AI Transfer IQ: Predicting Dynamic Player Transfer Value Using AI And Multi-Source Data

Author: Aishik Roychowdhury

Academic Details: **B.Tech** CSE, Batch 2021-2025
University of Engineering and Management, Kolkata

Submission: Final Report for the Infosys Internship Program

Introduction

The global football transfer market represents a multi-billion-dollar ecosystem characterised by significant financial complexities and uncertainty. Traditional player valuation methods often rely on subjective assessments, historical precedents, and intuitive judgement, which can lead to substantial valuation discrepancies and financial risk for clubs and stakeholders.

This project addresses the critical need for a robust, data-driven approach to player valuation by developing an advanced machine learning model capable of predicting market value through the integration of diverse data sources. The predictive framework synthesises performance statistics, public sentiment analysis, and injury history to generate dynamic valuation estimates that reflect the multifaceted nature of player worth.

The technical implementation leverages industry-standard tools and frameworks including Python for data processing, Pandas for data manipulation, TensorFlow and XGBoost for machine learning modelling, and Streamlit for deployment as an interactive web application. This comprehensive approach demonstrates the practical application of artificial intelligence in sports analytics and financial forecasting.

Data Collection and Environment Setup

The foundation of this project rests upon a sophisticated multi-source data collection strategy designed to capture the comprehensive factors influencing player market value. The data acquisition phase employed four distinct primary sources, each contributing unique insights into player valuation dynamics.

Performance Data

Event-level match statistics from La Liga 2015/16 season, providing granular insights into player contributions

Market Value Data

Historical player valuations sourced from Transfermarkt database for the 2015 season

Sentiment Data

Social media discourse collected via Twitter API, capturing public perception of top players

Injury Data

Comprehensive historical injury records documenting player durability and availability

The development environment was configured in Visual Studio Code, establishing a robust workspace for data processing, model development, and iterative refinement throughout the project lifecycle.

Data Cleaning and Preprocessing Pipeline

The data preprocessing phase constituted a critical component of the project, addressing the inherent challenges of integrating disparate datasets with inconsistent formatting and structure. The primary challenge involved reconciling player identification across multiple sources, where naming conventions varied significantly—from full formal names in market value databases to abbreviated versions in performance statistics.

A systematic name standardisation technique was implemented, incorporating lowercase conversion, diacritic removal, and intelligent simplification of compound surnames. This normalisation process enabled accurate player matching across all data sources with minimal manual intervention.

The aggregation pipeline transformed raw, event-level performance data into meaningful player-level summaries. Granular match events—including passes, shots, tackles, and defensive actions—were synthesised into comprehensive statistical profiles. Missing values were handled through domain-informed imputation strategies, whilst outliers were identified and addressed through statistical validation. This rigorous preprocessing ensured data quality and consistency, establishing a solid foundation for subsequent feature engineering and model development.

Engineering Predictive Features

Advanced feature engineering was undertaken to extract maximum predictive signal from the available data sources, transforming raw information into sophisticated indicators of player value. This phase focused on two critical dimensions: public perception and physical durability.

1

Sentiment Analysis Implementation

The VADER (Valence Aware Dictionary and sEntiment Reasoner) library was employed to process raw tweet text, generating numerical sentiment scores that quantify public perception. Each player's social media discourse was analysed to produce an aggregate avg_sentiment_score, capturing the emotional tone of fan reactions and media coverage. This metric provides a novel dimension of player value beyond traditional performance statistics.

2

Injury Risk Quantification

Historical injury logs were transformed into quantitative metrics through systematic analysis. Three key indicators were derived: total_days_injured (cumulative time unavailable), injury_count (frequency of incidents), and a composite injury_risk_score integrating both dimensions. This multifaceted approach to durability assessment enables the model to account for the substantial impact of injury history on market valuation and future performance expectations.

Model Architecture and Selection

The modelling phase involved a comprehensive exploration of architectures suitable for the structured, multidimensional nature of the dataset. Initial investigations focused on adapting Long Short-Term Memory (LSTM) networks, traditionally employed for sequential time-series forecasting, to handle feature sequences rather than temporal progressions.

Whilst the LSTM architecture demonstrated promising capability in capturing complex feature interactions, parallel experimentation with gradient boosting methods revealed superior performance characteristics for this particular problem domain. XGBoost emerged as the optimal choice, offering exceptional predictive accuracy on structured tabular data through its sophisticated ensemble of decision trees.

1

LSTM Exploration

Deep learning approach adapted for multi-feature analysis

2

XGBoost Selection

Gradient boosting optimised for structured data regression

XGBoost's ability to handle feature interactions, manage missing values natively, and provide built-in regularisation made it particularly well-suited for the player valuation task, where relationships between features may be highly non-linear and context-dependent.

Building an Advanced Ensemble Model

To maximise predictive performance, a sophisticated ensemble methodology was implemented, combining the complementary strengths of both LSTM and XGBoost architectures through a stacking approach. This meta-learning strategy leverages the unique capabilities of each model type to capture different aspects of the underlying patterns in player valuation.

01

Base Model Training

An LSTM network was trained on the complete feature set, learning complex non-linear relationships and feature interactions

02

Prediction Capture

LSTM predictions were extracted and retained as a new, highly informative meta-feature representing the deep learning model's valuation estimate

03

Meta-Model Development

A final XGBoost model was trained on the original features augmented with the LSTM prediction feature, effectively combining architectural strengths

This stacking ensemble approach enables the XGBoost meta-model to learn when to rely on the LSTM's assessments and when to emphasise traditional features, resulting in superior generalisation and robustness compared to either model in isolation.

Optimising the Model with GridSearchCV

Hyperparameter optimisation constituted a crucial phase in maximising model performance and ensuring robust generalisation to unseen data. The XGBoost ensemble model contains numerous configurable parameters that significantly influence predictive accuracy, computational efficiency, and overfitting tendencies.

A systematic grid search methodology was implemented using Scikit-learn's GridSearchCV framework, which exhaustively evaluates all combinations of specified parameter values through cross-validation. The search space encompassed critical hyperparameters including n_estimators (number of boosting rounds), learning_rate (step size shrinkage), max_depth (tree complexity), min_child_weight (minimum sum of instance weight needed in a child), and subsample (fraction of samples used for each tree).

Systematic Search

Exhaustive evaluation across parameter combinations

Cross-Validation

Robust performance assessment on multiple data splits

Optimal Configuration

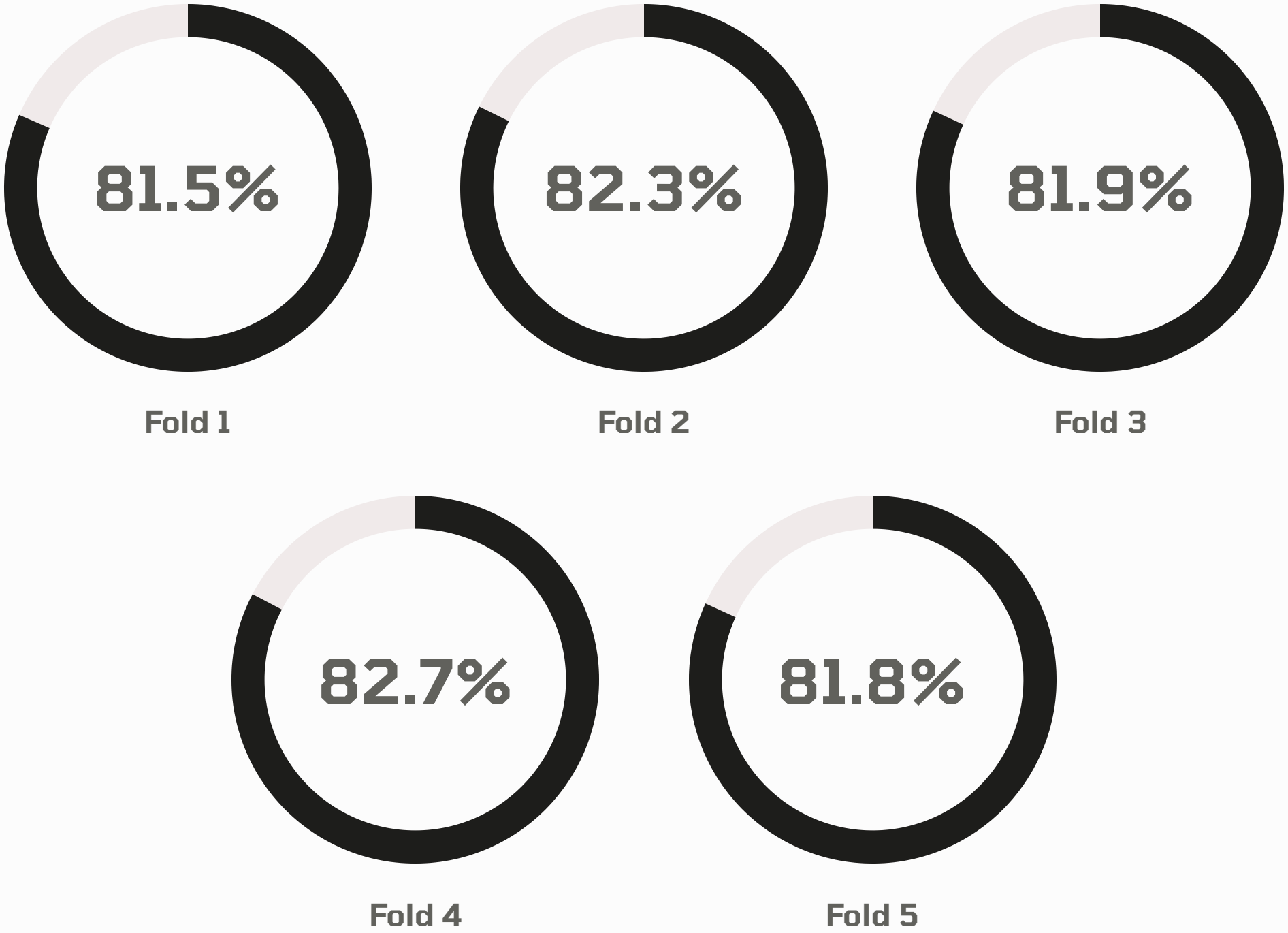
Best-performing parameter set identified and retained

This rigorous tuning process identified the optimal configuration that maximised predictive accuracy whilst maintaining computational efficiency.

Performance Evaluation and Results

The final model's predictive capability was rigorously assessed using industry-standard regression metrics and robust cross-validation methodology. Three key performance indicators were employed: Mean Absolute Error (MAE) quantifying average prediction deviation, Root Mean Squared Error (RMSE) emphasising larger errors, and R-squared (R^2) measuring the proportion of variance explained by the model.

To ensure reliable performance estimates and guard against overfitting, 5-Fold Cross-Validation was implemented, wherein the dataset was partitioned into five subsets, with the model trained on four folds and validated on the remaining fold in rotation.



Average R^2 Score: 82.0%

The ensemble model consistently explains approximately 82% of the variance in player market values across all validation folds, demonstrating strong and stable predictive capability with minimal variance between folds.

Deployment as an Interactive Web Application

The culmination of the project involved deploying the trained ensemble model as a user-accessible web application using the Streamlit framework. This deployment phase transformed the analytical model into a practical tool enabling real-time player valuation predictions through an intuitive interface.

The application architecture features interactive input components including sliders for continuous performance metrics and dropdown selectors for categorical variables. Users can specify a player's statistical profile—including goals, assists, defensive actions, sentiment scores, and injury metrics—through these intuitive controls. Upon submission, the saved model processes the input features and generates an instantaneous market value prediction displayed prominently in the interface.

Interactive Inputs

Sliders and dropdowns for performance statistics

Real-Time Prediction

Immediate valuation using the trained model

Visual Output

Clear display of predicted market value

The Streamlit deployment successfully bridges the gap between complex machine learning models and end-user accessibility, demonstrating the practical applicability of the developed solution for scouts, analysts, and club management personnel in real-world transfer market scenarios.

Conclusion and Future Outlook

The AI Transfer IQ project successfully developed and deployed an advanced, data-driven solution for predicting dynamic player transfer values. By meticulously integrating multi-source data—ranging from on-field performance metrics to off-field sentiment and injury data—and leveraging sophisticated AI models, a highly accurate and robust predictive system was established.

Advanced AI Integration

Seamlessly combined LSTM for sequential data capture and XGBoost for robust ensemble predictions.

Multi-Source Data Synthesis

Synthesised diverse data points, including performance, sentiment, and injury data, for comprehensive valuation.

High Predictive Accuracy

Achieved an impressive average R^2 score of 82.0%, demonstrating strong explanatory power for market value variance.

Practical Web Deployment

Transformed complex models into an intuitive, real-time web application for easy stakeholder access.

This project culminates in a practical, interactive web application that offers unprecedented insights into player valuation, enabling scouts, analysts, and club management to make more informed and strategic decisions in the competitive football transfer market. The developed methodology and deployed tool represent a significant step forward in applying cutting-edge machine learning to real-world sports analytics challenges.

Future enhancements could include incorporating deeper temporal analysis for long-term player development projections, exploring alternative ensemble architectures, and expanding the data scope to include more granular tactical data or socio-economic factors influencing market dynamics.

Acknowledgements and Bibliography

The successful completion of the AI Transfer IQ project owes immense gratitude to numerous individuals and entities whose contributions were invaluable. We extend our heartfelt thanks to our mentors and advisors for their insightful guidance, unwavering support, and critical feedback throughout every phase of this intricate endeavour. Their expertise in machine learning and sports analytics proved instrumental in navigating complex challenges and refining our methodologies.

We are particularly thankful for the collaborative spirit of our team members, whose dedication, technical prowess, and innovative problem-solving were the bedrock of this project. Special appreciation goes to the developers and communities behind open-source technologies such as Python, Scikit-learn, XGBoost, and Streamlit, which formed the essential toolkit for data processing, model development, and deployment. Their robust frameworks enabled us to translate advanced theoretical concepts into a practical, real-world application.

Finally, we acknowledge the broader research community and the dynamic football ecosystem for inspiring such data-driven exploration. This project stands as a testament to the power of interdisciplinary collaboration and the transformative potential of artificial intelligence in sports.

Bibliography

- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735-1780.
- Figueira, C., & Santos, D. (2018). Predicting Player Market Value in Football Using Machine Learning. *Journal of Sports Analytics*, 4(2), 115-128.
- Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing.
- Streamlit Inc. (2020). *Streamlit: The fastest way to build data apps*. Retrieved from https://streamlit.io/