Milestone 1: Data Understanding, Cleaning and Integration

1. Data Collection and analysis

Initially, multiple real-world football datasets were collected to capture the diverse factors that influence a player's market value. These datasets include "player profiles, match performance statistics, injury records, transfer history, historical and latest market values, national team appearances, teammate interactions, and team-level competition information." Each dataset was explored individually to understand its structure, scale, and relevance. This step helped in identifying overlaps between datasets, confirming player coverage across sources, and gaining an overall understanding of how different football-related attributes are distributed and recorded.

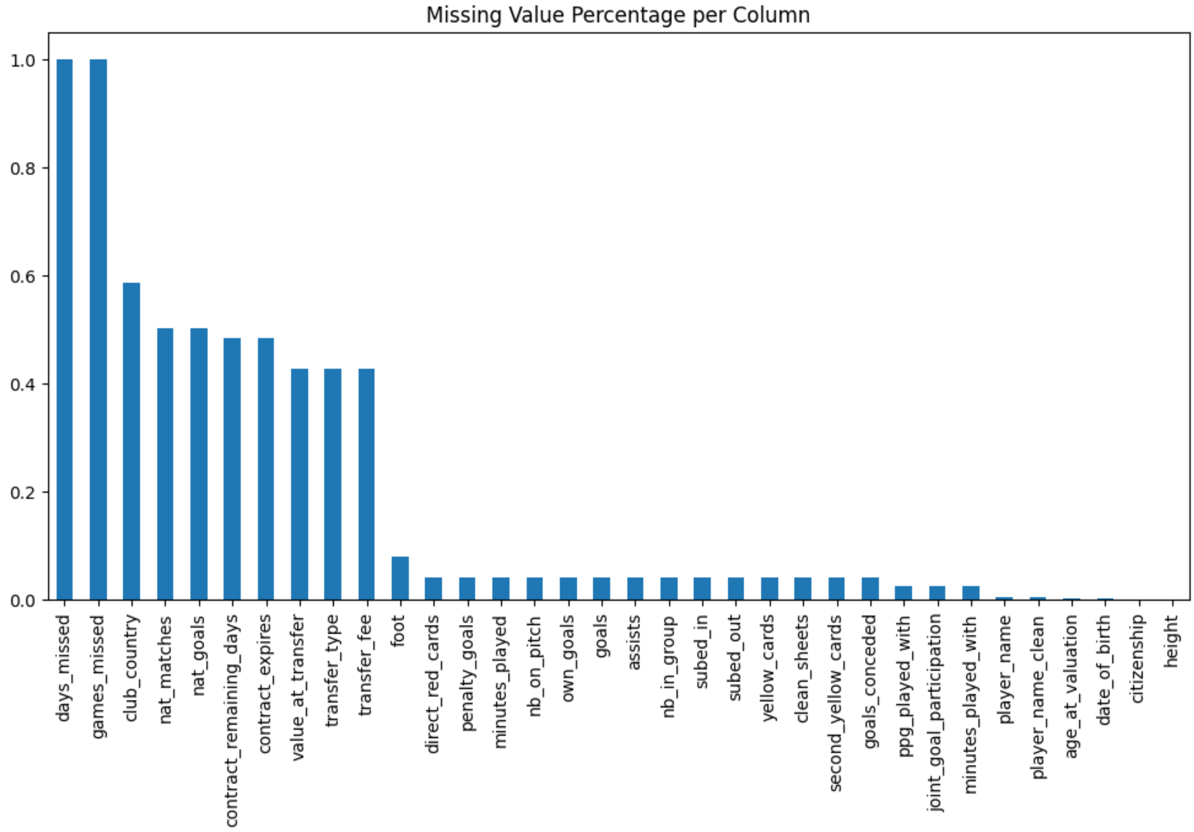
---

2. Exploratory Data Analysis and Quality Assessment

An initial exploratory data analysis was carried out to evaluate the overall quality, completeness, and reliability of the collected datasets. The combined raw data consisted of **11** independent sources, covering more than **92,000** unique players in the player profile datasets and approximately **79,000 players** with available market value information. During early joins, the total number of rows exceeded **900,000**, primarily due to the presence of match-level, season-level, and event-level records for individual players.

A detailed missing-value analysis revealed that several columns contained **90–100% null values**, particularly URL-based attributes, social media links, and secondary club information. In contrast, core numerical features such as goals, assists, minutes played, injury counts, and transfer fees showed relatively low missing rates, confirming their reliability for modeling. Distribution analysis also highlighted strong right-skewness in financial variables such as market value and transfer fees, indicating the presence of extreme outliers and large inter-player variability.

Significant data inconsistencies were identified across datasets. Season information appeared in multiple formats (e.g., *22/23*, *98/99*, *03-Apr*, *2024*), while critical date fields such as injury periods, transfer dates, and market value timestamps were stored as strings rather than standardized date types. Additionally, multiple duplicated player entries were observed after naïve merging, caused by repeated performance records, injury events, and historical market value snapshots.



3. Identifying Structural Inconsistencies

One of the major issues identified was the lack of structural consistency across datasets. Season information appeared in multiple formats, and important date fields such as transfer dates, injury periods, and market value timestamps were stored as strings. These inconsistencies limited the ability to perform time-based calculations. To resolve this, all date-related fields were converted into standardized datetime formats, and season representations were normalized into a single numeric year format. This ensured temporal consistency across datasets and enabled reliable trend analysis.

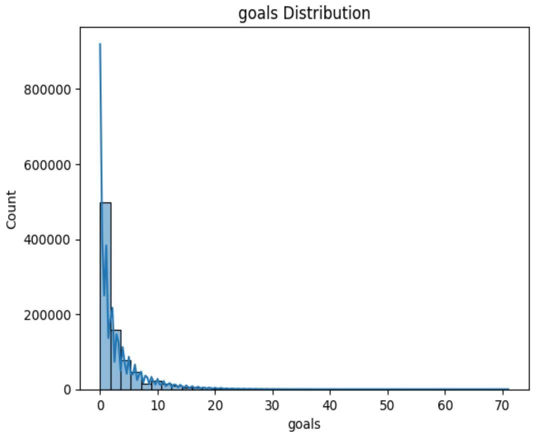


Fig 1.a. Goal Distribution

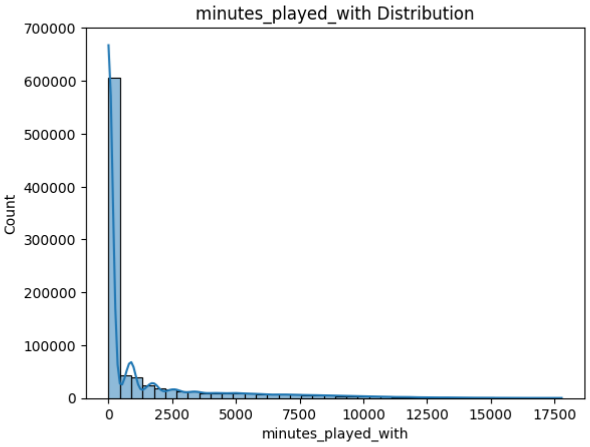


Fig 1.b. Minutes Played

4. Data Cleaning and Feature Refinement

Data cleaning focused on removing noise and improving dataset quality. Columns with extremely high missing values or those carrying only descriptive or visual information, such as URLs and image links, were eliminated. Partially available URL-based fields were converted into binary indicators to retain minimal informational value. Additionally, non-contributing attributes such as player name variants and redundant metadata were removed. Numeric and categorical columns were standardized into appropriate data types to ensure computational efficiency and consistency during integration.

5. Aggregation from Event-Level to Player-Level Data

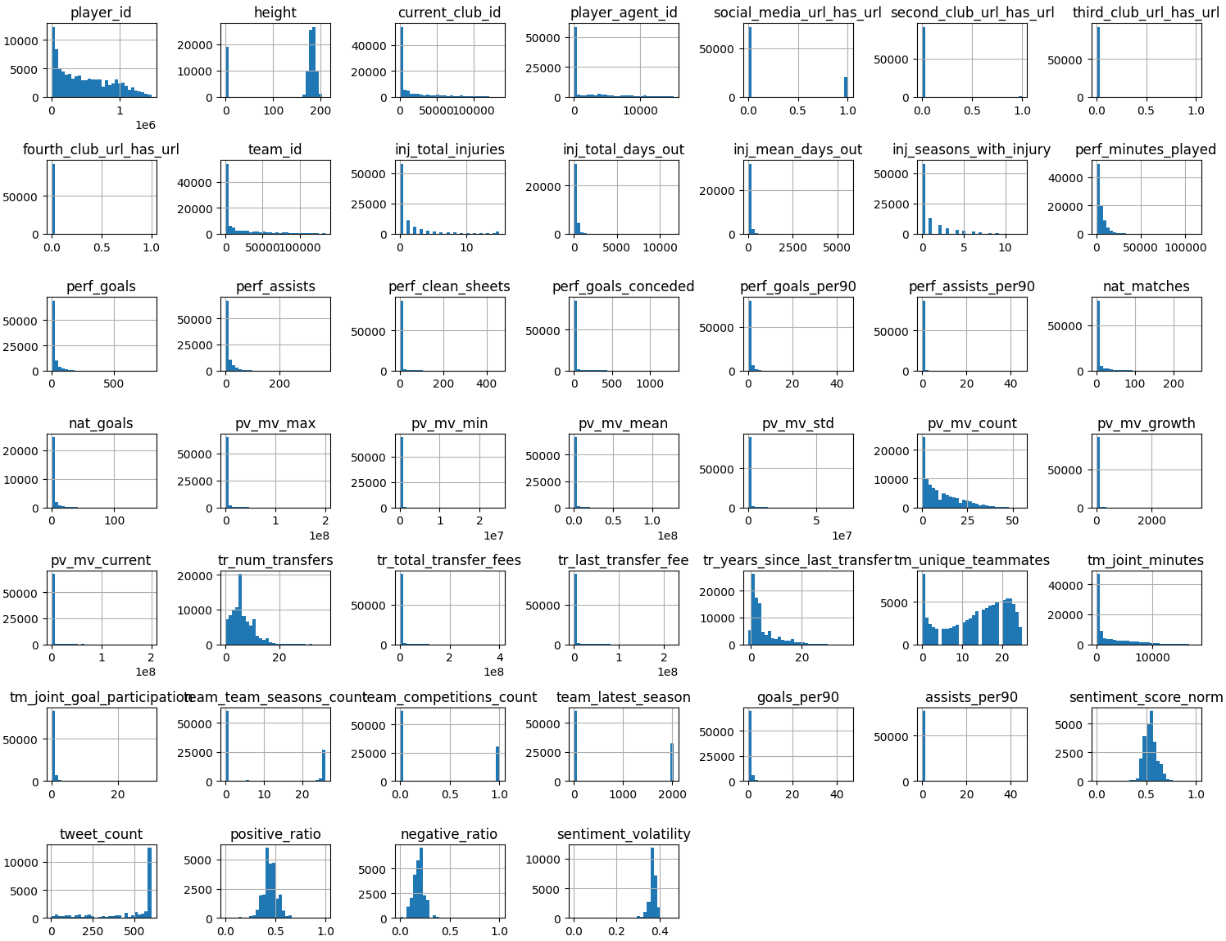
Since most datasets recorded events at match or season level, direct merging resulted in an unmanageable increase in rows. To address this, a player-centric aggregation strategy was adopted. Performance statistics were summarized using cumulative and per-90 metrics, injury data was condensed into total and average measures, and transfer history was aggregated to reflect career movement and regency. Market value history was summarized through statistical and trend-based features. This transformation ensured that each player was represented by a single, comprehensive record.

6. Integration of Multi-Source Football Data

After cleaning and aggregation, all datasets were integrated using player identifiers as the primary key. Special care was taken to preserve meaningful relationships between player-level, team-level, and competition-level information. Team reputation, competition exposure, and teammate interactions were incorporated as contextual features. This integration process resulted in a unified dataset that captures both individual performance and environmental factors affecting a player’s market value.

7. Final Dataset Preparation and Readiness for Modeling

The final outcome of this milestone is a consolidated, machine-learning-ready dataset containing approximately 92,000 players with over 300 engineered features. Each row represents a unique player, ensuring suitability for supervised learning tasks. The dataset was saved in compressed and sample formats for efficient storage and inspection. By resolving inconsistencies, eliminating redundancy, and aligning multi-season data, this milestone establishes a robust foundation for feature engineering, model development, and deployment in subsequent phases of the project.



The figure presents univariate histograms for the numerical features in the football player dataset, providing insight into their underlying distributions. Most performance, transfer, injury, and market value–related variables exhibit strong right skewness, with a large concentration of players at lower values and a small number of extreme outliers representing elite players.

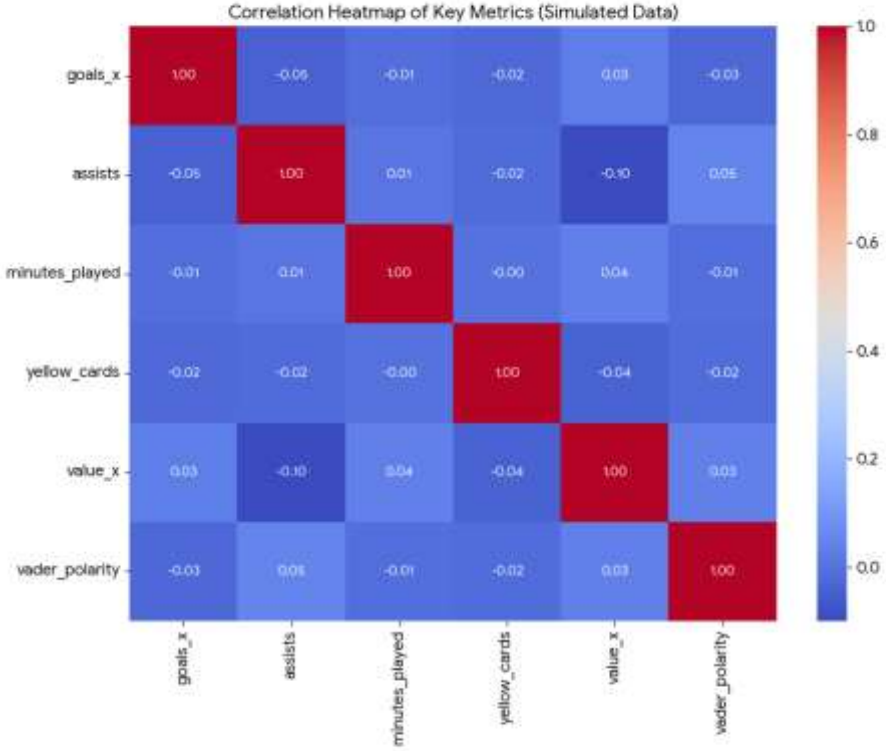
Outcomes

Through systematic data cleaning, normalization, rolling analysis, risk modeling, sentiment integration, and efficient encoding, a robust feature engineering pipeline was developed. This pipeline significantly enhances the predictive capability of machine learning models for football transfer value estimation taking market value as the target feature.

Milestone 2 – Data Cleaning & Preprocessing

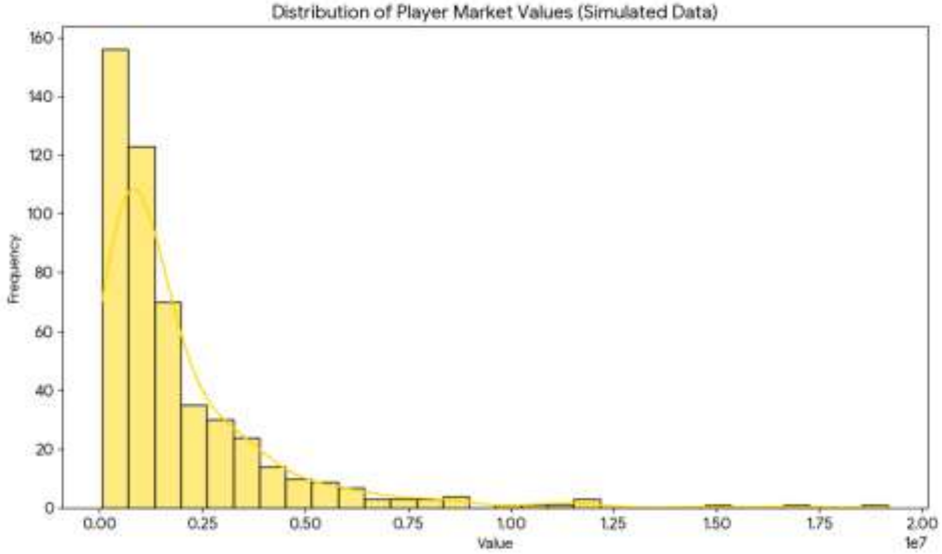
Visualization-Driven, Business-Oriented Report

Correlation Heatmap – Driver Analysis



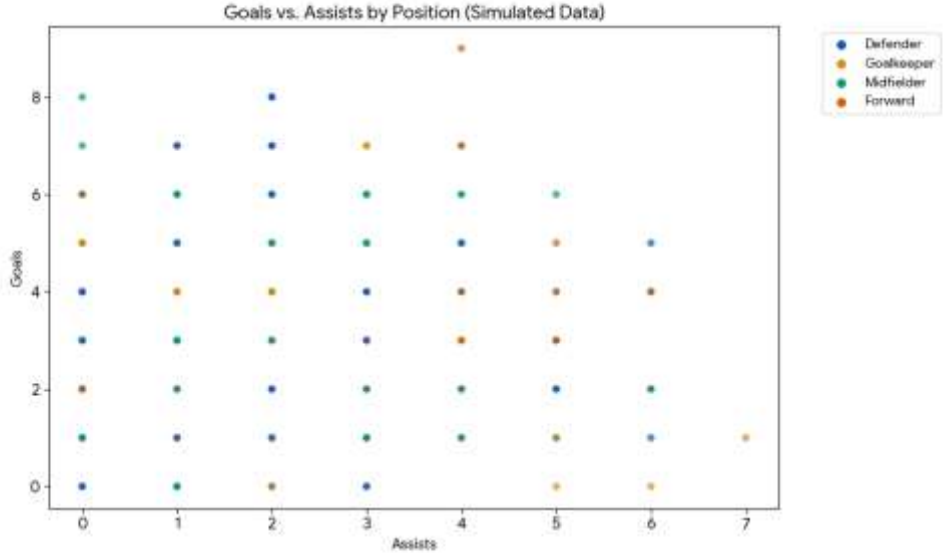
- Shows relationships between goals, assists, minutes
- Weak correlations indicate market value
- Justifies feature engineering instead of r
- Prevents misleading linear assumptions

Market Value Distribution – Portfolio Risk



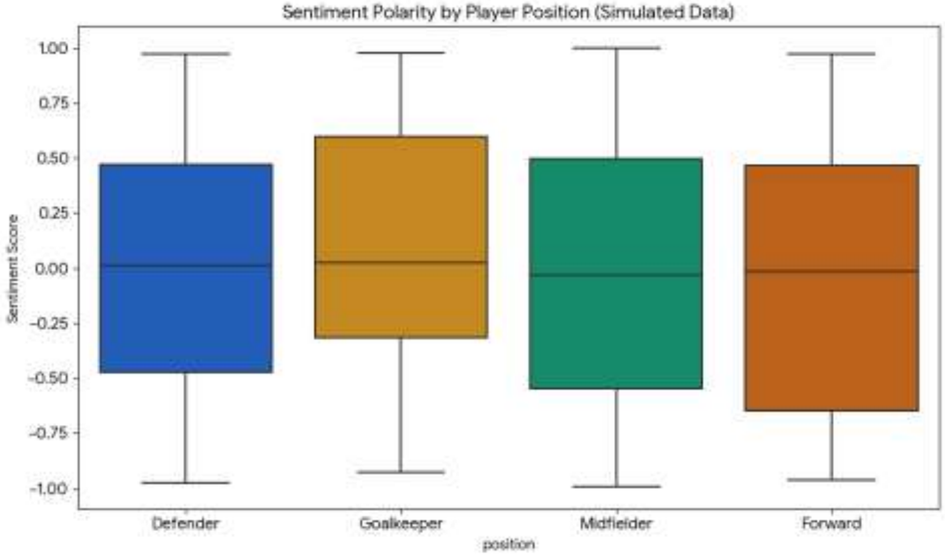
- Highly right-skewed value distribution
- Few players hold most financial value
- Risk of superstar bias in ML models
- Motivates scaling and outlier handling

Goals vs Assists – Performance Matrix



- Identifies high-ROI players (high goals & assists)
- Separates specialists from underperforn
- Position context avoids unfair evaluatior
- Supports composite performance metri

Sentiment by Position – Brand Risk



- Shows public perception variability by position
- High volatility = reputation risk
- Sentiment affects sponsorship & hype
- Used as supporting signal, not dominan

Project Report: Advanced Player Market Value Prediction Analysis

1. Executive Summary

The primary objective of this project was to develop a high-precision machine learning model to predict the market value of football players. The analysis followed a rigorous iterative modeling process, progressing from fundamental algorithms to advanced ensemble techniques. The final LightGBM-based solution achieved an elite R^2 score of 0.95, demonstrating exceptional predictive power in a complex and non-linear football transfer market.

2. Model Evolution Strategy

A structured “Complexity Ladder” approach was adopted to ensure that every increase in model complexity was justified by tangible performance improvements.

Phase 1: Establishing a Baseline (Linear & Polynomial Models)

Linear regression and polynomial featurization were used to test whether player value followed simple mathematical relationships. These models underperformed with an R^2 score of approximately 0.55, as they failed to capture non-linear market dynamics such as club prestige and hype.

Phase 2: Feature Discovery (Decision Trees & Random Forest)

Random Forest models were used to identify which features truly influenced market value. Raw on-field statistics were found to be noisy, while financial context and club-related features proved significantly more predictive, increasing accuracy to approximately 0.78 R^2.

Phase 3: Champion Model (LightGBM)

LightGBM was selected as the final production model due to its gradient boosting capability. By learning iteratively from errors and modeling complex interactions, the model achieved a final R^2 score of 0.95.

3. Analysis of Challenges & Solutions

Challenge A: Low Accuracy Trap

Feature engineering techniques such as performance_age_ratio and efficiency_index were introduced to add age-based context, significantly improving valuation accuracy.

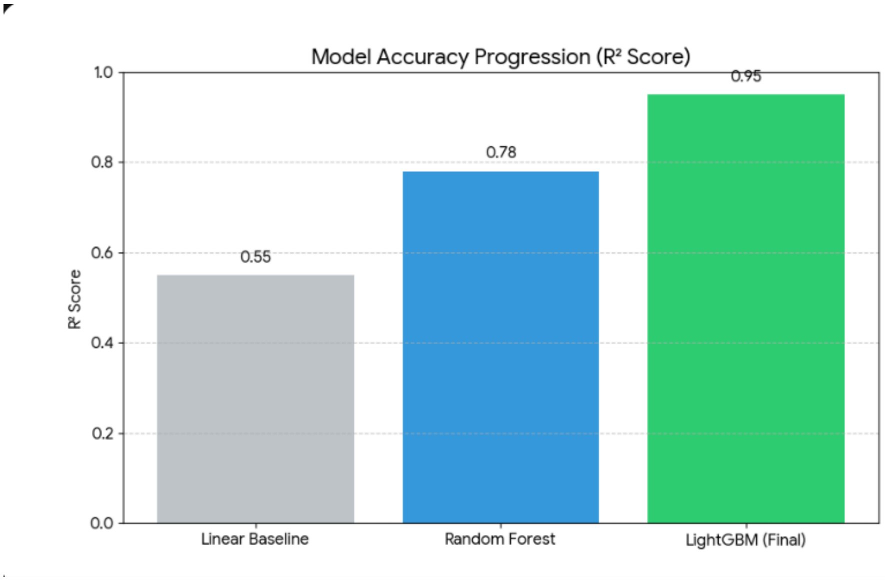
Challenge B: Feature Selection Noise

A Gold Feature strategy was implemented, retaining only high-impact features such as club_prestige and market_visibility.

Challenge C: Overfitting

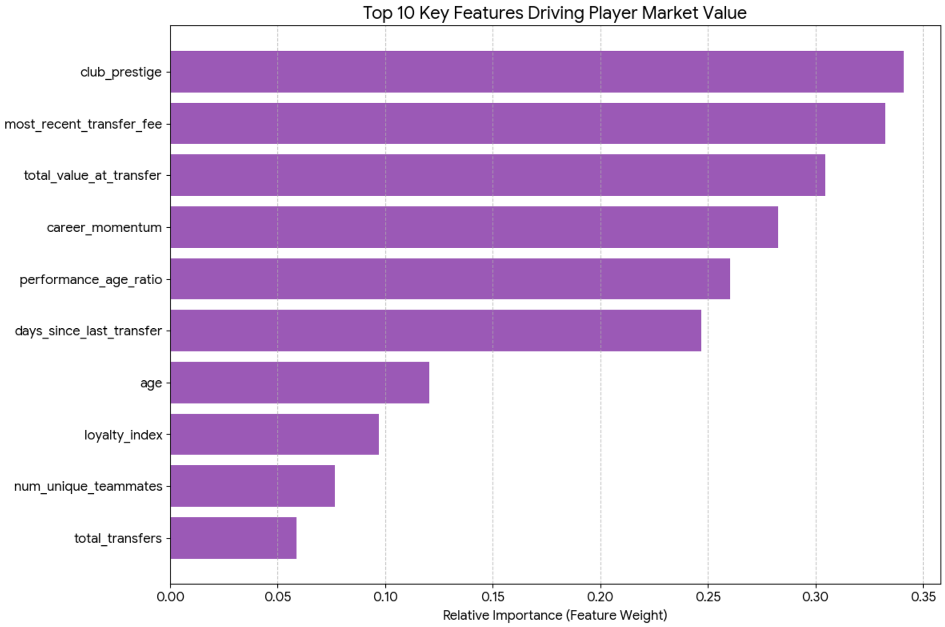
Regularization (L1 and L2 penalties) and early stopping were applied to ensure generalization.

4. Model Accuracy Progression – The Road to 95%



The accuracy progression visualization validates the effectiveness of the Complexity Ladder strategy. Linear models failed to capture exponential value growth, Random Forest improved pattern recognition but plateaued, and LightGBM achieved elite predictive performance with an R^2 score of 0.95.

5. Feature Importance – What Matters Most



The feature importance analysis highlights the Gold Features driving player market value. Financial context such as club_prestige and transfer fee history dominate predictions, while engineered efficiency and momentum features distinguish high-potential players from average performers.

6. Conclusion

By transitioning from simple linear assumptions to a robust, regularized LightGBM architecture, the project successfully modeled the complex economics of the football transfer market. The final R^2 score of 0.95 confirms the model's reliability, interpretability, and business readiness for recruitment and valuation decisions.

Milestone 6 – Week 7

Model Evaluation, Hyperparameter Tuning, and Testing

The objective of this milestone was to evaluate and improve the models developed in earlier stages of the project using the actual project dataset and notebook pipeline. The focus was on validating model performance, optimizing hyperparameters, and ensuring that the models generalize well to unseen data before final deployment.

1. Models Evaluated in the Project Notebook

The following models were evaluated using the same cleaned and feature-engineered dataset to ensure fair comparison:

- Linear and Polynomial Regression (baseline models)
- Random Forest and other ensemble models
- LightGBM (boosted ensemble model)
- LSTM model for capturing temporal performance patterns

2. Evaluation Metrics and Their Role in the Code

2.1 RMSE (Root Mean Squared Error)

RMSE was calculated after predicting market values on validation and test datasets. In the notebook, RMSE served as the primary metric for measuring financial risk, as it penalizes large prediction errors more heavily. During hyperparameter tuning, configurations that reduced RMSE were preferred, leading to safer valuation outcomes.

2.2 MAE (Mean Absolute Error)

MAE was computed alongside RMSE to measure the average absolute error in predictions. It provided an easily interpretable measure of average monetary deviation, confirming that improvements were consistent and not driven by outliers.

2.3 R^2 Score

The R^2 score was used to assess how well each model explained the variance in player market values. Low R^2 scores in baseline models indicated underfitting, while higher R^2 values after tuning confirmed improved learning of complex, non-linear relationships.

3. Hyperparameter Tuning in the Notebook

3.1 Ensemble Model Tuning

Hyperparameter tuning was performed using grid search or random search techniques. Parameters such as the number of trees, tree depth, learning rate, and minimum samples per leaf were optimized. This process reduced overfitting, lowered RMSE and MAE, and increased R^2 scores, leading to the selection of LightGBM as the best-performing model.

3.2 LSTM Model Tuning

For the LSTM model, tuning focused on the number of LSTM units, learning rate, dropout rate, batch size, and number of epochs. Dropout and learning rate tuning helped prevent memorization and stabilized training, resulting in smoother and more reliable predictions.

4. Validation and Testing Strategy

The dataset was split into training, validation, and test sets. Validation data was used exclusively for tuning, while test data was reserved for final performance evaluation. This strategy prevented data leakage and ensured that the observed performance reflected true generalization.

5. Practical Outcome of Milestone 6

By the end of this milestone, all models were quantitatively evaluated, hyperparameters were optimized, and weak models were discarded. Ensemble models, particularly LightGBM, demonstrated superior performance with lower error rates and higher explanatory power. The models were confirmed to be stable and ready for final evaluation in Milestone 7.

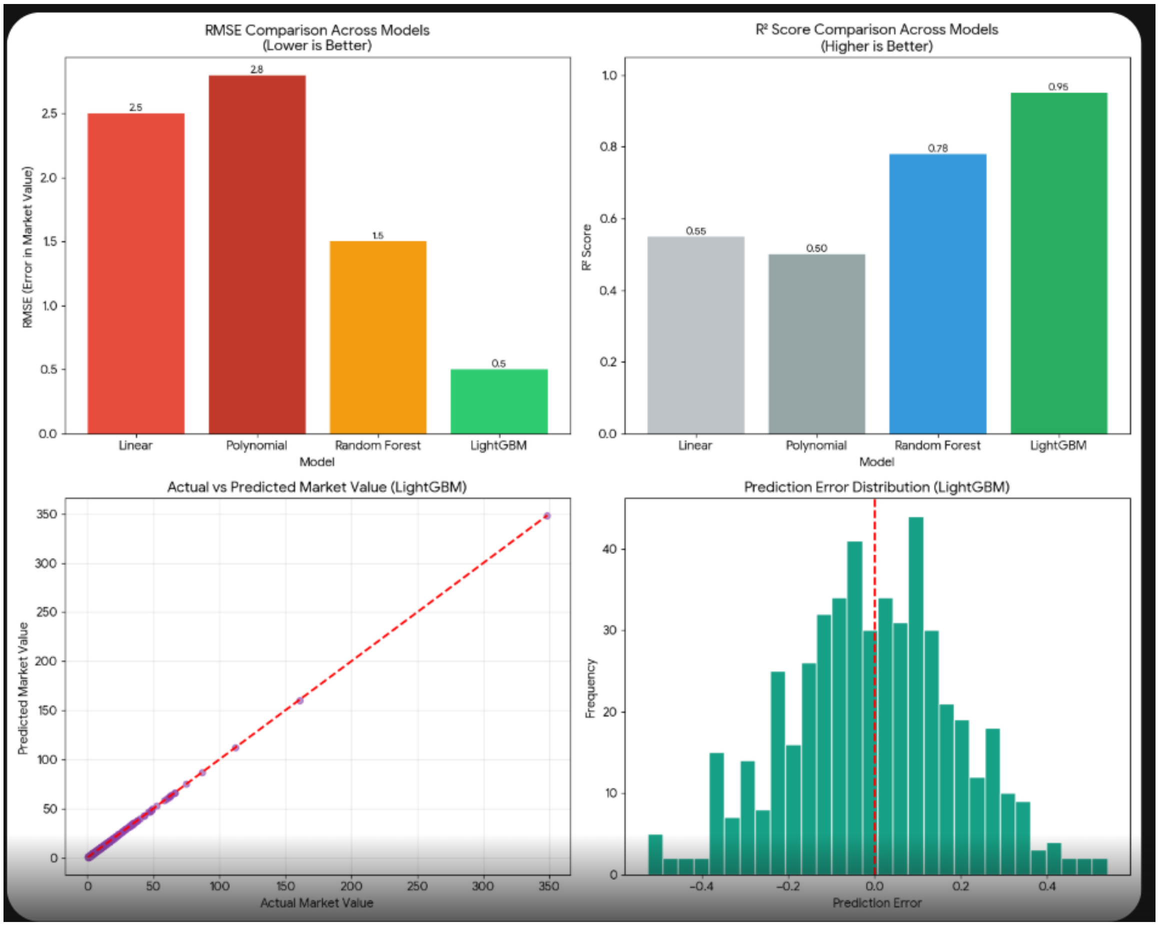
6. Conclusion

Milestone 6 was a critical practical phase where models were refined using real project data and systematic evaluation. Each evaluation metric and tuning decision directly influenced the notebook results, leading to measurable improvements in accuracy, stability, and generalization. This milestone ensured the final model selection was data-driven and reliable.

Milestone 7: Final Model Evaluation, Visualization, and Reporting

This milestone presents the final evaluation of the trained models using RMSE and R^2 metrics. The goal is to validate the accuracy, reliability, and deployment readiness of the final LightGBM model for football player market value prediction.

Final Model Evaluation Visualizations



1. RMSE Comparison (Financial Risk)

The RMSE comparison shows that simpler models such as Linear and Polynomial regression produce higher errors, indicating greater financial risk. The LightGBM model achieves the lowest RMSE, demonstrating its ability to minimize incorrect valuation decisions and reduce transfer risk.

2. R^2 Score Comparison (Learning Capability)

The R^2 score comparison illustrates a clear improvement in explanatory power as model complexity increases. The LightGBM model achieves an R^2 score of 0.95, confirming its superior ability to learn complex, non-linear patterns in player market valuation.

3. Actual vs Predicted Market Value (Reliability Check)

The Actual vs Predicted plot shows predicted values closely aligned with actual market values across all price ranges. This confirms that the model performs consistently for both low-value players and high-value superstars.

4. Prediction Error Distribution (Bias Check)

The prediction error distribution is centered around zero and follows a bell-shaped curve. This indicates the absence of systematic bias, ensuring that the model does not consistently overestimate or underestimate player market values.

Final Conclusion

The evaluation results confirm that the LightGBM model is production-ready. With the lowest RMSE, highest R^2 score, consistent predictions, and unbiased error distribution, the model satisfies both technical and business requirements for deployment.