DSC680 Project 2: Final Paper

Soccer Transfer Market Player Value Prediction in Soccer Leagues

Pragathi Porawakara Arachchige

Bellevue University

DSC680-T301 Applied Data Science (2257-1)

Dr. Matthew Metzger

07/16/2025

Business Problem

Given billions of dollars invested yearly, professional football teams come under great pressure regarding player transfers. A financial and strategic choice that can affect the performance and budget of a team is knowing a player's actual market value. But today's valuation approaches usually rely on media impact, subjective opinions, and inadequate data. By creating an exact, data driven model, clubs may lower financial risk, strengthen negotiating leverage, and identify undervalued talent more effectively. This project solves the business issue: can machine learning predict a soccer player's market value using performance and demographic data?

Background / History

The global multibillion-dollar industry of soccer transfers has developed such that player prices are shaped by both on-field performance and off-field brand power. Although player valuation is still mostly subjective, analytics has become popular in sports over the past ten years. Although sites like Transfermarkt offer crowdsourced estimates, they lack openness and might be prejudiced. It is now feasible to construct strong machine learning models to quantify player value depending on measurable criteria thanks to increased availability of structured data through APIs and publicly available sources such as FBref and FIFA databases. The best opportunity to investigate predictive modeling in this high-stakes setting is the growing focus on data-driven decision-making in sports.

Data Explanation

Accessed via Kaggle, the main merged collection from Transfermarkt and FBref used in this project covers 2017 to 2020 and has comprehensive records for more than 7,000 European football players. Almost 400 characteristics per player, the dataset reflects a broad spectrum of features including appearances, goals, assists, minutes played, age, nationality, position, and estimated market value.

Data preparation was an important first phase that included normalizing numerical values, eliminating duplicates, and cleaning missing data. Onehot encoding was used on foot preference and country among other categorical variables. Furthermore applied to handle extreme values especially among elite players were outlier detection methods. Continuous numerical characteristics had standardization methods applied to guarantee model-wide comparability. To help contextualize player ratings including categories like Attacking, Mentality, and Goalkeeping skills, an extra FIFA 21 auxiliary dataset was cited.

Methods

To find which model would most accurately forecast market value, a varied collection of machine learning techniques was evaluated. These included linear regression, decision trees, random forests, K-nearest neighbors (KNN), support vector regression (SVR), gradient boosting, and Gaussian process regression. Standard regression performance indicators including the coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE) were used to assess each model.

The best performer was the Gradient Boosting Regressor, and RandomizedSearchCV helped to further fine-tune it. Optimized hyperparameters included subsample ratio, maximum tree depth, learning rate, and number of estimators. Model assumptions were validated and prediction accuracy assessed across various value ranges using visual tools such scatter plots and residual plots.

Analysis

Model evaluation showed that with default settings Gradient Boosting Regressor had the greatest predictive power R^2 of 0.75 and increased to 0.789 after hyperparameter adjustment. The MAE was roughly 5.19 million and the corresponding RMSE averaged 9.05 million. Considering the complexity and fluctuation of the soccer industry, these numbers indicate a great level of accuracy.

Though they were somewhat decent, alternative models like decision trees and random forests were finally surpassed by Gradient Boosting. Even after scaling, linear regression models battled with multicollinearity problems and outlier sensitivity. Owing to the huge dataset and nonlinearity in the relationships between characteristics, SVR and Gaussian Process Regression underperformed. Analysis of feature importance revealed that predicted values depended much on factors including player position, age, goals scored, minutes played, and appearances. Patterns and outliers in the data were highlighted by means of visualizations including scatter plots, bar charts, and violin plots.

Conclusion

With a tuned R^2 of 0.789, the Gradient Boosting Regressor showed excellent projection capability for forecasting player value based on accessible data. This model aids clubs in talent acquisition, appraisal, and investment planning by quantifying subjective judgments. It improves decision-making with data driven insight, even if it doesn't completely replace human intuition. The results confirm the application of machine learning in sports economics and offer a replicable model for predictive modeling in other sports fields.

Assumptions

This project holds that past player performance reflects their current market value. It also assumes that the major elements affecting player value are reflected in the available data, barring aspects like media perception or player endorsements. The model also makes the assumption, based on historical data, that market circumstances such as transfer policies and inflation remain fairly constant throughout the forecasting horizon.

Limitations

This initiative has several constraints. First, the model ignores unexpected market changes resulting from contracts changes, injuries, or political considerations influencing transfers. Second, it not only lacks real time data integration which could improve prediction accuracy. Third, there is no consideration of attributes like leadership, attitude, or locker room impact significant but difficult to measure traits. Finally, since public databases are used, some figures could be predicted instead of validated.

Challenges

One of the most demanding obstacles was data quality. Outliers especially for high-value players like Lionel Messi or Cristiano Ronaldo skewed model performance and needed unusual handling. Furthermore, choosing the most pertinent characteristics from more than 400 available requires careful evaluation and subject matter expertise. Ensuring model generalization was complicated by imbalanced representation throughout leagues and positions. Particularly with ensemble models like Gradient Boosting, model interpretability proved difficult as well.

Future Uses / Applications

Agents, scouts, or clubs' dynamic real-time valuation tools can be modeled by adapting this one. It could also be extended to incorporate injury records, psychological assessments, or social sentiment data from media sources. Another possible use is on fantasy football sites to fairly price players depending on current performance. The model could act as a recommendation engine for transfers or scouting decisions with appropriate integration and retraining.

Recommendations

Given its great performance and adaptability, Gradient Boosting should be used as the primary algorithm for predictive work in this field. Non-Technical stakeholders should have access to a friendly dashboard or API. Maintaining the model's accuracy also depends on retraining it with data from every new season. Clarifying the model’s rationale helps to build stakeholder trust by using explainability technologies like SHAP or LIME.

Implementation Plan

Phase one through four make up the implementation plan. Phase 1 saw the establishment of a baseline model and cleaning of historical data, Phase 2 saw model tuning and performance validation, which has been finished. The emphasis of Phase 3 will be either developing the model as an API or creating an interactive dashboard. At last, in Phase 4, the model will be incorporated into a soccer analysis pipeline to help with simulation of several transfer possibilities.

Ethical Assessment

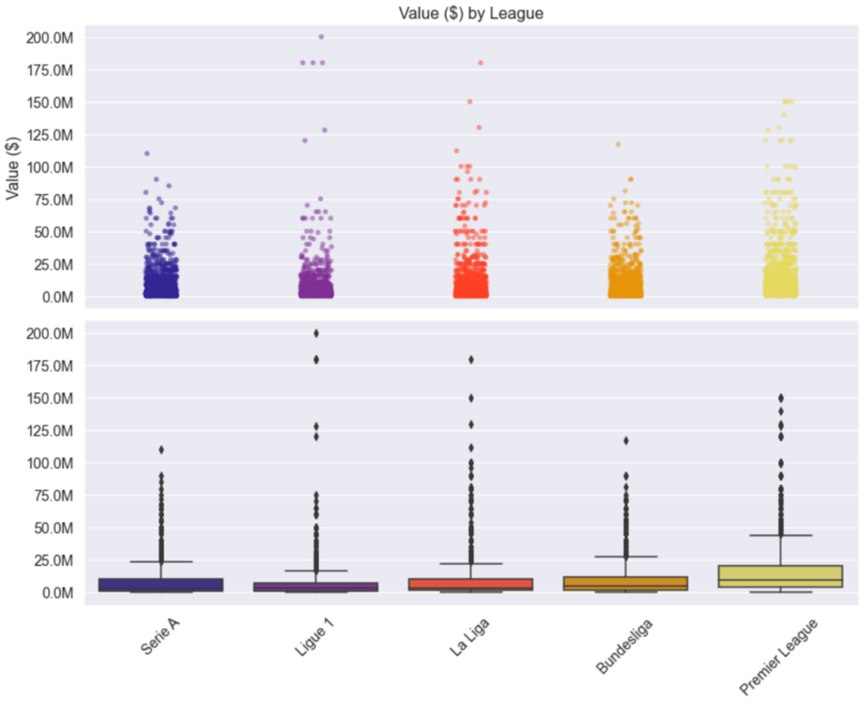
This initiative raises significant ethical issues. Although statistically relevant, including demographic characteristics like age, ethnicity, or physical traits could introduce or perpetuate societal and cultural prejudices if not rigorously controlled. For instance, if a model gives players of a certain ethnicity lesser predicted values owing to historical underrepresentation or team selection bias, this would unfairly disadvantage members of that background.

Furthermore, a thorough verification of the openness and justice of machine learning systems in high stakes financial choices is needed. Clubs depending only on model forecasts could miss qualitative elements like leadership or teamwork that are not reflected in the data. It's also crucial to solve data privacy issues, particularly if the tool is ever grown to cover private medical or financial information. Data-driven understanding should always aim to enhance, not substitute, human judgment, and the model should be constantly reviewed for fairness, accuracy, and possible bias.

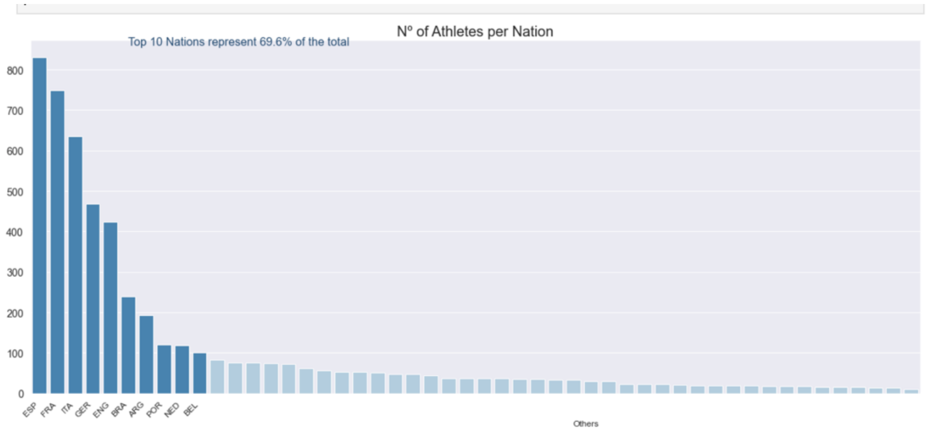
Visualizations and Storytelling

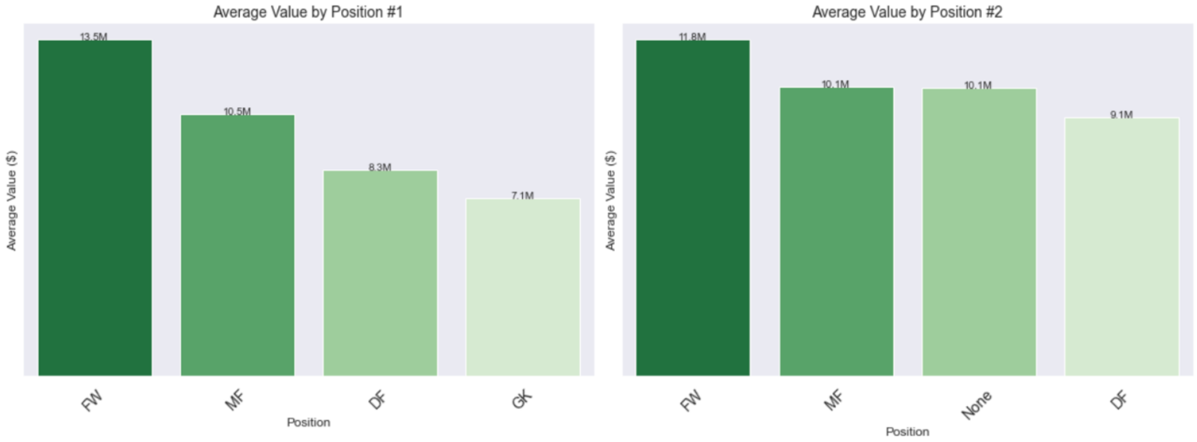
Effective storytelling through visualizations played a key role in conveying the insights of this project. Several types of plots and charts were used to aid interpretation and engage the audience.

- Visualization to show the average value by player league.



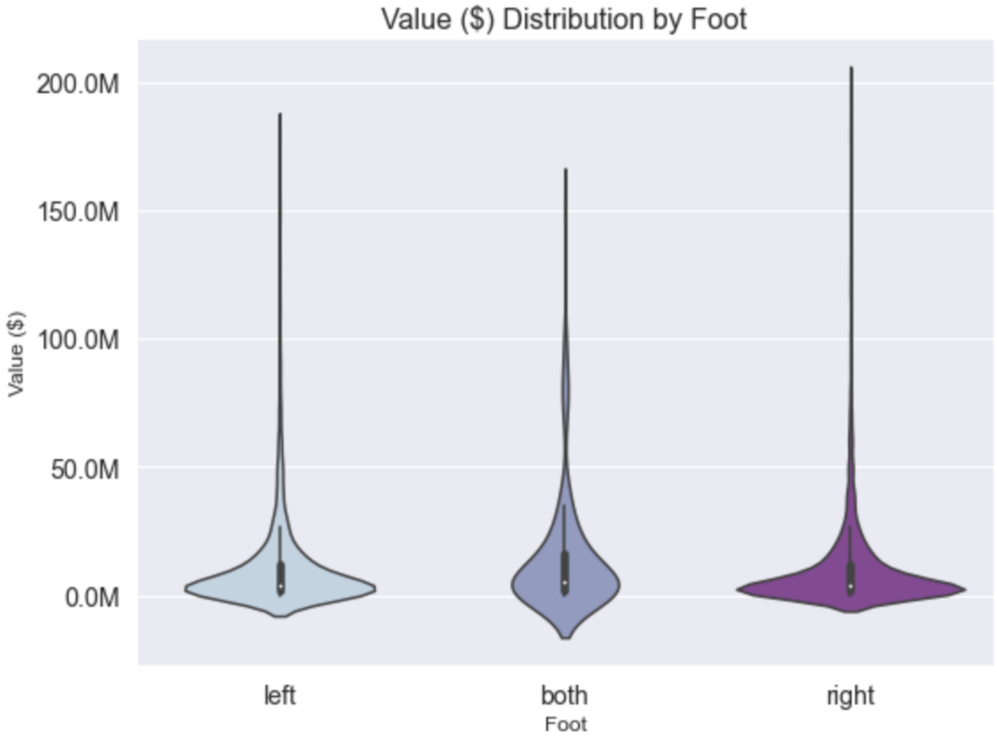
- Visualization to visualize the distribution of player market values based on categorical variables like position or nationality.



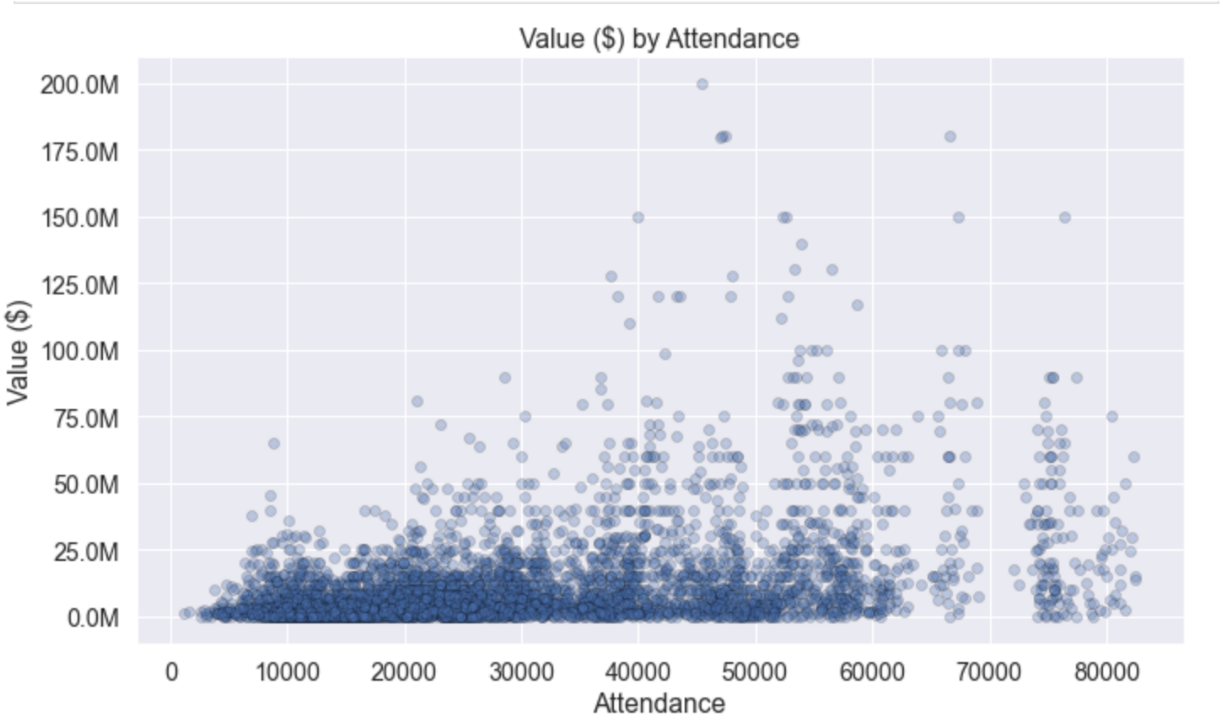


We can see that is a strong relationship between the 1st position and value, but not so much with the second position.

- Visualization to show the average value by player Foot.



- **Scatter plots** to explore correlations between continuous variables such as player attendance or goals vs. market value.



These visualizations provided a narrative flow from data exploration to model diagnostics, and ultimately strengthened the clarity and impact of the findings.

Questions and Answers Audience Would Ask

Anticipating potential questions from an academic or professional audience is crucial to fully understanding the strengths and limitations of this project. Below are ten likely questions that may arise during the final presentation:

1. **What were the key features that most strongly influenced the player value predictions?**

   The most influential features included appearances, minutes played, age, goals scored, and player position. These had high feature importance in the Gradient Boosting Regressor, highlighting their predictive power.

2. **Why did the linear regression model perform so poorly, even after scaling?**

   Linear regression assumes linear relationships and is sensitive to outliers. Despite scaling, the model returned an R^2 of -4.06e+24 and an RMSE of 3.65e+19 due to the dataset's non-linear nature and large variance in values.

3. **How was the dataset validated to ensure accuracy and completeness?**

   The dataset underwent rigorous validation including null value handling, duplicate removal, and outlier inspection. The preprocessing steps ensured only quality, non-anomalous data were fed into the models.

4. **Can the model be generalized to leagues or seasons not represented in the dataset?**

   While the current model is built on 2017–2020 European data, it can be generalized with retraining. Periodic model updates are essential to account for evolving league dynamics and player attributes.

5. **What measures were taken to control for bias in player nationality, position, or league?**

   One-hot encoding and careful feature selection prioritized performance metrics over demographic identifiers. Categorical features were encoded neutrally to mitigate systemic bias.

6. **How interpretable is the final model, and what tools can help explain individual predictions?**

   Although ensemble methods are complex, tools like SHAP and LIME offer transparency by showing how each feature contributes to a player's valuation, helping stakeholders interpret results.

7. **How does this model compare to existing valuation systems like Transfermarkt in terms of accuracy?**

   Unlike Transfermarkt crowdsourced estimates, this model provides reproducible, data-driven valuations with an R^2 of 0.789, making it a strong alternative for predictive reliability.

8. **Could external factors such as injuries, team changes, or popularity be integrated to improve predictions?**

   Yes, integrating variables like injury history, social media sentiment, or transfer rumors would enhance prediction quality, particularly for real-time applications.

9. **How will the model be updated or retrained with future data to maintain accuracy?**

   A retraining schedule using data from new seasons is planned. Automating updates ensures the model adapts to market trends and retains predictive accuracy.

10. **What ethical safeguards should be in place when using such models in high-stakes transfer decisions?**

    Models must be audited for fairness and explainability. Sensitive attributes should be handled with caution, and model results should inform, not dictate, high-stakes decisions.

These are the answers to the questions which I asked in Milestone 4.

Appendix

This section includes visual illustrations such as model comparison charts, residual error plots, and violin plots showing distribution by player attributes. A data dictionary is also provided that summarizes key features used in the model along with their definitions and types. Finally, code snippets for the modeling pipeline, hyperparameter tuning, and evaluation metrics are documented to support reproducibility.

References

- Kriegsmaschine. (2020). *Soccer players' values and their statistics (Kaggle)*. https://www.kaggle.com/kriegsmaschine/soccer-players-values-and-their-statistics
- FBref.com and Transfermarkt.com for raw and comparative soccer data
- FIFA 21 Dataset – available from various public scraping projects and Kaggle datasets
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*
- Articles on fairness and bias in sports analytics, e.g., Sloan Sports Conference papers