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

1. **Introduction**

Player transfer markets are a vital aspect of professional sports, influencing team dynamics and competitive success. Estimating a player's transfer value is a complex task that depends on various factors such as performance, age, team, public sentiment, and market trends. In this project we propose the development of an AI driven model that predicts player transfer values by integrating multi-source data including player performance statistics, social media sentiment, injury history and historical market data. By using advanced machine learning and time-series forecasting techniques, this project will provide a dynamic, data-driven approach to predicting player transfer values.

2. **Methodology**

The project will be divided into several stages, each involving the application of different AI techniques to predict transfer values. The focus will be on integrating multi-source data and leveraging time-series forecasting methods to model transfer value trends.

2.1.Data Collection and Preprocessing

Data Sources:

Player Performance Data: StatsBomb Open Data

Market Value Data: Transfermarkt datascrapping

Social Media Sentiment: Twitter API for sentiment analysis of player mentions.

Injury Data: Historical player injury records

Data Cleaning and Feature Engineering:

Handling missing values, scaling numerical data, and performing one-hot encoding for categorical variables.

Feature engineering for performance trends, injury risk, and contract details (eg:- remaining contract duration).

Sentiment analysis using Natural Language Processing (NLP) tools like VADER to quantify public perception of players.

2.2.Model Development

Time-Series Data Modelling:

Use Long Short-Term Memory (LSTM) network to capture temporal dependencies in player performance and market trends over time.

Multivariate Time-Series Models:

Player Performance Model: Incorporate performance statistics, injury history and sentiment data into an LSTM model to predict player value.

Merket Sentiment Analysis: Use public sentiment data as an additional feature to refine value prediction based on media perception.

Ensemble methods:

Use XGBoost or LightGBM to create ensemble models that combune player performance data, market trends and social sentiment to generate final transfer value prediction.

2.3.Time-Series Forecasting with LSTMs

Univariate LSTM:

Train a basic LSTM model on historical player performance data to predict future transfer values.

Multivariate LSTM:

Expand the model to include additional feature such as injury history and sentiment analysis for more accurate predictions.

Encoder-decoder LSTM:

Implement an encoder-decoder structure for muti-step forecasting, predicting player values over the next several transfer windows.

2.4.Sentiment Analysis

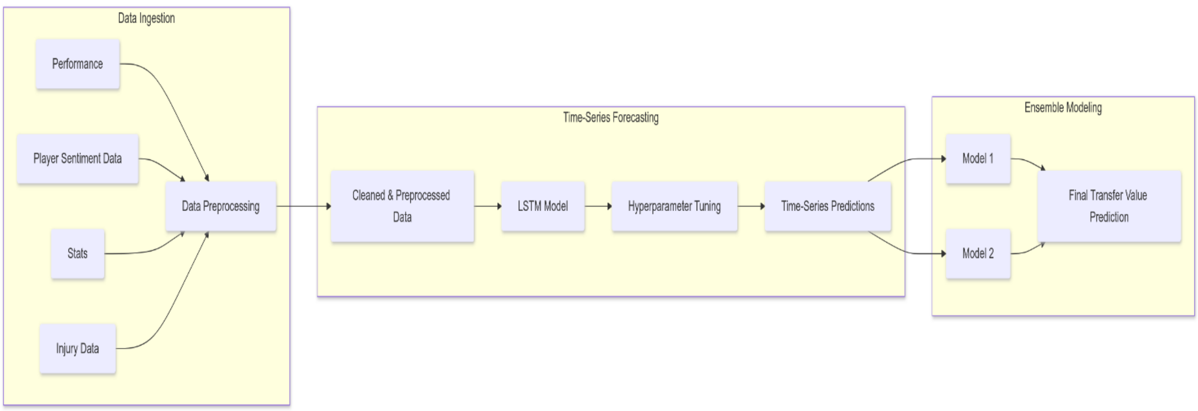
NLP Techniques:

Perform sentiment analysis on social media posts using the VADER or TextBlob libraries.

Create features that quantify player popularity and sentiment impact on transfer value.

3. Architecture Diagram

An architecture diagram illustrating the model components:



4. Expected Deliverables

A dynamic prediction model capable of estimating player transfer values based in multi-source data.

Trained LSTM and XGBoost models for player value prediction and market trend analysis.

Evaluation reports comparing the performance of different machine learning models.

Sentimental analysis reports indicating the impact of public perception on player market value.

Interactive visualization of player performance trends and predicted transfer values.

Comprehensive documentation detailing methodology, implementation and findings.

5. Week-wise module implementation and high-level requirements with output screenshots

Milestone 1: Week 1

Data Collection and Initial Exploration

- **Tasks**:
  - Collect player performance data from StatsBomb Open Data.
  - Scrape market value data from Transfermarkt using web scraping techniques.
  - Fetch sentiment data from social media using the Twitter API and perform sentiment analysis with NLP tools (VADER/TextBlob).
  - Gather injury history data from relevant sources.
- **Deliverables**:
  - Raw datasets from all sources.
  - Initial data exploration (distributions, missing data).
  - Exploration report on the structure and content of the datasets.

Milestone 2: Week 2

Data Cleaning and Preprocessing

- **Tasks**:
  - Clean the datasets (handle missing values, duplicate data).
  - Perform feature engineering:
    - Create performance trend features, injury risk metrics, and contract-related features (e.g., contract duration).
  - Process and scale numerical data; one-hot encode categorical variables.
  - Begin sentiment analysis by processing social media data using NLP.
- **Deliverables**:
  - Cleaned and pre-processed datasets.
  - Feature-engineered datasets with new metrics.

    - Preliminary sentiment analysis report.

Milestone 3: Weeks 3-4

Advanced Feature Engineering and Sentiment Analysis

- **Tasks**:
    - Refine feature engineering with advanced metrics such as performance trends over time and injury impact on value.
    - Perform sentiment analysis using NLP (VADER/TextBlob) on social media data, creating sentiment scores for players.
    - Generate features from social sentiment that impact market value.
- **Deliverables**:
    - Final feature set for modeling (performance metrics, sentiment scores, injury data).
    - Sentiment analysis report with insights on public perception's effect on player value.

Milestone 4: Week 5

LSTM Model Development for Time-Series Prediction

- **Tasks**:
    - Develop a univariate LSTM model for predicting player transfer values based on historical performance data.
    - Expand to multivariate LSTM, incorporating additional features like injury history and sentiment analysis.
    - Implement an encoder-decoder LSTM for multi-step forecasting to predict future player values over multiple transfer windows.
- **Deliverables**:
    - Trained univariate and multivariate LSTM models.
    - Initial prediction results using LSTM.
    - Model performance evaluation (loss curves, accuracy metrics).

Milestone 5: Week 6

Development of Ensemble Models and Integration

- **Tasks**:
    - Implement ensemble models using XGBoost or LightGBM.
    - Integrate the results of the LSTM models with XGBoost to create an ensemble model combining performance data, market trends, and social sentiment.
    - Test model performance on validation datasets.
- **Deliverables**:
    - Trained XGBoost models and integrated ensemble models.
    - Ensemble model performance evaluation report.

Milestone 6: Week 7

Model Evaluation, Hyperparameter Tuning, and Testing

- **Tasks**:

    - Evaluate model performance using metrics such as RMSE, MAE, and R-squared.
    - Conduct hyperparameter tuning for both LSTM and ensemble models (grid search or random search).
    - Test models on validation datasets and assess generalization.
- **Deliverables**:
    - Final tuned models with optimized parameters.
    - Model evaluation report comparing the performance of different models (LSTM, XGBoost, and ensemble).

Milestone 7: Week 8

Final Model Evaluation, Visualization, and Reporting

- **Tasks**:
    - Conduct final model evaluation and generate transfer value predictions.
    - Create interactive visualizations of player performance trends, market values, and prediction results.
    - Prepare comprehensive documentation detailing the methodology, model implementation, and findings.
    - Finalize project presentation and report.
- **Deliverables**:
    - Final model predictions and visualizations.
    - Complete project documentation.
    - Presentation deck summarizing the project.

6. Evaluation Criteria

Milestone 1: Week 1

- **Data collection**: Successful acquisition of player performance, market value, sentiment, and injury data.
- **Initial exploration**: Summary of data characteristics and early insights.
- **EDA with visualizations**: Key trends and distributions visualized

Milestone 2: Week 2

- **Preprocessing**: Handling missing values and data anomalies.
- **Feature engineering**: Creation of features like performance trends, injury risks, and sentiment scores.
- **Sentiment analysis**: Initial results from NLP sentiment analysis.

Milestone 3: Weeks 3-4

- **Advanced features**: Comprehensive set of engineered features (contract details, performance trends).
- **Sentiment integration**: Sentiment scores successfully integrated with other features.
- **Finalized dataset**: Ready for model training.

Milestone 4: Week 5

- **Univariate and multivariate LSTM**: Successful training and initial results of time-series models.
- **Model evaluation**: Early performance results (loss curves, RMSE, etc.).
- **Multi-step forecasting**: Predictions for future transfer windows.

Milestone 5: Week 6

- **Ensemble models**: Successful integration of LSTM results with XGBoost or LightGBM.
- **Validation**: Performance comparison between models on validation datasets.
- **Improvement**: Enhanced model results through ensemble techniques.

Milestone 6: Week 7

- **Hyperparameter tuning**: Optimization of LSTM and ensemble models.
- **Final model selection**: Best model chosen based on RMSE, MAE, or other metrics.
- **Validation results**: Testing on holdout datasets.

Milestone 7: Week 8

- **Model deployment**: Successful deployment of the final model.
- **Visualization**: Interactive plots showing predictions and trends.
- **Final documentation**: Complete report on methodology, results, and findings.

7. Conclusion

The proposed project aims to develop a robust AI-driven model for predicting player transfer values using a combination of time-series forecasting, sentiment analysis and muti-source data. By leveraging advanced machine learning techniques and multi-variate models this project seeks to provide a dynamic and accurate system for forecasting player market values. The finding will contribute to the growing field of sports analytics and player valuation methodologies.