The University of Texas at Austin
Cockrell School of Engineering

Premier League Match Predictions

Group 3: Alex Lozano, Luke Venkataramanan, Max Wiesenfeld

Introduction

Problem Overview

- The English Premier League is the most watched sports league in the world
- Billions of dollars flow through soccer decisions
  - Analysts and fans rely on predictive models for decision-making and entertainment
- Our Goal: Engineer features and create machine learning models to predict Premier League match outcomes

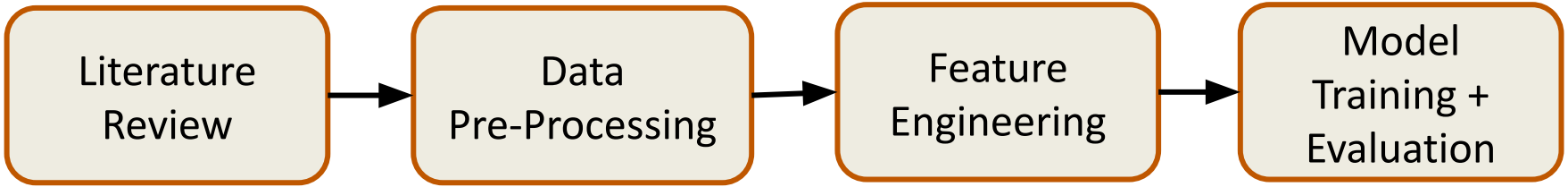
Why Predicting Outcomes is Hard

Soccer is very unpredictable, presenting many challenges:

- Due to the nature of the sport, results are notoriously noisy
- Recent results (“form”) strongly influence outcomes
- It is difficult to find and structure relevant data sources
    - Injuries, management, and tactics greatly impact games

Project Scope

- Review literature and the state-of-the-art
- Identify features and models to explore
- Gather and clean data
- Engineer feature matrix and principal component analysis (PCA)
- Train and tune models
- Analysis
- Future extensions: Integrate with web server and containerize



Methodologies

Current State-of-the-Art

"Predicting Football Match Outcomes Using Machine Learning Algorithms" (Heijboer, 2022)

- Used Premier League data up to 2016
- Compared classic ML models: Logistic Regression, SVM, Random Forest
- Best model: Random Forest (53.7% accuracy)

"Survey of Soft Computing Methods to Predict Soccer Win/Loss Probability" (Morgan, 2024)

- Compared classic ML to fuzzy logic and other soft computing fields on many data sets
- Highest accuracy was in the low-50s to low-60s using only pre-match features
- Bookmaker odds + rolling form features are most important
- Chronological splits and ensembles methods performed the best

Data and Features

Data sources:

- Football-Data.co.uk
- TransferMarkt
- FootballCritic

Feature Engineering:

- Previous game rolling statistics to quantify form
  - Wins, points, goals scored, goals conceded, shots, fouls, etc.
- Computed pre-match Elo ratings
- Averaged bookmaker odds
- Total squad value
- Possession

Models

Models:

1. Logistic Regression
2. Random Forest
3. Support Vector Machine (SVM)
4. XGBoost
5. Multilayer Perceptron Feedforward Neural Network (MLPFFNN)
6. Naive Bayes
7. Voting Ensemble

Significance to State-of-the-Art

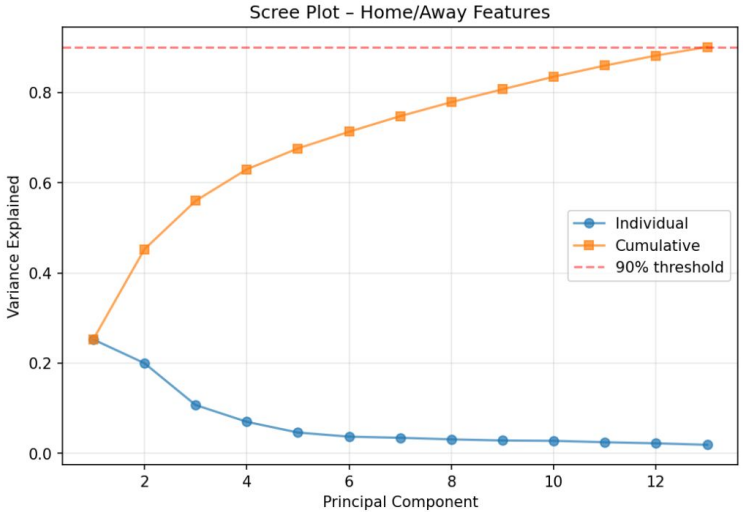
- Highly configurable and works across various time frames
  - Deals with new teams
- Added models (XGBoost, MLPFFNN, Naive Bayes, Voting)
- New Features (Average odds, Squad value, Possession)
- Consistently updatable and can be used in real time

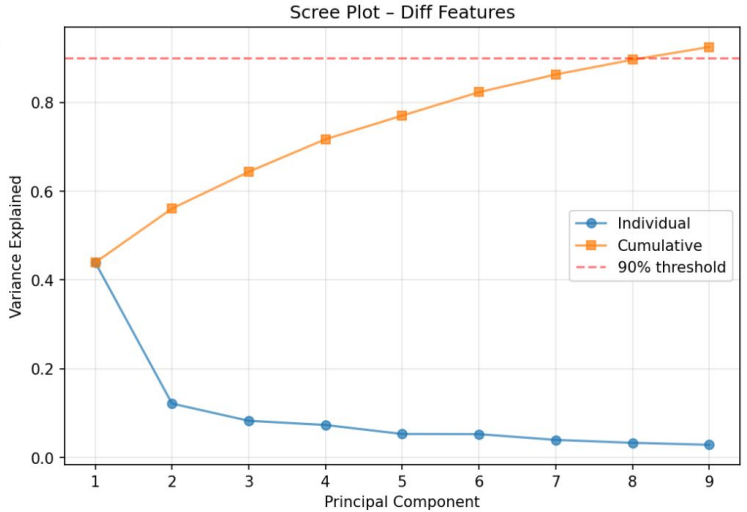
Results

Principal Component Analysis

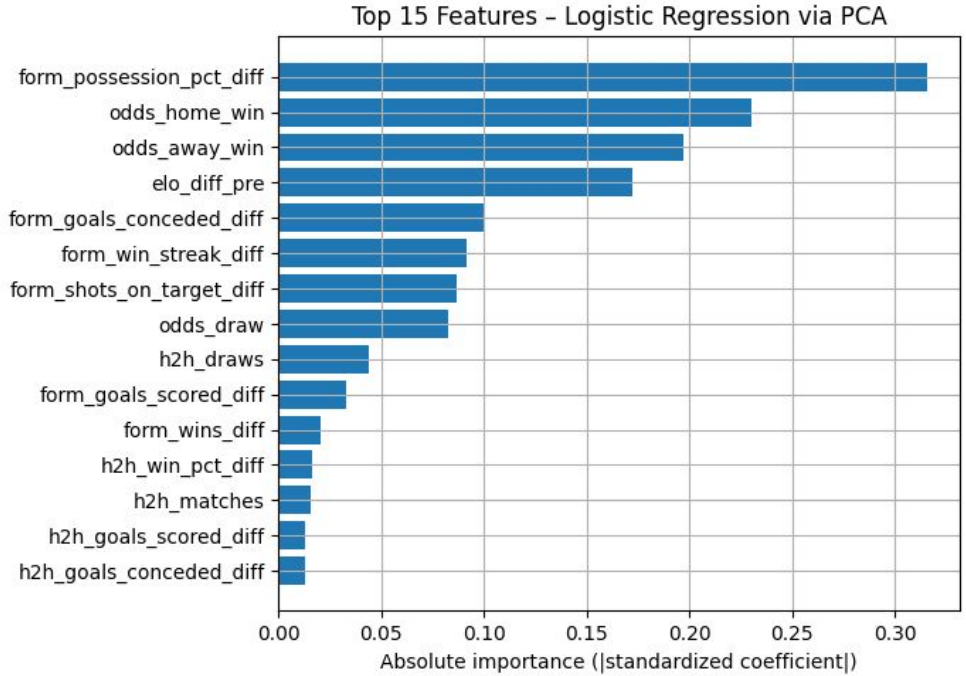
Principal Component Analysis (PCA) was used to reduce the dimensionality of our data:

- Elo
- Squad values
- Team head-to-head (H2H) history
- Possession statistics
- Diff features (*home - away* instead of separate home and away)

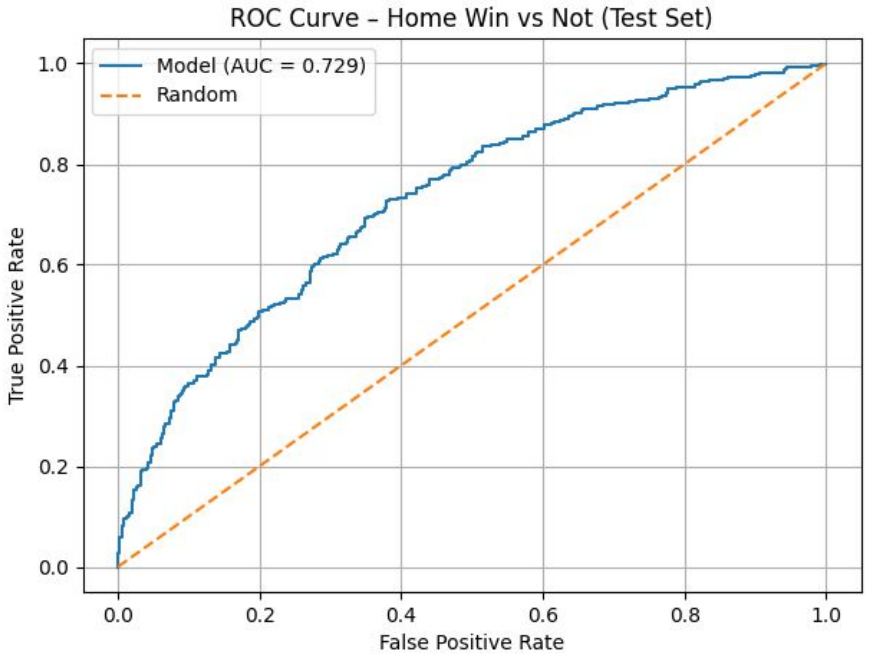




Principal Components and ROC Curve



PCs by significance



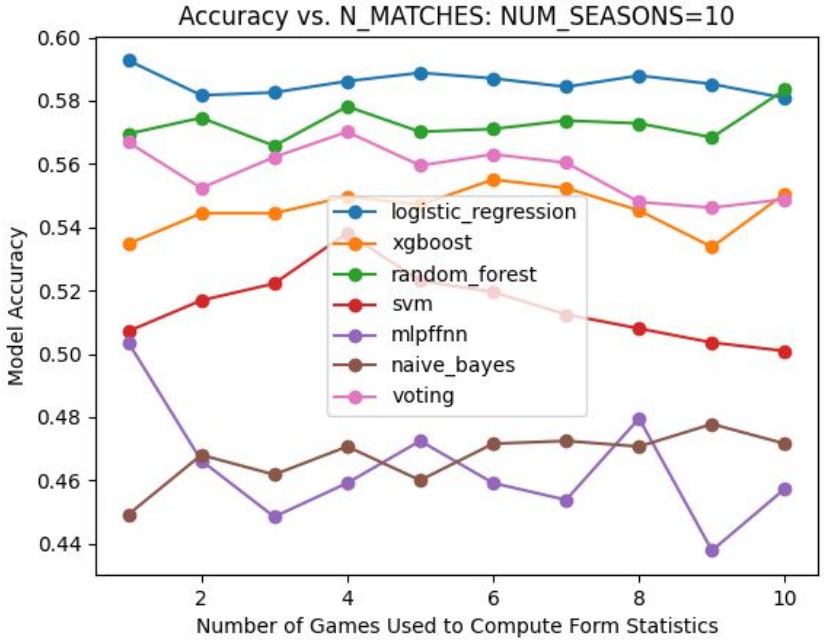
ROC Curve

Model Performance - Table

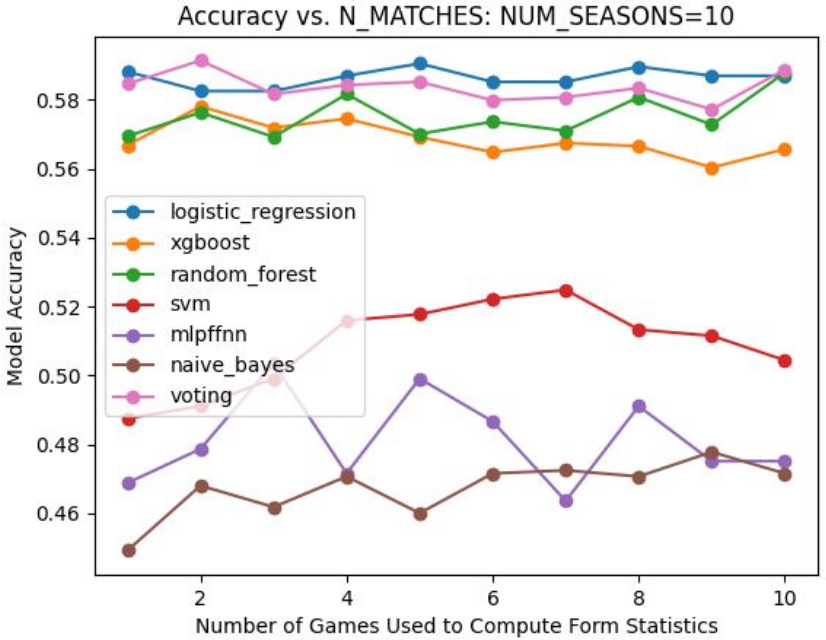
Model	Peak Training Accuracy	Peak Testing Accuracy
Logistic Regression	0.58531	0.59058
Random Forest (RF)	0.57046	0.58259
XGBoost	0.56778	0.58170
Support Vector Machine	0.52285	0.52486
MLP Feedforward NN	0.50266	0.50355
Naive Bayes	0.47373	0.47779
Voting Ensemble	0.58454	0.59058
Heijboer SOTA RF		0.537

Model Performance - Plots

- Accuracy was lower for models with higher draw prediction percentage
- After tuning voting weights and hyperparameters, the Voting Ensemble also performed very well and XGBoost saw notable upticks



Before tuning hyperparameters

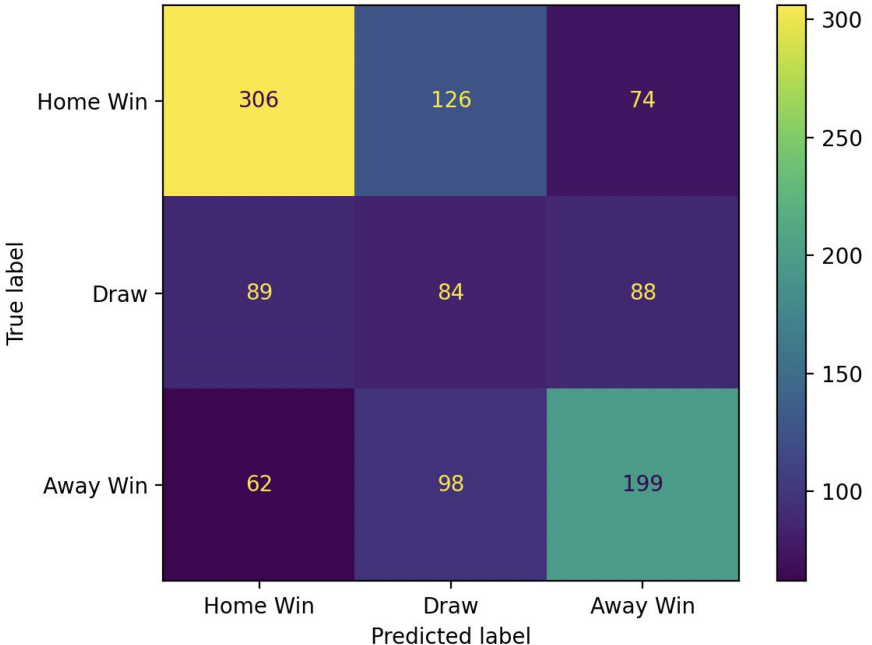


After tuning hyperparameters

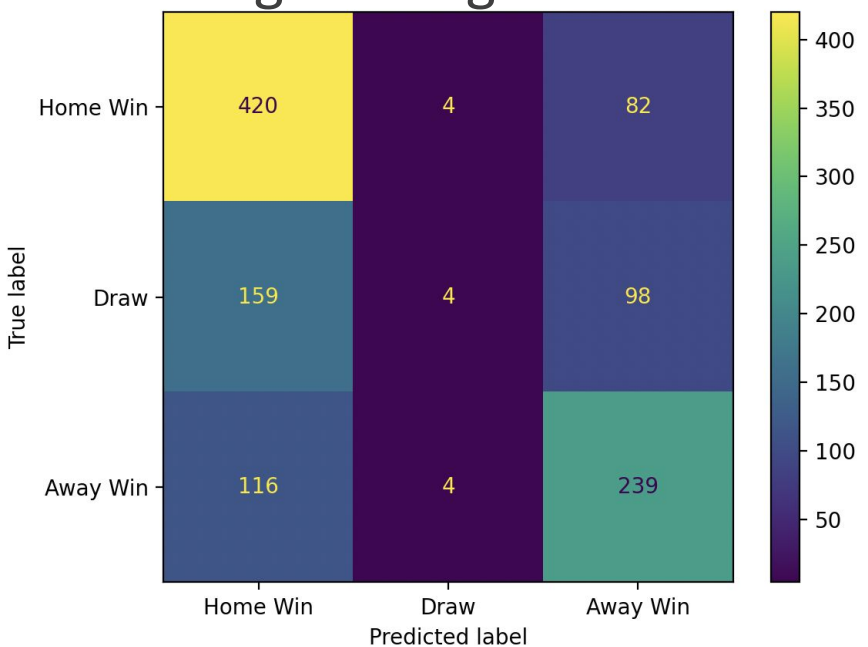
Draw Comparison

- Draws are difficult to predict – with significantly lower odds, linear models like logistic regression struggle to ever probabilistically predict a draw
- SVM and other nonlinear models carve out more favorable decision boundaries for predicting draws

SVM



Logistic Regression



Conclusion

- Logistic Regression and a Voting ensemble reach about 59% accuracy, beating accuracy of similar featured SOTA models
  - With highly engineered inputs (Elo, squad value, odds), linear decision boundaries perform best
  - XGBoost and Random Forest likely chase short-lived nonlinear patterns (say, a brief period of stellar coaching) instead of improving true out-of-sample accuracy
- 59% accuracy is vastly better than guessing (outcomes being win/loss/draw means the trivial guessing model only achieves 33% accuracy)
  - Independence from in-game statistics like halftime goals enables true pre-match prediction unlike many SOTA models

Outlook

- Model accuracy could still be refined by adding informative features which are difficult or costly to obtain
  - Injury data (requires all-time roster database, injuries, starters, etc.)
  - Weather (large database & costly for precise data)
- With our current accuracy, the model could be used in various applications like sports betting
  - Consistently beating predictions opens opportunities for informed bets
  - Eventually, this predictive model could be used to arbitrage across sportsbooks

References

Heijboer, M. (2022). *Predicting football match outcomes with classical machine learning methods: A comparative study across European leagues* (Master's thesis, Tilburg University). Tilburg University Repository. https://arno.uvt.nl/show.cgi?fid=160932

Morgan, K. A., Grant, E. S., & Kim, E. (2024). *Survey of soft computing methods to predict soccer win/loss probability*. In **Proceedings of the 2024 International Conference on Information System and Data Mining (ICISDM '24)** (pp. 1–7). ACM. https://doi.org/10.1145/3686397.3686398

Football-Data.co.uk. (n.d.). *Historical football results & betting odds data archive*. Retrieved December 1, 2025, from Football-Data.co.uk: https://www.football-data.co.uk/

Transfermarkt. (n.d.). *FAQ / Data administration*. Retrieved December 1, 2025, from Transfermarkt: https://www.transfermarkt.com/intern/faq

Q&A