A SOFT-ROUTING PIPELINE TO PREDICT THE VALUE OF FOOTBALL PLAYERS

Lorenzo Caputi 3240499, Andrea Procopio 3248158, Filippo Strub 3261811, Andrea Pettenon 3257784, Paride Lauretti 3247740

ABSTRACT

This study addresses the problem of predicting the market value of professional football players using a high-dimensional dataset comprising demographic, contractual, and performance-related features. The presence of nonlinear relationships, multimodal feature distributions, and distributional outliers among top-valued players limits the effectiveness of standard regression models. To address these challenges, a soft-routing pipeline is proposed, consisting of a binary classifier to identify high-value players, a base regressor trained on the full dataset, and a specialized regressor trained exclusively on top-valued players. The proposed model produces final predictions through a weighted combination of the two regressors, with weights derived from the classifier's output. Empirical results show that the proposed model improves predictive accuracy and reduces variance across cross-validation folds, outperforming conventional tree-based methods and demonstrating the benefits of structure-aware modeling in heterogeneous regression settings.

DATA AND TASK

The dataset provided consists of two main parts: a training set (*train.csv*) and a test set (*test.csv*), both of which contain information on professional football players. The training set has approximately 14,000 observations, each representing a unique player described by a high-dimensional feature set that provides demographic details (e.g., age, height, weight), contract details (e.g., wage, release clause, club), positional information, and over 40 technical performance indicators (e.g., passing, dribbling, positioning).

The most significant difference between the two datasets is that the target variable *value_eur* (a player's market value in Euros) is present only in the training dataset. The task is to forecast this target variable for players in the test dataset, using observed features. The prediction task is thus a regression problem with a high-dimensional data of mixed types (numeric, categorical, ordinal). The loss function considered for the problem is the Root Mean Squared Error (RMSE).

EXPLORATORY DATA ANALYSIS (EDA)

The exploratory data analysis reveals several key characteristics of the dataset that have direct implications for our modeling strategy.

First, regarding the distribution of the variables, the target, *value_eur*, is highly right-skewed, as shown in the kernel density plots on log-transform scales **(Figure 1)**. This **skewness** extends to many explanatory variables, such as wage_eur and release_clause_eur, motivating a log transformation to both the target and selected features to approximate a more normal-like distribution and stabilize variance.
Moreover, the distribution of explanatory variables, particularly those describing players' technical attributes, are often **multimodal**. These distributions reflect the heterogeneity of player roles: goalkeepers, for instance, have high values in attributes related to their role, such as goalkeeping_diving, while outfield players show near-zero values. Other attributes such as defending or attacking_finishing are shared among roles but with varying distributions. This suggests that the dataset contains a few subpopulations with different dynamics, and it has important implications for modeling, since models must be capable of capturing this multimodality.

As regards the relationship of features with the target, the scatterplots in **Figure 7** show that they are often nonlinear, supporting further the use of log-based transformations and flexible models able to capture complex input-output relationships.
There are numerous features with strong correlations with value_eur, particularly those reflecting on-field performance and contract terms like overall, potential, and wage_eur. But the strength of these correlations can vary significantly across player positions. For

instance, the overall correlation of shooting with the label is 0.29, while for attackers it is 0.5.
The analysis also reveals strong multicollinearity between technical attributes, particularly for role-specific subgroups. Features like dribbling and skill_dribbling are very highly correlated, especially for the midfielder position (corr=0.97). This high redundancy can lead to a higher risk of overfitting and decreased model interpretability.

PREPROCESSING PIPELINE

A. Format normalization and variable pruning

Training and test data were equally preprocessed. All variables in the datasets were converted to lower-case snake_case to enforce a consistent naming and remove white-space or punctuation artefacts. Furthermore nineteen identifier-type and media-link columns carrying no predictive signal for market value (e.g. *player_face_url*, *club_logo_url*) were discarded, alongside *nation_position* as it was missing for the 96% of players in the datasets. Moreover, observations missing the target variable *value_eur* were removed.

B. Missing data strategy

As the proportion of missing values in each feature, when present, was always above 5%, all numeric features with missing values were zero imputed. For each such feature, a structural missing flag was appended, indicating that "missing" may itself be informative rather than a measurement error (e.g. players missing *shooting* are all goalkeepers).

C. Domain specific feature engineering

Features specific to the football domain were engineered, extracting relevant information from high-cardinality variables by storing them in low-cardinality ones. This cardinality reduction reduces the overall dimensionality of the feature space, and prevents overfitting. The table in **Figure 6** illustrates all the features extracted, the process, and the rationale.

D. Encoding Categorical Predictors

Categorical predictors were partitioned by cardinality. For low-cardinality features (≤15 classes) One-Hot-Encoding was performed, while a 5-fold cross-validated target encoding was performed on high-cardinality features. The choice for a 5-fold encoding was made to prevent data leakage. In this case, data was split into 5 folds, then for every fold, for every variable, every category is replaced with the target mean of the other folds of that category.
Original string columns were then dropped to prevent any duplication.

E. Non-linear transformations and metric scaling

Continuous features previously explored that exhibited possible exponential or quadratic relations with the market value were transformed. Specifically, alongside with the square root of *age*, log-transformed version of *wage*, *release_clause_eur*, *overall*, *potential*, and *international_reputation*, was obtained. While a log-transformed version of target *value_eur* was also obtained, training experiments reported negative results on the viability of employing *log_value_eur* as the target variable.

Original raw variables were not dropped from the datasets to allow models for flexibility in identifying the most informative features during training.

Although tree ensembles are scale-invariant, z-score standardisation (mean 0, variance 1) was applied to all remaining numeric predictors excluding targets, to allow for a more flexible experimentation in the model selection phase. Standardisation preserves flexibility to experiment with distance-based or gradient-based learners, such as k-NNs or neural nets, without re-engineering the pipeline.

MODEL SELECTION

A. Evaluation Method

To compare model performance, we adopted 5-fold cross-validation with Root Mean Squared Error (RMSE) as the evaluation metric. This approach provides an unbiased estimator of out-of-sample performance and has smaller variance compared to a simple holdout (train/test) split. Unlike out-of-bag (OOB) error estimation, which is model-specific for ensemble methods such as Random Forest, cross-validation can be applied evenly across all classes of models, enabling a consistent and fair comparison.

B. Base Models

We tested three main categories of models: linear, non-parametric, and tree based.

The **linear**, ridge, and lasso models all performed quite similarly, with validation RMSEs of around 1.355 million euros and training RMSEs also well above 1.35 million. This is a clear indication of underfitting, as the models fail to capture the underlying data structure. This is consistent with the exploratory analysis: the relationship between features and the target is highly nonlinear, and there are several features with complex interactions, multimodality, and conditional effects based on player roles.

A **K-nearest neighbors (K-NN)** model was also attempted to determine if a basic non-parametric model would be able to learn local structure. However, even in the best case of K=4, it performed poorly, with a validation RMSE of 1.56 million and a training RMSE of 1.23 million. This difference suggests that while the model is fitting the training data slightly better than linear models, it is not yet generalizing. This can be explained by high dimensionality and multimodality. On one hand, in high-dimensional spaces, the concept of distance becomes ambiguous, a phenomenon known as “curse of dimensionality”. As the number of features grows, distances between data points tend to homogenize, making it difficult for models like k-nearest neighbors (k-NN) to identify truly relevant neighbors. On the other hand, distance metrics (like Euclidean) are not meaningful when different dimensions dominate in different modes and the method cannot distinguish between semantic proximity (being truly similar) and numerical proximity (being close in Euclidean space).

Tree-based ensemble methods achieved the best results by a wide margin. All of the ensemble methods (Random Forests, Gradient Boosting, LightGBM, XGBoost, and CatBoost) achieved validation RMSEs below 800k, with the best-performing model, Gradient Boosting, at 687,281 euros with smallest standard deviation (114,170) among folds.
These models perform better as they are naturally able to deal with nonlinearities, feature interactions, and categorical features. They also do not require making strong features independence or distribution shape assumptions. Furthermore, they are able to implicitly leverage the multimodality of the feature space revealed through our exploratory analysis—e.g., by splitting differently for goalkeepers than for outfield players.

C. Soft-routing pipeline

To further improve performance, we investigated where even the best-performing tree-based models were failing. The goal was to identify systematic sources of instability across folds and high RMSE.
To do this, we trained an **autoencoder** to project the dataset onto a 2-dimensional nonlinear manifold by minimizing reconstruction error. This embedding (**Figure 2**) provided a visual summary of the dataset structure and confirmed the multimodality of the feature space, something we had previously addressed during feature selection.
It also revealed a deeper issue: the highest-valued players are not only outliers in terms of market value, but also **distributional anomalies** in the feature space. These players lay in low-density regions of the input space, with a different structure from the main clusters of typical players.
This was supported by a **leave-one-bin-out** error analysis (**Figure 3-4**): we partitioned the data into bins by target quantiles, leaving one bin out as the validation set. This revealed that error in prediction is highly localized in the top quantile of the target distribution, even after normalizing for the mean target of each bin. Moreover, removing the highest 20% of players from the data, RMSE dropped by 90% to approximately 70k, demonstrating that these outliers disproportionately dominate the loss function. Their rare and uneven presence also explained the high variance in validation RMSE across cross-validation folds. In “lucky” splits where most top players appeared in the training set, even simple models like random forests achieved validation RMSEs as low as 400k. In “unlucky” splits, where those same players fell into the validation set, RMSE often exceeded 1 million.
It was clear that to improve performances, greater attention should be placed in predicting outliers, which are distributional anomalies. This motivated our solution in the form of a 3-model **soft-routing pipeline** (**Figure 5**), to explicitly account for the upper outliers. The pipeline includes:

1. A binary classifier trained to predict whether a player is a "top player," defined as being in the top 5% of the value_eur distribution;
2. A base regressor trained on the full dataset;
3. A top regressor trained on top players only.

At prediction time, we compute a weighted average of the outputs of the top and base regressors, where the weight is given by the estimated probability of the classifier that a player belongs to the top group. The 0.95 quantile cut-off was selected (again with 5-fold

CV) as a compromise: narrow enough to isolate genuine outliers and detect their dissimilar behavior, but wide enough to have sufficient data to train the tailored model without excessive overfitting.

Tree-based base models generally work better when merged with a suitable top model in the soft-routing pipeline. The best-balanced, most generalizable combination uses Random Forests as the classifier and base regressor, and CatBoost as the top regressor, offering a good balance of stability and specialization. This yields a RMSE of 609265 on 5-fold CV.

The Random Forest model was chosen as both the classifier and the base regressor, mainly due to its robustness to overfitting and class imbalance.
In fact, both the binary classification and regression excluding the upper outliers are relatively easy tasks, with dense data and stable patterns, where even a simple model can achieve low bias. So the main focus here was to avoid overfitting.

The CatBoost top regressor is appropriate to pick up on the small, structurally variant subset of valuable players. Its ordered boosting and regularizer function especially in the low-data conditions where overfitting problems happen, such as in the upper 5% of the value distribution.

The configuration allows for each model to operate on its best set of data, providing improved performance and generalization, as shown in **Figure 8**.
We can notice how the soft-routing pipeline improves performances of all tree-based methods, reducing both the bias (average RMSE in 5-fold CV) and the variance (standard deviation of RMSE across folds).

CONCLUSION AND FUTURE WORK

This project addressed the challenge of predicting football players' market value from a rich, high-dimensional dataset of demographic, contractual, and technical features. Initial exploratory analysis revealed the presence of nonlinear relationships in the data as well as multimodality and heterogeneity, driven by different roles and upper outliers, highlighting the need for models that can accommodate complex, heterogeneous relationships. While ensemble tree-based algorithms like Random Forest and Gradient Boosting already performed the best over linear and non-parametric baselines, further improvements were achieved with a soft-routing pipeline. The framework explicitly addressed structural variability across player subpopulations and significantly improved accuracy and consistency in cross-validation.

In the future, there are a few directions possible for further work. One is to improve the current heuristic-based averaging process with a sophisticated meta-model capable of learning how best to combine the predictions of the base and top regressors with the input features. This would potentially enable context-dependent ensembling, dynamic weighting of the aggregation weights against player attributes rather than static probability. In addition, integrating richer external information like recent performance data and records of injuries, might provide valuable signals, particularly for high-value assets. Together, these extensions offer a path toward an intelligent and more generalizable valuation model.

COMPLEMENTARY FIGURES

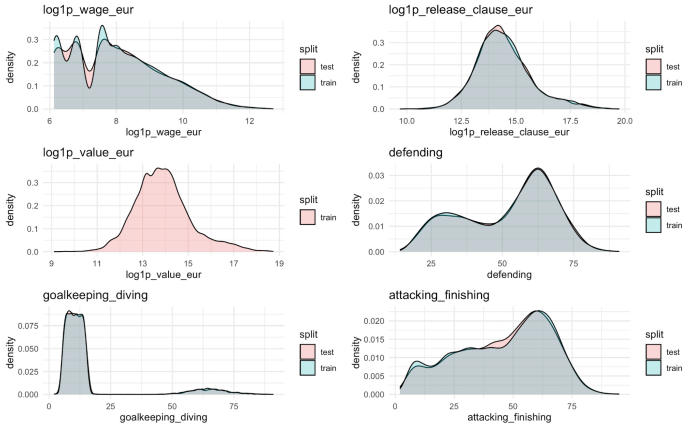


Figure 1: Density plots of relevant selected variables overlapped between train and test set, showing matching distributions for all.

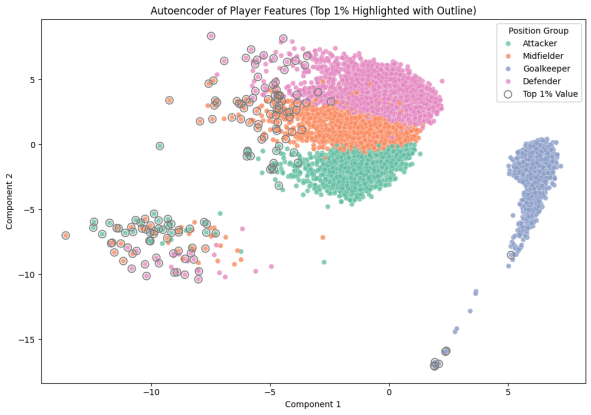


Figure 2: Autencoder clustering scatterplot, evidentiating the need for a separated outlier-speicific regression model

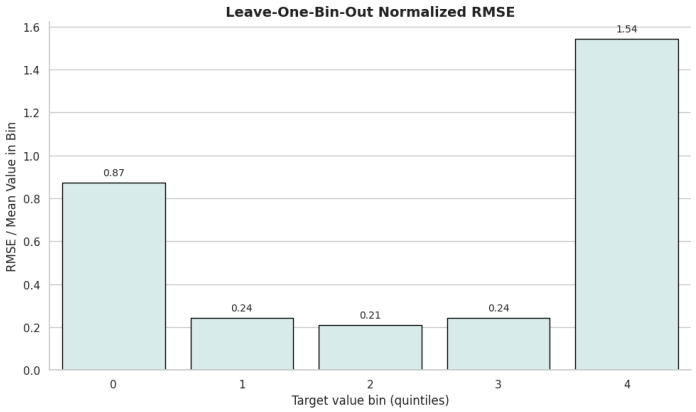


Figure 3: Barplot showcasing Leave-One-Bin-Out normalized Root Mean Squared Error

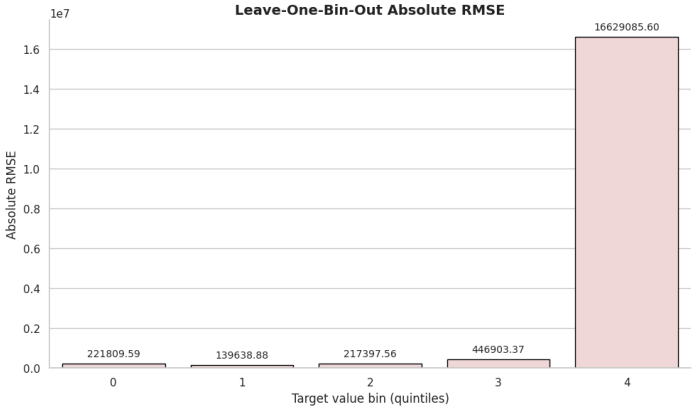


Figure 4: Barplot showcasing Leave-One-Bin-Out absolute Root Mean Squared Error

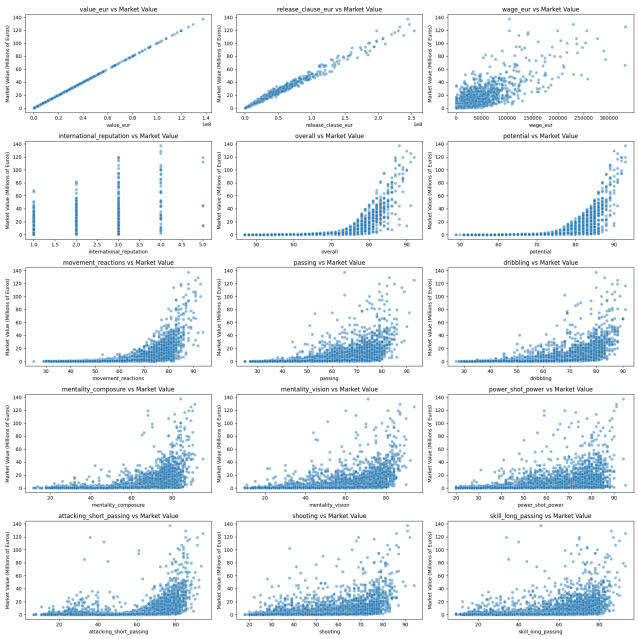
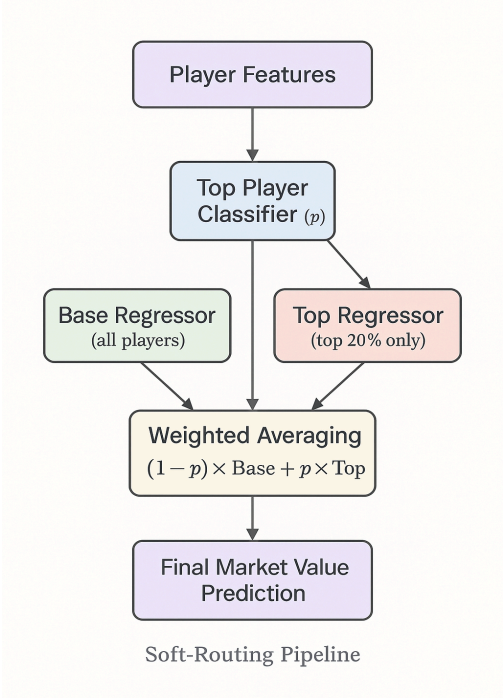


Figure 7: Scatterplot showcasing linear and nonlinear relationships between selected variables and the arget



Soft-Routing Pipeline

Figure 5: Proposed model's architecture, comprehensive of classifier, standard and outlier-specific regressors, and weight averaging

Constructed Feature	Definition	Motivation
position_group	Raw position code mapped into: {Goalkeeper, Defender, Midfielder, Attacker}	Reduces cardinality of player_positions while capturing tactical roles
bmi	weight_kg / (height_cm*100)^2	Normalised body mass indicator related to physical performance combining height and weight
priority	Ordinal' status extracted from club_position. 1 for reserve, 2 for substitute, 3 for starting team	Make explicit whether a player is central for its team. Important since we don't have access to performance metrics.
contract_remaining	Computed as contract expiry date minus minimum expiry year	Proxy for bargaining position and residual contract value. Teams usually set up long contracts for their top players.
years_at_club	represents the calendar difference (as of Jan 1st 2022) between club_joined and reference date	clubs specific tenure may capture related loyalty premiums and is an indication of consistency in performance
defense_work_rate & attack_work_rate	ordinal interpretation split of composite work_rate string encoded as integers (Low = 0, Medium = 1, High = 2)	Separates effort orientation into phase-specific intensities

Figure 6: table comprehensively illustrating the specific features engineered, the extraction process, and the rationale behind it.

base model	train RMSE	val RMSE	Std RMSE
soft-routing	261 736	609 265	119 587
Gradient Boosting	317 989	687 281	114 170
LightGBM	173 631	726 637	154 426
Catboost	188 517	757 617	139 852
XGBoost	118 890	774 346	136 481
Random Forest	301 066	792 855	115 976
Linear regression	1 351 167	1 355 672	210 195
Lasso Regression	1 351 167	1 355 672	210 195
Ridge regression	1 351 168	1 355 681	210 082
4-NN	1 227 723	1 559 706	129 888

Figure 8: table summarising results of all models benchmarked on 5-fold cross validation