

Data Preprocessing & Advanced Feature Engineering

Feature engineering was crucial for transforming raw statistics into powerful predictors, particularly focusing on efficiency and historical performance metrics.

League Winner Model

- **Efficiency Ratios:** Calculated win_ratio, draw_ratio, loss_ratio, and points_per_match.
- **Strength Metrics:** Derived attack_strength, defense_strength, and the combined attack_defense_ratio.
- **Historical Context:** Incorporated previous season's points (prev_points) and final league position (prev_position).
- **Labels:** Established binary labels for isChampion and isRelegated.

Top Assists Model

- **Handling Missing Data:** Imputed missing numerical values with the mean; categorized unknowns as "Unknown."
- **Per 90 Metrics:** Normalized metrics by playing time (e.g., Assists_prev_per_90, Key_Passes_per_90) to ensure fair comparison.
- **Positional Data:** Created 'Minutes_Attack' to quantify forward contribution and removed leakage features.
- **Categorical Encoding:** Processed categorical features like position and club affiliation.

Model Design and Optimization Approach

The models selected—Random Forest Classifier and Regressor—are favored for their robustness and ability to capture non-linear feature interactions, optimized through extensive search methods.

League Winner Model: Random Forest Classifier

Integrated into an **imblearn.Pipeline** with **SMOTE** to address the high class imbalance (few champions relative to non-champions).

- Tuning Method: RandomizedSearchCV (20 iterations, 3-fold CV)
- Best Parameters: n_estimators = 50, min_samples_leaf = 4, bootstrap = True

Top Assists Model: Random Forest Regressor

Employed within a robust Pipeline utilizing a **ColumnTransformer** for parallel preprocessing of numeric and categorical features.

- Tuning Method: RandomizedSearchCV (50 iterations, 5-fold CV)
- Objective: Minimize prediction error (RMSE)

Random Forest was chosen for both tasks due to its stability, resistance to overfitting, and ease of interpretation of feature importance.

Model Evaluation and Key Results

The models demonstrated strong performance in their respective domains, achieving high accuracy in classification and moderate variance explanation in regression.

League Winner Classification

0.9615

Overall Accuracy

0.702

Mean CV F1 Score

The model excels at identifying non-champions (Precision = 0.99) but maintains a respectable recall for the champion class (Recall = 0.83). Early season predictions successfully identified winners like Manchester United and Blackburn Rovers.

Top Assists Regression

2.74

RMSE (Root Mean Square Error)

0.393

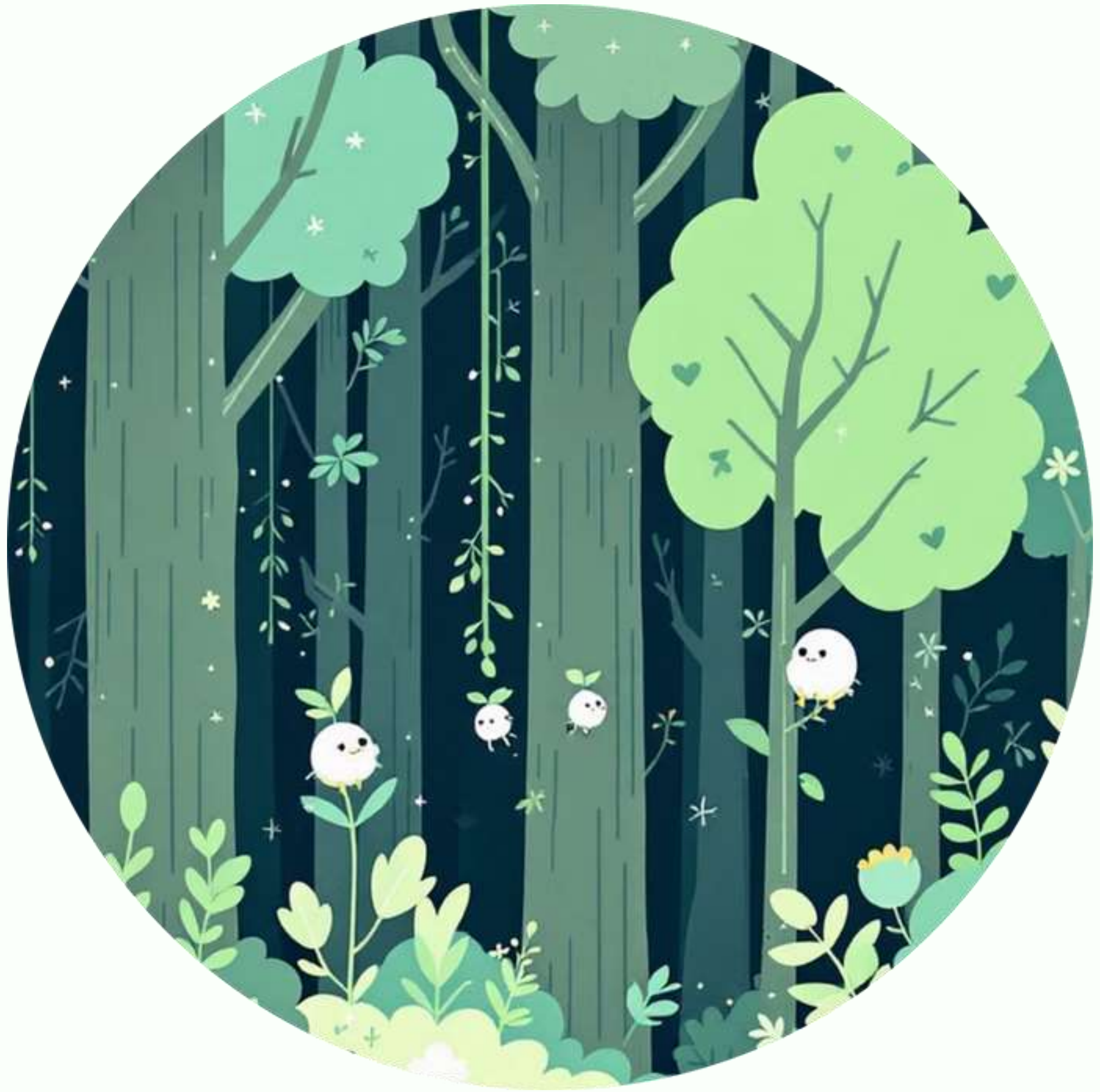
R^2 Score

The R^2 score indicates the model explains approximately 39% of the variance in total assists, which is reasonable given the inherent unpredictability of human performance.

- **Key Features:** Key_Passes_per_90, Assists_prev_per_90, Minutes_Played, Club_xG, Big6_Club_Feature.

Insights and Interpretation

Analyzing feature importance and prediction patterns provides valuable insights into the key statistical drivers of both team success and individual playmaking ability.



Dominant Club Identification

The League Winner model effectively utilizes historical strength metrics (attack_strength, prev_points) to reliably identify historically dominant clubs.

Assists Driven by Playtime and Opportunity

The Top Assists model confirms that prediction power is strongly tied to playing time, volume of key passes, and the attacking quality of the player's club (Club_xG).

Model Suitability

The Random Forest algorithm proved highly effective for both tasks due to its inherent ability to handle complex feature interactions typical of football data.

Conclusions and Future Work

The AI ScoreSight project successfully demonstrated the application of robust ML techniques to complex football prediction problems, setting a foundation for advanced analytics tools.

Project Success

Successfully built effective data preprocessing, feature engineering, and ML pipelines for both classification and regression tasks.

Integrate Live Data

Future models should incorporate real-time match data, such as live xG differentials, possession metrics, and defensive pressures, for in-season adjustments.

Advanced Modeling

Testing higher-performing gradient boosting models like XGBoost and LightGBM to potentially increase R^2 and classification metrics.

Real-World Deployment

The models are foundational for developing analytical dashboards, betting algorithms, or fan prediction apps that leverage data-driven insights.