Product Requirements Document (PRD)  
Premier League Winner 2025/26 Prediction

# 1. Executive Summary

This project delivers a machine‑learning service that predicts, in real‑time, the probability that each Premier League club will win the 2025/26 season outright. The model ingests historical match results, squad valuations, wages, injuries, manager tenure, advanced player metrics and betting odds to generate calibrated title‑winning probabilities for the 2025/2026 Premiier League season.

# 2. Goals & Success Metrics

**Primary Goal:**Achieve a model that correctly identifies the eventual champion within the top‑3 predicted clubs before the first ball of the season (August 2025) and maintains a Brier score ≤ 0.20 across the season.  
**Success Metrics:**

* • Top‑1 accuracy on historical back‑tests ≥ 30 % (random baseline 5 %).
* • Top‑3 recall ≥ 75 %.
* • Brier score ≤ 0.20 and log‑loss ≤ 0.85 on weekly probability updates.
* • Calibration slope between 0.9 and 1.1.
* • Stakeholder adoption: at least 5 internal users consume the dashboard weekly.

# 3. Stakeholders

• \*\*Product Owner\*\*: Shaheen Behbehani  
• \*\*Data Science Lead\*\*: TBD  
• \*\*Engineering\*\*: Data Engineering squad for ETL, MLOps squad for deployment  
• \*\*Design & UX\*\*: Dashboard design team  
• \*\*End Users\*\*: Internal analytics team, sports journalists, partner betting platforms

# 4. Personas & Use Cases

• \*\*Analyst Alex\*\* wants a weekly probability table to contextualize match previews and long‑term narratives.  
• \*\*Trader Taylor\*\* at a partner sportsbook wants a sanity‑check signal against market odds and early alerts to mispriced outright lines.  
• \*\*Executive Eve\*\* at a club wants to benchmark her team’s title chances versus investments in wages and transfers.

# 5. Scope

\*\*In‑Scope for MVP:\*\*

* • Automated ETL that refreshes all source datasets daily.
* • Feature engineering pipelines that aggregate club‑level season and match‑level stats.
* • Gradient Boosting (XGBoost) ensemble trained on seasons 2010/11‑2024/25.
* • Batch scoring every Friday 06:00 UTC and ad‑hoc scoring after each match.
* • Streamlit or Tableau dashboard showing probabilities, trend charts and key drivers.

\*\*Out of Scope (Phase 2+):\*\*

* • Player‑level contribution model (SHAP) to show per‑player impact.
* • Simulation of full table positions beyond champion.
* • Integration with live betting odds API for market‑level comparison.

# 6. Data Sources

• \*\*premier\_league\_fixtures\_2025\_2026.csv\*\* – full 2025/26 season fixture list with dates, home/away teams, kickoff times, and scheduling notes.

Datasets catalogued in the \*Premier League Data Inventory & Schema Reference\*:  
• \*\*E0.csv Match Results & Odds\*\* (1993/94‑2023/24) – base match outcomes and bookmaker prices.  
• \*\*Attendance Data.csv\*\* – stadium capacity & attendance (club‑level).  
• \*\*Club Value.csv\*\* – squad market valuations as of 1 Jul 2025.  
• \*\*Club wages.csv\*\* – wage bill 2024/25.  
• \*\*Injury list 2002‑2016.csv\*\* – injury incidents for historical burden features.  
• \*\*player\_salaries.csv\*\* – player‑level wages 2024/25.  
• \*\*players\_epl\_14‑15.csv\*\* – player performance metrics 2014/15‑2021/22.  
• \*\*shots\_epl\_14‑15.csv\*\* – event‑level shot data for xG features.  
• \*\*Possession data 24‑25.csv\*\* – 2024/25 team stats.  
• \*\*Premier League Managers.csv\*\* – manager tenure & experience.

# 7. Technical Approach

Feature engineering will blend historical season aggregates (e.g., last‑5‑season rolling xG differential), current squad strength (market value, wage bill), manager stability, injury burden, and bookmaker implied probabilities. The model will predict the log‑odds of winning the title and convert to calibrated probabilities via isotonic regression.

# 8. Evaluation & Monitoring

* • Cross‑validation with time‑based split (train up to season N‑1, validate on season N).
* • Back‑test on seasons 2014/15‑2024/25 and compare against implied odds baseline.
* • Drift monitoring on feature distributions (KL divergence) and probability calibration (expected vs. observed).
* • Weekly champion probability movement alerts > 5 pp.

# 9. Timeline & Milestones

25 Jul 2025: Dataset ingestion & schema validation

10 Aug 2025: Feature engineering v1 complete

24 Aug 2025: Model benchmark & back‑test

31 Aug 2025 (season kick‑off): Dashboard MVP launch

30 Sep 2025: First monthly review & calibration check

# 10. Risks & Mitigations

• Data freshness lag for player injuries and transfers — \*Mitigation\*: Automate web‑scraping & manual QC twice‑weekly

• Transfer window volatility alters squad strength — \*Mitigation\*: Include transfer net spend proxy and update player valuations post‑window

• Overfitting to bookmaker odds features — \*Mitigation\*: Apply regularization, validate against seasons without odds leak

• Model interpretability concerns — \*Mitigation\*: Provide SHAP summary plots and feature documentation

# 11. Open Questions

* • Do we have legal clearance to display bookmaker odds on the public dashboard?
* • What is the business priority for second‑place and top‑4 probabilities?
* • Should we expose API endpoints for external partners?

# 12. Sign‑Off

Product Owner: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_

Data Science Lead: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_

Engineering Lead: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_