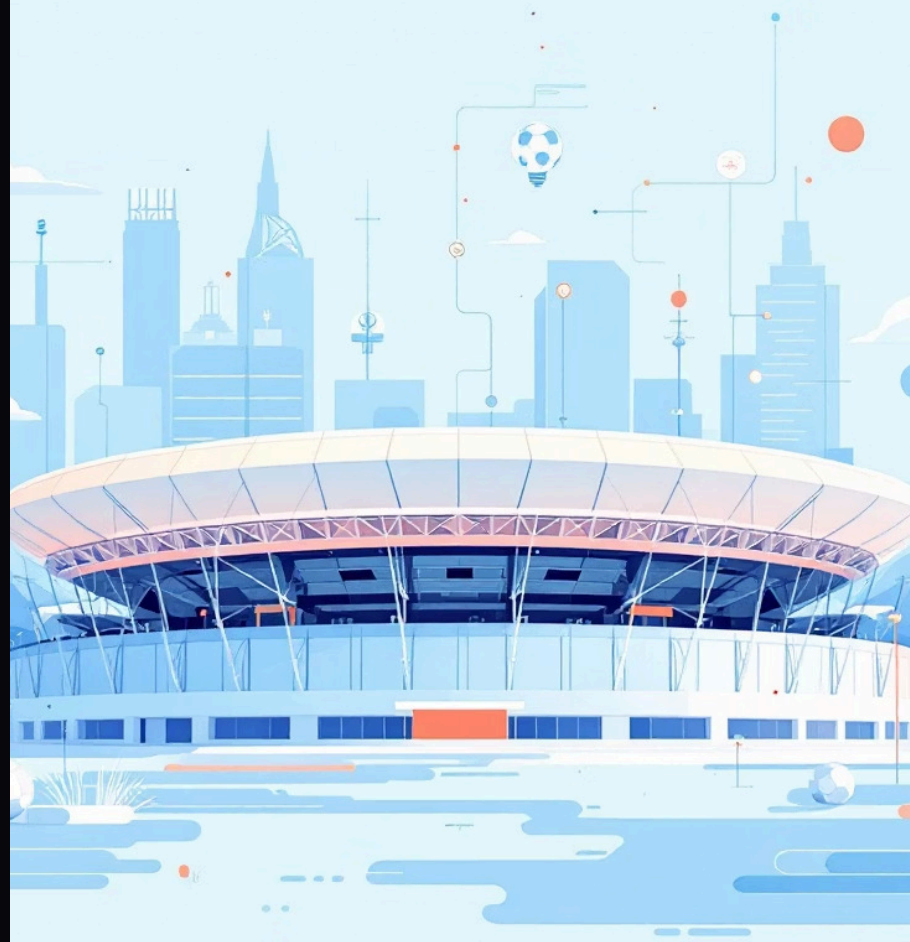
AI Football Match Predictor

A Data-Driven Engineering Journey

From raw data to deployed web application - a comprehensive exploration of machine learning engineering principles applied to sports prediction.

Arvind K N
B.Tech Computer Science(2nd year)
National Institute of Technology Calicut
AI intern at Infosys springboard



Project Motivation: Beyond Simple Prediction

End-to-End Pipeline

Master the complete ML workflow: data collection through deployment, experiencing every challenge of production systems.

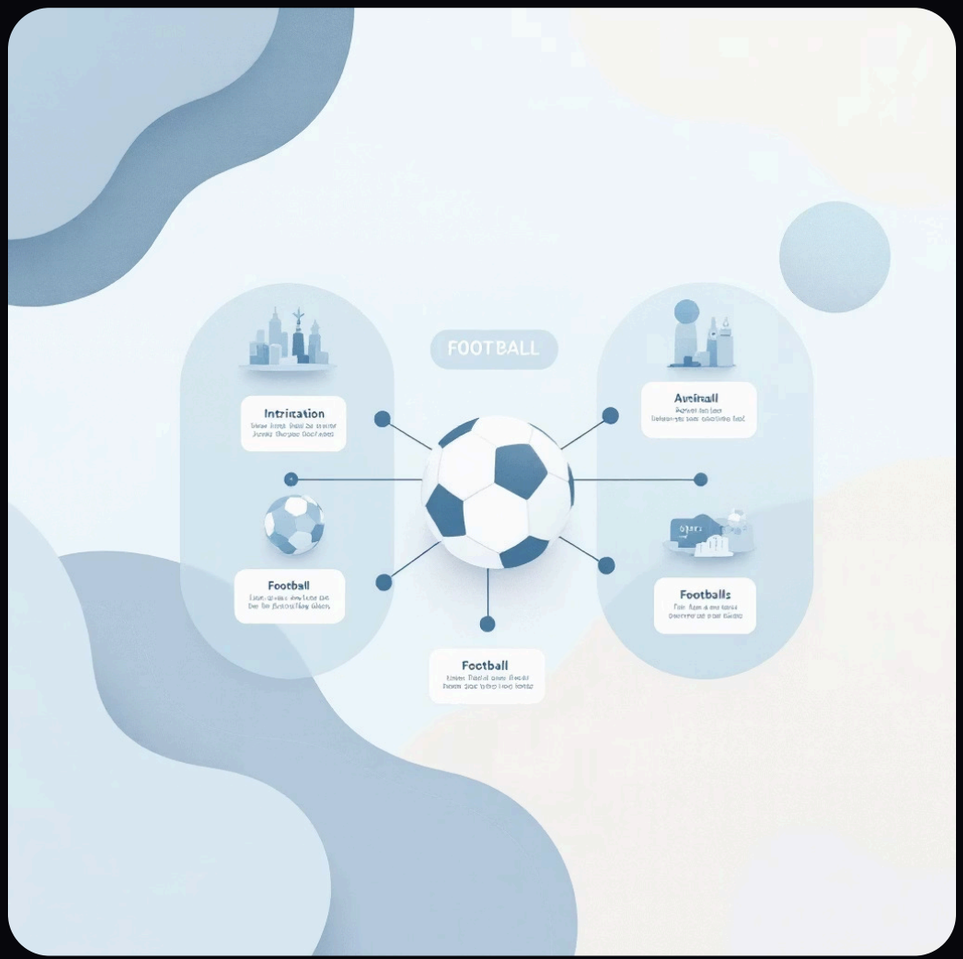
Portfolio Excellence

Build a showcase project demonstrating professional AI engineering skills and real-world problem-solving capabilities.

The Challenge: Why Football Prediction is Hard

The Complexity Problem

Football represents a chaotic system with countless hidden variables. Success requires accounting for this inherent complexity through sophisticated modeling approaches.



High Randomness

Lucky bounces, deflected shots, and referee decisions can completely alter match outcomes

Hidden Human Element

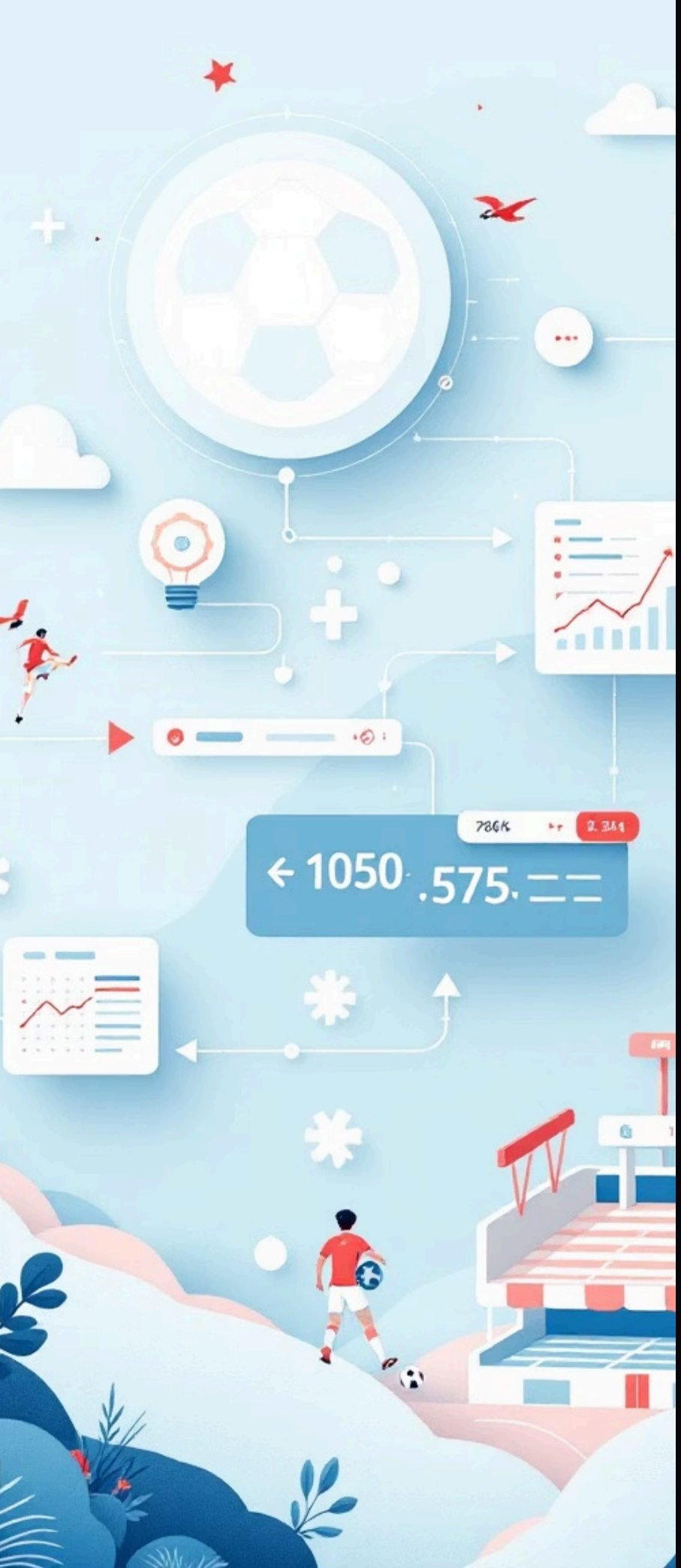
Team morale, psychology, and real-time tactical adjustments remain invisible to historical data

Draw Problem

Draws create severe class imbalance, making accurate prediction extremely challenging

Dynamic Systems

Teams and players constantly evolve - form and ability change throughout seasons



Advanced Feature Engineering: The Foundation

The most critical breakthrough came from creating powerful features that provide deep contextual understanding, going far beyond simple statistics.

1

Base Features

Head-to-head records, betting odds, and recent form metrics using rolling averages of goals, shots, and possession statistics.

2

Temporal League Rank

Revolutionary approach: reconstructed the complete league table state before every single historical match to capture true team standings.

3

Temporal Team Strength (Elo)

Adaptive rating system measuring teams' true, evolving power levels over time using dynamic strength calculations.

Technical Deep Dive: Elo rating formula: New_Rating = Old_Rating + K × (Actual_Score - Expected_Score)

The Engineering Journey: Three Model Approach

Following scientific methodology, I implemented rigorous experimentation to identify the optimal solution through systematic comparison.

Model	Approach	Key Finding
Logistic Regression	Simple Linear Baseline	Failed completely - proved high non-linearity (Draw F1: 0.03)
XGBoost Hybrid	Complex "Divide & Conquer"	Promising real-world scenarios but lower test performance
Pure CatBoost	Powerful All-Rounder	**Champion:** Best balanced performance and Draw F1-score

> Classic Occam's Razor: After exploring complex architectures, the data proved a single powerful model was the most elegant solution.

Champion Model Performance

64.59%

Overall system accuracy on held-out test set

Classification Results

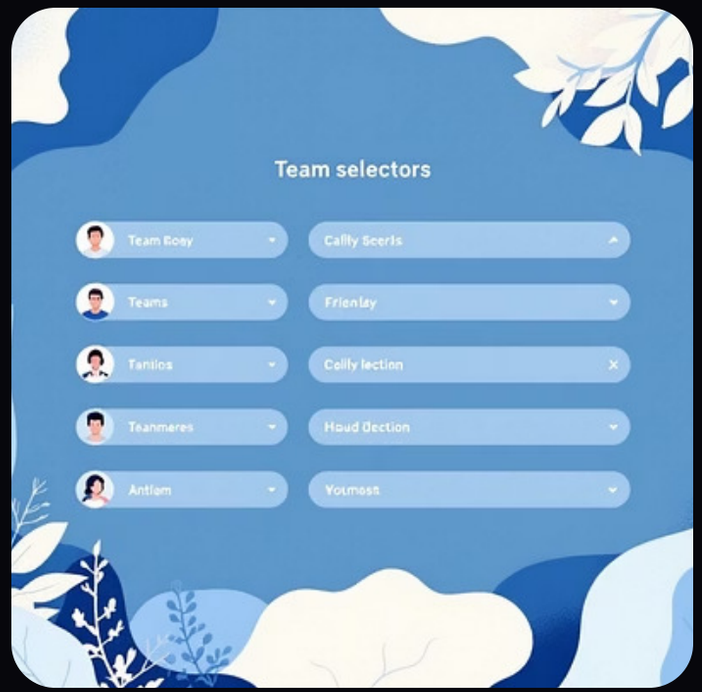
Outcome	Precision	Recall	F1-Score
Away Win	0.72	0.69	0.70
Draw	0.39	0.49	0.43
Home Win	0.79	0.70	0.75

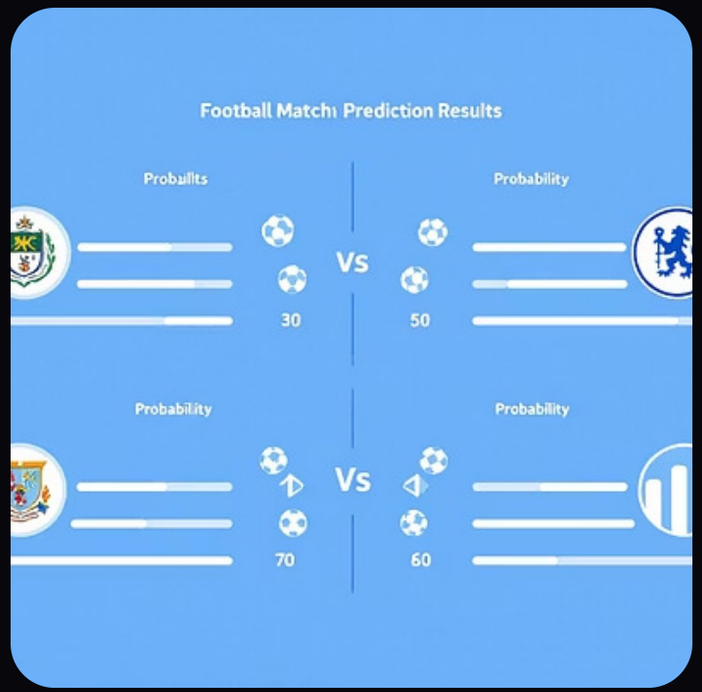
Key Achievements

- **Weighted Avg F1: 0.66** - Strong overall model health
- **Draw F1: 0.43** - Exceptional performance on hardest prediction class
- **Balanced approach** - No single outcome dominates predictions

From Model to Product: Live Application

The champion model was integrated into a professional, user-friendly web application featuring real-time data integration and intuitive interface design.





Interactive Interface

Clean team selection with visual feedback and professional UI components

Probability Visualization

Professional probability bars with team logos and celebratory confetti effects

Real-Time Data

Live ranking integration ensuring predictions use current league standings

Engineering Wisdom & Technical Insights

Feature Engineering is King

Quality of features, especially temporal ones, had greater impact than any single model architecture choice. Data preparation determines success.

Power of Experimentation

Rigorous testing of multiple architectures was essential for discovering the true champion model. Scientific methodology beats intuition.

Information Gap Reality

Primary limitation: lack of live player-level data (injuries, current form) explains performance differences from commercial prediction services.

Thank You

Ready to discuss technical implementation details, model architecture decisions, and future enhancement opportunities.