Infosys Internship Report

TransferIQ: Dynamic Player Transfer Value Prediction using AI and Multi-Source Data

By: P. Veerababu**Organization:** Infosys**Duration:** August – October 2025

Abstract / Executive Summary

The football transfer market is increasingly volatile, with player valuations influenced by a complex interplay of performance metrics, injury history, contract details, and public sentiment. Traditional scouting and valuation approaches are limited by their reliance on historical performance and expert intuition, failing to capture dynamic market trends.

This project, *TransferIQ*, proposes a hybrid machine learning framework combining traditional regression models (Random Forest, XGBoost) with advanced time-series deep learning (Multivariate LSTM with Attention) to predict football player market values in Euro (€). The framework integrates heterogeneous data sources, including detailed match statistics (from StatsBomb), financial information (Transfermarkt), social media sentiment (Twitter, Reddit), and injury history.

Key Findings: The Multivariate LSTM model with Attention outperformed traditional models, accurately capturing temporal patterns in player performance, injury impacts, and sentiment-driven market fluctuations. Feature analysis via SHAP and Random Forest importance highlighted contract risk, recent performance metrics, and injury exposure as critical predictors. This methodology provides a scalable, interpretable, and robust framework for data-driven transfer valuation.

Table of Contents

I. Introduction

1.1 Problem Statement and Motivation

The modern football transfer market is characterized by high volatility and substantial financial stakes. Player valuation is affected by dynamic factors, including short-term form, injury history, and public sentiment, which are difficult to capture with traditional scouting or statistical methods. Existing approaches often rely on intuition, past performance, or historical market trends, failing to account for temporal shifts or external influences.

Research Goal: Develop a robust, data-driven machine learning model capable of predicting football player market values (€) by integrating diverse data sources—performance metrics, injury data, financial information, and social sentiment.

1.2 Project Objectives and Scope

- **Objective 1 (Data):** Collect, merge, and clean heterogeneous datasets to build a master dataset covering player performance, financial contracts, sentiment, and health (Day 1–11).
- **Objective 2 (Modeling):** Implement and compare multiple regression and time-series models (RF, XGBoost, LSTM) to predict market values (Day 12–20).
- **Objective 3 (Interpretability):** Apply Feature Importance and SHAP for model interpretability, ensuring actionable insights (Day 12, 32).

Scope: Focused on top-tier football players (Top5 Players and Futbin data) with advanced performance metrics from StatsBomb.

II. Data Acquisition and Preparation

2.1 Data Sources

- **Performance Data:** Futbin, FBref, StatsBomb Open Data (detailed passes, shots, pressure events).
- **Financial Data:** Transfermarkt (Market value, contracts, transfer fees).
- **Sentiment Data:** Twitter API and Reddit scraping to quantify fan buzz and sentiment.
- **Health Data:** Injury history, quantified by total days missed.

2.2 Data Ingestion and ETL Process

- **Combined Dataset Creation (Day 3):** Futbin and Top5 datasets merged, prioritizing ‘similar player’ consistency. (Figure 1: Dataset Schema)
- **Data Cleaning (Day 6–8):** Missing value handling, type standardization, duplicate removal.
- **Versioning (Day 23):** Snapshots of CSVs for reproducibility.

III. Feature Engineering

In this phase, the goal was to transform raw data into meaningful features that capture both player performance trends and contextual factors influencing market value.

Domain-Specific Features

1. **age_experience:**
    - Combines a player’s age with total career appearances to quantify experience relative to age.
    - Helps capture early-career potential versus late-career consistency.
1. **contract_risk:**
    - Derived from the number of months remaining on a player’s contract.
    - Players nearing contract expiration often experience market value fluctuations, making this a key financial driver.
1. **total_days_missed:**

Summarizes injury history by calculating total days a player has missed due to injuries.

Provides a quantitative measure of injury risk, which negatively impacts transfer value.

Sentiment Feature

Raw social media data from Twitter and Reddit were processed using natural language processing techniques to generate numeric sentiment scores.

Positive sentiment trends were observed to correlate moderately with increases in player market value, reflecting fan and media influence.

IV. Exploratory Data Analysis (EDA) and Dimensionality Reduction

4.1 Exploratory Data Analysis (Day 13, 24, 25)

Target Variable: `market value` is highly skewed, necessitating robust models like RF/XGBoost.

Correlation Analysis: Examined KPI relationships (Goals, Assists) with market value; low correlation observed for sentiment, moderate for injury days.

Position-Specific Insights: Forwards tend to maintain higher market value despite more missed games.

4.2 Dimensionality Reduction (Day 14)

PCA: Reduced numeric redundancy while retaining explained variance (Figure 2).

LDA: Class separation by player position, informing feature selection.

V. Modeling Methodology and Results

5.1 Traditional Regression Models

Linear Regression: Baseline model.

Random Forest (RF): Captures non-linear relationships, robust to outliers.

XGBoost: Gradient boosting for superior performance.

Ensemble: Average RF + XGBoost slightly improved accuracy.

Metrics: RMSE, MAE, R^2 (Table 1).

5.2 Time-Series Modeling with LSTM

Data Preparation: 5-match sequential windows, MinMax scaling, One-Hot encoding.

Univariate LSTM: Baseline using past value only.

Multivariate LSTM: Inputs include performance, injury, sentiment (Day 18–20).

Advanced Techniques: Bi-LSTM (Day 29) and Attention Layer (Day 31).

Results: Multivariate LSTM with Attention outperformed traditional models (Table 2, Figure 3).

VI. Model Interpretability and Deployment

6.1 Random Forest Feature Importances

Top features: recent goals, assists, contract risk, injury days (Figure 4).

6.2 SHAP Analysis (Day 32)

Performance positively influences market value; contract risk and injuries negatively.

Feature influence varies by player position (Figure 5).

6.3 Prediction Pipeline

Pipeline: Preprocessing → Feature Selection → Prediction

Robustness: Stable predictions on unseen data.

Final Output: `master_list_cleaned.csv` with top predicted players (Day 35).

VII. Conclusion and Future Work

7.1 Conclusion

Successfully integrated performance, sentiment, and injury data to predict market values.

Multivariate LSTM with Attention captured temporal trends effectively.

Contract risk and recent performance are the most predictive features.

7.2 Future Scope

Deploy as an interactive web dashboard for analysts.

Incorporate opposition strength and match context into XGboost.

Expand sentiment to include news and media releases.

Explore Graph Neural Networks to model inter-player and inter-club relationships.

VIII. References

Futbin, FBref, StatsBomb Open Data, Transfermarkt.

Relevant research papers on machine learning in sports analytics.