Dynamic Player Transfer Value Prediction using AI and Multi-source Data

---

1. Introduction

Football player transfer valuation is a complex and dynamic process influenced by on-field performance, market dynamics, public perception, and player health. Traditional valuation approaches rely heavily on historical prices and expert judgment, often ignoring qualitative and temporal factors.

This project aims to build an AI-driven system that predicts football players' market values by integrating **multi-source data** including performance statistics, market values, injury history, and social media sentiment. The project is executed in multiple milestones, each contributing a critical layer to the final predictive system.

This consolidated report integrates **Milestones 1 to 4** into a single, cohesive project document.

2.: Data Collection and Initial Exploration

2.1 Objectives

- Collect player performance data from open football datasets
- Extract real-world market value data through web scraping
- Gather social media sentiment using NLP techniques
- Collect injury history data
- Perform initial data exploration and quality assessment

2.2 Data Sources

- **StatsBomb Open Data** – Player and match performance statistics
- **Transfermarkt** – Player market value and contract information
- **Twitter API** – Fan and media sentiment data
- **Public Injury Databases** – Injury type, duration, and recurrence

2.3 Initial Data Exploration

- Distribution analysis of market values and performance metrics
- Missing value assessment
- Identification of inconsistencies across datasets

2.4 Key Insights

- Market values follow a right-skewed distribution
- Injury data contained higher missing values
- Social sentiment varied significantly based on player popularity

---

3.: Data Cleaning, Preprocessing, and Feature Engineering

3.1 Objectives

- Clean and standardize multi-source datasets
- Handle missing, inconsistent, and duplicate records
- Engineer meaningful predictive features
- Prepare data for machine learning models

3.2 Data Cleaning

- Missing values handled using statistical imputation
- Duplicate player records removed
- Currency, date formats, and categorical values normalized
- Player identity standardized across datasets

3.3 Feature Engineering

Performance Features

- Rolling averages of goals, assists, and xG
- Recent form indicators
- Match-to-match performance variance

Injury Features

- Injury frequency per season
- Average recovery duration
- Availability ratio

Market and Contract Features

- Remaining contract duration
- Age-to-market-value ratio
- Club-level influence factors

3.4 Sentiment Processing

- Text cleaning and normalization
- Sentiment polarity classification (positive, neutral, negative)
- Aggregation of sentiment at player level

4.: Advanced Feature Engineering and Sentiment Analysis

4.1 Objectives

- Enhance temporal and statistical features

- Quantify injury impact on market value
- Perform advanced sentiment analysis
- Consolidate final feature set for modeling

4.2 Advanced Performance Metrics

- Season-over-season growth rate
- Momentum and stability indices
- Short-term vs long-term form comparison

4.3 Injury Impact Metrics

- Value drop ratio post-injury
- Injury severity index
- Risk-adjusted availability score

4.4 Advanced Sentiment Features

- Sentiment volatility index
- Public perception score
- Sentiment momentum over time
- Engagement-weighted sentiment

4.5 Key Findings

- Positive sentiment correlated with market value stability
- Injury events caused sharp sentiment and value drops
- Sentiment enhanced explanatory power beyond performance data

---

5.: LSTM Model Development for Time-Series Prediction

5.1 Objectives

- Develop time-series models for market value prediction
- Implement univariate and multivariate LSTM models
- Build encoder–decoder LSTM for multi-step forecasting
- Evaluate model performance

5.2 Model Architectures

Univariate LSTM

- Input: Historical market value only
- Captured long-term trends
- Served as baseline model

Multivariate LSTM

- Inputs: Performance, injury, and sentiment features
- Improved responsiveness to real-world events
- Lower prediction error

Encoder–Decoder LSTM

- Multi-step forecasting across future seasons
- Suitable for long-term transfer planning

5.3 Evaluation Metrics

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Training vs validation loss analysis

5.4 Observations

- Multivariate LSTM outperformed univariate model
- Encoder–decoder model effective for long-term trends
- Stable convergence with minimal overfitting

---

6. Overall Deliverables

- ✔️Cleaned and consolidated multi-source datasets
- ✔️Engineered feature-rich dataset
- ✔️Player-level sentiment analytics
- ✔️Trained LSTM-based prediction models
- ✔️Model evaluation and insights

---

7. Conclusion

This project successfully demonstrates how AI and multi-source data can be combined to model the complex dynamics of football player transfer valuation. By progressively advancing from data collection to deep learning–based forecasting, the system captures technical performance, injury risk, market conditions, and public perception in a unified framework.

The consolidated milestones collectively establish a strong foundation for deploying a real-world, data-driven player valuation system with potential applications in scouting, club management, and sports analytics.

---

8. Future Scope

- Integration with Spring Boot backend APIs
- Real-time data ingestion
- Deployment as a web-based analytics platform
- Model explainability and interpretability enhancements