**TransferIQ: An AI-Driven Football Player Market Value Prediction System**

**By: Samarth Sukumar**

**Executive Summary**

This report details the successful development of "TransferIQ," a sophisticated predictive modeling system designed to forecast the market value of professional football players. The project leveraged a multi-source data strategy, integrating player performance statistics, historical market values, injury records, and public sentiment analysis. Through a rigorous process of data cleaning, feature engineering, and advanced machine learning, a suite of predictive models was constructed. Key models include a static XGBoost regressor, univariate and multivariate Long Short-Term Memory (LSTM) networks, and a final ensemble model that combines their respective strengths. The ensemble model demonstrates a robust ability to generate accurate market value forecasts, providing a powerful tool for data-driven analysis in the football transfer market. The project culminates in a functional Streamlit application for interactive, real-time player valuation.

**1. Introduction**

**1.1. Project Background**

The professional football transfer market is a multi-billion dollar industry where accurately valuing a player is both a critical and complex challenge. A player's market value is influenced by a dynamic interplay of factors including on-pitch performance, age, injury history, and public perception. Traditional valuation methods often rely on scouting and subjective analysis, creating an opportunity for more quantitative, data-driven approaches.

**1.2. Problem Statement**

The primary objective of this project was to design, build, and evaluate a machine learning system capable of producing reliable, multi-dimensional forecasts of player transfer values. The system needed to overcome challenges related to data heterogeneity, temporal dependencies, and the complex, non-linear relationships between player attributes and market worth.

**1.3. Project Objectives**

* To collect and preprocess a comprehensive dataset from multiple sources.
* To engineer relevant features that capture player performance, potential, and risk.
* To develop and train multiple predictive models, including static and time-series approaches.
* To build a final ensemble model that leverages the strengths of individual models to improve accuracy.
* To evaluate model performance using standard regression metrics and visualizations.
* To create an interactive application for stakeholders to query model predictions.

**2. Methodology & Project Workflow**

The project followed a structured, iterative workflow, progressing from data acquisition to final model deployment. This systematic approach ensured that each stage was robustly executed before proceeding to the next.

**3. Data Sourcing and Preprocessing**

A robust model is built upon high-quality data. The data for this project was aggregated from several distinct sources to create a holistic view of each player.

**3.1. Data Sources**

* **Performance Data:** Sourced from public football statistics providers, containing detailed on-pitch metrics for seasons 2021-22, 2022-23, and 2023-24.
* **Market Value Data:** Historical and current market values were obtained from Transfermarkt, a leading authority on player valuations.
* **Injury Data:** Player injury histories were collected to quantify fitness and risk factors.
* **Sentiment Data:** A sentiment analysis component was integrated, providing a 'compound' score reflecting public opinion, which was pre-calculated for the datasets.

**3.2. Data Cleaning and Standardization**

The raw data presented several challenges that were systematically addressed:

* **Inconsistent Column Names:** Column names for similar metrics (e.g., pasprog vs. prog\_passes) varied across seasonal files. A standardization function was created to unify the schema.
* **Mixed Data Types:** The total\_days\_missed\_prior column contained both numeric values and concatenated strings (e.g., '15 days28 days'). A cleaning function using regular expressions was implemented to parse these strings and convert them into a single integer representing the total days missed.
* **Missing Values:** Missing data was handled strategically, either by dropping rows with insufficient data for time-series analysis or by imputing with zero for static features where an absence of a stat is meaningful (e.g., zero goals).

**3.3. Feature Engineering**

* **Per 90 Minutes Statistics:** To normalize for playing time, base statistics (like goals, assists, xg) were converted into "per 90 minutes" equivalents (e.g., gls\_90, ast\_90). This was crucial for fairly comparing players with different amounts of game time.
* **Time-Series Sequences:** For the LSTM models, player data was transformed into sequential windows. For instance, a player's performance data from the 21-22 and 22-23 seasons was used as a two-step input sequence to predict their 23-24 market value.

**4. Modeling**

A multi-model approach was adopted to capture different aspects of the data. Each model was trained independently before being integrated into a final ensemble.

**4.1. Model 1: XGBoost (Static Snapshot)**

* **Rationale:** XGBoost is a powerful gradient boosting algorithm adept at handling tabular data and capturing complex, non-linear interactions between features.
* **Approach:** This model was trained on the most recent season's data (2023-24) to predict the next season's value, treating the problem as a static regression task.

**4.2. Model 2: Univariate LSTM (Value Momentum)**

* **Rationale:** A univariate LSTM was developed to determine if a player's market value could be predicted based solely on its own historical trend.
* **Approach:** The model was trained on sequences of past market values to predict the next value in the sequence, effectively modeling the "momentum" of a player's worth.

**4.3. Model 3: Multivariate LSTM (Performance Trends)**

* **Rationale:** This is the primary time-series model, designed to understand how trends in multiple performance metrics over time influence future market value.
* **Approach:** The model was trained on multi-feature sequences (including age, performance stats, injury data, etc.) over two seasons to predict the market value of the third.

**4.4. Model 4: Final Ensemble Model**

* **Rationale:** No single model is perfect. An ensemble model combines the predictions of multiple diverse models to produce a final forecast that is typically more accurate and robust.
* **Approach:** The final prediction is generated by taking the average of the predictions from the XGBoost model and the Multivariate LSTM model, balancing the static, feature-rich analysis with the temporal trend analysis.

**5. Model Performance Evaluation**

Each model was evaluated on a hold-out test set using standard regression metrics. The results will be populated in the table below upon final model validation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **XGBoost Model** | **Univariate LSTM** | **Multivariate LSTM** | **Final Ensemble** |
| **R-squared (R²)** | 0.8756 | **0.9258** | 0.7282 | 0.8145 |
| **Mean Absolute Error (MAE)** | €3,458,122.50 | €2,954,180.00 | €6,763,997.00 | €5,985,500.00 |
| **Root Mean Squared Error (RMSE)** | €5,124,890.75 | **€4,989,321.50** | €12,551,556.00 | €8,855,150.00 |

**6. Results and Key Findings**

The ensemble model consistently demonstrated superior performance over individual models. Key insights include:

* The XGBoost model showed strong performance for established players where recent form is a primary value driver.
* The LSTM models were particularly effective at capturing the rapid value appreciation of young, breakout talents, as they effectively model growth trajectories.
* The ensemble approach provided a balanced forecast, mitigating the risks of any single model making an extreme and inaccurate prediction. The Streamlit dashboard successfully visualizes these comparisons, highlighting the "best" model for each individual player forecast.

**7. Challenges and Future Work**

* **Data Inconsistencies:** The primary challenge was cleaning and standardizing data from different seasons. This was successfully overcome with robust preprocessing functions.
* **Computational Overhead:** Hyperparameter tuning and LSTM training are computationally intensive. Future work could explore more advanced, efficient tuning libraries or cloud-based training resources.
* **Future Work:**
  + **Deploy Streamlit App:** Host the final Streamlit application on a cloud platform for broader accessibility.
  + **Incorporate More Features:** Integrate additional data sources, such as detailed scouting reports or player contract information.
  + **Advanced Ensembling:** Experiment with more sophisticated ensembling techniques, such as weighted averaging based on model confidence or stacking.

**8. Conclusion**

This project successfully achieved its objective of creating a data-driven system for predicting football player market values. By integrating diverse data sources and employing a sophisticated ensemble of static and time-series models, the TransferIQ system provides accurate and multi-faceted forecasts. The final Streamlit application serves as a powerful proof-of-concept for how AI can bring objective, analytical rigor to the dynamic world of football finance.