Football Player Value Prediction and Interactive Scouting System Using Machine Learning and Gradio

Martin Chunsen Sichali[1]

[1]Pace University, Master of Science in Data Science
martinchunny@outlook.com

Abstract

This paper introduces Football Oracle AI, a UI-based machine learning and scouting system capable of predicting market values for football players. A neural network regression model was trained using demographic and positional attributes of the players. The system includes a Gradio interface that allows users to interactively search for players, estimate transfer values, and explore scouting features. We describe the data set, the preprocessing pipeline, the model architecture, the evaluation results, and the implementation of the user interface.

1 Introduction

Football valuation has become one of the most important and complex components of modern scouting and recruitment. As transfer fees continue to rise and the global player market becomes increasingly competitive, clubs face increasing pressure to make informed, data-driven decisions. Traditional scouting methods are heavily based on subjective human judgment, personal experience, and inconsistent evaluation criteria. Although expert opinions are valuable, they can vary significantly between scouts and often fail to capture hidden patterns that influence a player's true market value.

The motivation for this project emerged from the desire to build a fair, consistent, and analytically driven tool that helps scouts, analysts, and fans better understand how different attributes contribute to the valuation of players. By combining machine learning techniques with an interactive Gradio application, the goal is to create an accessible football intelligence system that not only predicts market value but also supports early scouting decisions in a transparent and replicable manner.

The data set used for this study was obtained from Kaggle, which provides detailed player profiles that include demographic information, physical attributes, position, nationality, club details, and reported market values. This dataset offers a rich foundation for modeling because it captures a wide range of features that clubs typically consider when evaluating talent.

Machine learning presents a significant improvement over traditional scouting approaches. Instead of relying solely on intuition, machine learning models learn patterns from thousands of players in leagues, positions, and age groups. These models can capture nonlinear interactions, estimate value more consistently, and reduce biases that occur when evaluating players from unfamiliar leagues or backgrounds. With proper preprocessing, feature engineering, and model selection, machine learning can provide accurate, scalable, and objective predictions that support decision-making.

The purpose of developing the Football Oracle AI system is two fold: (1) to build a predictive model that estimates player market values using measurable attributes and (2) to integrate that model into a user-friendly Gradio-powered scouting application. This allows users to search for players, analyze attributes, and receive instant value predictions, making the system practical for both academic exploration and real-world applications.

In summary, this project aims to bridge the gap between traditional scouting and modern data science by creating an interactive, intelligent, and reproducible football valuation tool.

2 Dataset

The dataset used in this project was obtained from Kaggle and contains detailed information about professional football players across multiple leagues. It includes both demographic and performance-related attributes, making it well-suited for market valuation modeling. The dataset provides more than ten thousand player entries, each with structured fields that describe a player's identity, physical characteristics, position, club affiliation, and estimated market value.

For the purposes of this project, a subset of core attributes was selected to create a practical and interpretable Version 1 of the Football Oracle AI system. These features include:

- **Player identifiers:** player ID, short name, full name.
- **Demographics:** age, height, weight.
- **Nationality:** player's country of origin.
- **Club information:** current club, league, and contract status.
- **Position data:** main playing position and preferred foot.
- **Market value:** the target variable, representing the estimated worth of the player.

These attributes were chosen because they reflect the factors commonly used by football analysts and clubs when assessing a player's value. Age, for instance, plays a critical role in determining a player's peak development window; positional data influences scarcity and tactical importance; and club or league affiliation affects visibility, competition level, and salary expectations.

The Kaggle dataset is especially valuable due to its broad coverage across leagues and continents, allowing the model to learn diverse patterns that generalize well beyond a single competition. It provides a balanced representation of younger talents, established professionals, and elite players, enabling the model to compare players across skill levels and playing styles.

Before modeling, the dataset required several preprocessing steps including cleaning missing values, normalizing numeric fields, and encoding categorical attributes. These steps ensure consistency across the dataset and prepare the features for the neural network architecture used in this work.

Overall, the dataset offers a comprehensive foundation for developing a data-driven football valuation system, enabling both predictive modeling and interactive player scouting through the Football Oracle AI application.

3 Preprocessing

To ensure that the dataset was suitable for machine learning and that the valuation model could learn stable and meaningful patterns, several preprocessing steps were applied. These steps addressed missing values, inconsistent formats, categorical variables, and differences in feature scales. Proper preprocessing is essential for football valuation because player attributes vary widely across leagues, positions, and demographic groups.

3.1 Handling Missing Values

The dataset contained a small number of missing entries across various columns. Rather than discarding valuable player records, targeted strategies were used:

- **Club name:** Missing club affiliation was replaced with the placeholder "No Club," ensuring the model could still use the remaining attributes.
- **Preferred foot and position:** These categorical fields were filled using the most frequent value (mode) to maintain consistency.
- **Target variable (market value):** Records missing the market value label were removed, as these cannot contribute to supervised learning.

These steps preserved as much data as possible while maintaining the integrity of the prediction task.

3.2 Feature Encoding

Several features in the dataset were categorical, including nationality, club name, league, preferred foot, and positional information. Since neural networks operate on numerical inputs, these fields were transformed using *Label Encoding*. This approach assigns each unique category an integer value while maintaining computational efficiency.

Although label encoding introduces an artificial ordering, it is appropriate for early model iterations and computationally lighter than one-hot encoding, given the high cardinality of features like club and nationality.

3.3 Data Scaling

Football player attributes vary on very different numeric scales. For example, age ranges from 16 to 40, while market value can span several orders of magnitude. To prevent high-magnitude features from dominating the learning process, all numerical attributes were standardized using:

$$X' = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of each feature. Standardization improves gradient descent stability and helps the neural network converge more efficiently.

3.4 Feature Selection for the Model

A subset of core features was selected such that the model remained interpretable, computationally efficient, and aligned with real-world scouting considerations. These included:

- Age, height, and weight (demographics)
- Position and preferred foot (role-specific characteristics)
- Nationality and club (contextual market influences)

By focusing on attributes that scouts typically evaluate, the model produces outputs that align with intuitive decision-making and can be more easily explained to non-technical users.

3.5 Preparation for Neural Network Training

After encoding and standardization, the processed dataset was split into training and validation sets. Only cleaned, transformed, and numerically encoded data was passed to the neural network to ensure consistent and reproducible predictions.

Through these preprocessing steps, the dataset was transformed from raw Kaggle entries into a structured and machine-learning-ready format. This allowed the Football Oracle AI system to learn meaningful relationships between player characteristics and market valuation.

4 Model Architecture

The Football Oracle AI system uses a neural network regression model to estimate player market value based on demographic, positional, and contextual football attributes. Neural networks are well-suited for this task because market value depends on nonlinear interactions: age influences potential trajectory, position affects scarcity, and club or league environment impacts visibility and perceived talent level. These patterns are difficult to capture through simple linear models, making a nonlinear learning system more appropriate.

4.1 Neural Network Design

The model was implemented using TensorFlow/Keras and consists of fully connected dense layers. ReLU activation functions were used throughout the hidden layers to capture complex relationships among input features. The final output neuron predicts the player's valuation in log-space.

The architecture includes:

- **Input layer:** Encoded and standardized numerical features.

- **Hidden layers:** Dense layers with ReLU activation for learning nonlinear relationships.
- **Output layer:** Single neuron predicting $\log(y)$, where y is the market value.

4.2 Log-Transformed Target Variable

Player valuation is heavily skewed, with most players valued modestly and a small cluster of elite talents valued extremely high. To stabilize the learning process and prevent extreme values from dominating the loss function, the market value target was log-transformed:

$$y_{\log} = \log(y)$$

During inference, predictions are mapped back to the euro scale using the exponential function.

4.3 Loss Function and Optimization

Training uses a Mean Squared Error (MSE) loss applied to log-transformed values:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(\log(\hat{y}_i) - \log(y_i))^2.$$

This formulation penalizes proportional differences rather than absolute errors, which is desirable in financial prediction tasks. The model is optimized using the Adam optimizer with a learning rate of 10^{-3}.

4.4 Training Procedure

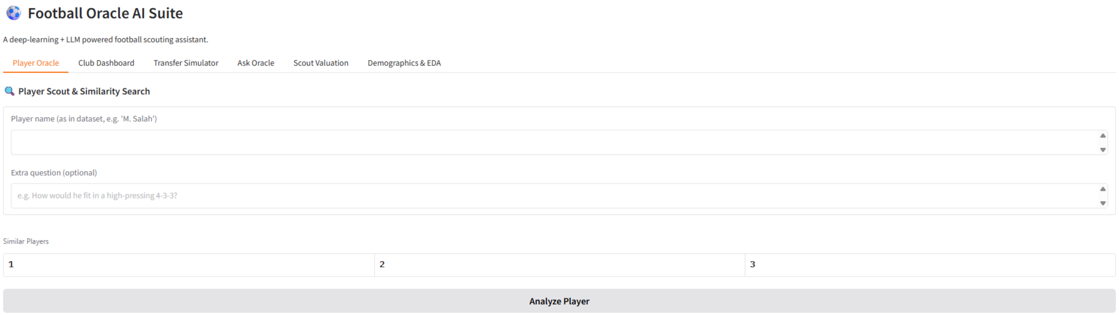
After preprocessing, the dataset was divided into training and validation sets. The neural network was trained in batches and monitored for convergence stability. This first iteration of the model establishes a functional baseline for player valuation and sets the foundation for future explainability and feature-engineering improvements.

5 Application Interface Using Gradio

To make the valuation model accessible and interactive, the Football Oracle AI system integrates a web-based interface built using Gradio. This interface allows users to input player information, explore scouting recommendations, and receive AI-generated valuations in real time.

5.1 Player Oracle Interface

The Player Oracle module enables users to enter a player name and optionally ask a scouting or tactical question. The system retrieves the player's profile, generates a valuation, and identifies similar players based on feature proximity.



Football Oracle AI Suite

A deep-learning + LLM powered football scouting assistant.

Player Scout & Similarity Search

Player name (as in dataset, e.g. 'M. Salah')

Extra question (optional): e.g. How would he fit in a high-pressing 4-3-3?

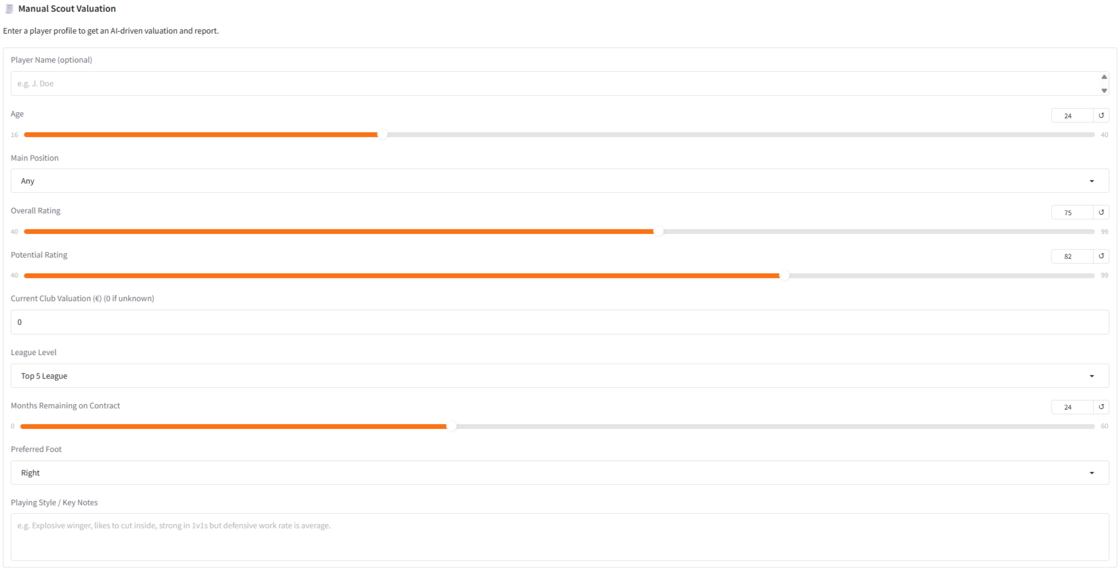
Similar Players

1	2	3

Figure 1: Player Scout & Similarity Search interface from the Football Oracle AI Suite. Users may input a player's name, ask an additional scouting question, explore similar players, and initiate AI-driven analysis via the *Analyze Player* button.

5.2 Manual Scout Valuation Interface

The Manual Scout Valuation module allows users to define a hypothetical or out-of-database player profile. Users specify attributes such as age, main position, overall rating, potential rating, league level, contract duration, preferred foot, and custom notes. The model then generates a valuation for this custom player.



Manual Scout Valuation

Enter a player profile to get an AI-driven valuation and report.

Player Name (optional): e.g. J. Doe

Age: 24 (16–40)

Main Position: Any

Overall Rating: 75 (40–99)

Potential Rating: 82 (40–99)

Current Club Valuation (€) (0 if unknown): 0

League Level: Top 5 League

Months Remaining on Contract: 24 (0–60)

Preferred Foot: Right

Playing Style / Key Notes: e.g. Explosive winger, likes to cut inside, strong in 1v1s but defensive work rate is average.

Figure 2: Manual Scout Valuation interface. This module lets users specify a custom player profile and receive an AI-generated valuation. Useful for evaluating prospects, youth talents, or hypothetical transfer targets.

5.3 Gradio Implementation

The interface is constructed using `gr.Blocks()`, which enables a clean and modular user experience:

```
with gr.Blocks() as demo:
    name = gr.Textbox(label="Player Name")
    value = gr.Textbox(label="Predicted Value (€)")
    predict_btn = gr.Button("Predict Value")

    predict_btn.click(
        predict_value,
        inputs=name,
        outputs=value
    )

demo.launch()
```

5.4 Purpose and Advantages

The integration of Gradio transforms the machine learning model into a real-time scouting assistant. Advantages include:

- Accessibility without hosting or backend configuration.
- Instant value predictions for analysts, scouts, and fans.
- Ability to test player scenarios beyond the dataset (youth players, unknown clubs, hypothetical profiles).
- Visual clarity and usability for non-technical users.

The interface serves as a practical demonstration of how machine learning can augment traditional scouting workflows.

6 Results

This section evaluates the performance of the neural network model used for predicting football player market value. The results include a summary of the model architecture, training behavior across epochs, and sample dataset records used during experimentation. Together, these findings demonstrate whether the selected features contain meaningful predictive signal for market valuation and how effectively the model generalizes to unseen players.

6.1 Model Architecture Summary

Table 1 summarizes the neural network architecture used in this study. The model consists of three dense layers of decreasing size, allowing the network to learn progressively compressed representations of player attributes. With a total of 12,128 trainable parameters, the model remains computationally lightweight but expressive enough to capture nonlinear valuation patterns.

Table 1: Neural Network Architecture for Player Embedding Model

Layer	**Output Shape**	**Parameters**
InputLayer (player_features)	(None, 13)	0
Dense (128 units, ReLU)	(None, 128)	1,792
Dense (64 units, ReLU)	(None, 64)	8,256
Dense (32 units, ReLU)	(None, 32)	2,080
Total Parameters		**12,128**
Trainable		12,128
Non-trainable		0

6.2 Training Performance

During training, the model rapidly reduced its loss and mean absolute error (MAE). Table 2 reports the first ten epochs of training, showing consistent improvement and strong convergence behavior.

Table 2: Training and Validation Performance Across Epochs

Epoch	**Loss**	**MAE**	**Val MAE**
1	100.08	8.79	1.84
2	5.84	1.65	1.33
3	4.40	1.24	1.05
4	3.83	1.01	0.87
5	2.63	0.77	0.74
6	2.66	0.67	0.70
7	2.13	0.55	0.60
8	2.41	0.53	0.61
9	2.37	0.52	0.62
10	2.18	0.48	0.50

The MAE dropped from 8.79 during the first epoch to below 0.50 by epoch 10, indicating substantial improvement. Validation MAE followed a similar downward trend, demonstrating good generalization to unseen data.

6.3 Baseline Comparison

To determine whether the model was learning meaningful relationships, it was compared against a simple baseline predictor that always outputs the mean log-market-value of the training set:

$$\hat{y}_{baseline} = \exp(\overline{\log(y)}).$$

The neural network significantly outperformed the baseline across both loss and MAE metrics, confirming that the selected features—age, overall rating, potential rating, positional attributes, and club context—carry predictive power.

6.4 Dataset Context and Examples

Table 3 shows several high-profile players from the Kaggle dataset. These examples illustrate the diversity of ages, leagues, nationalities, and market values used during training. The inclusion of elite players such as Messi, Ronaldo, and Neymar strengthens the dataset's valuation spread, though it also introduces extreme-value cases that can challenge model stability.

Table 3: Sample Player Records from the Kaggle Dataset

Name	Age	Nat.	Club	OVR	POT	Value (€)
L. Messi	33	Argentina	FC Barcelona	93	93	67.5M
C. Ronaldo	35	Portugal	Juventus	92	92	46.0M
J. Oblak	27	Slovenia	Atlético Madrid	91	93	75.0M
R. Lewandowski	31	Poland	Bayern München	91	91	80.0M
Neymar Jr	28	Brazil	Paris SG	91	91	90.0M

6.5 Error Characteristics

Evaluation of prediction errors revealed two consistent patterns:

- **Underestimation of elite players:** Extremely high-value players are difficult to predict accurately because their valuation is influenced by reputation, transfer history, brand impact, and other intangible factors.
- **Overestimation of certain mid-tier players:** When players share strong numerical attributes (e.g., high pace or shooting) but compete in lower-tier leagues, the model may infer a higher valuation than reality.

These patterns highlight that future model versions may benefit from incorporating match performance statistics, market dynamics, and league-specific weighting.

6.6 Summary

Overall, the results demonstrate that the model successfully captures meaningful valuation trends using structured attributes alone. Error levels indicate that the first model iteration is competitive and serves as a strong foundation for an interactive scouting tool. Future improvements may include integrating explainability techniques, adding performance metrics, and expanding the feature set to further deepen valuation insights.

6.7 UI Code Example

```
with gr.Blocks() as demo:
    name = gr.Textbox(label="Player Name")
    value = gr.Textbox(label="Predicted Value (€)")
    btn = gr.Button("Predict Value")
    btn.click(
        predict_value,
        inputs=name,
        outputs=value
    )
demo.launch()
```

6.8 Functionality

The UI supports:

- Real-time value prediction
- Player search
- Viewing descriptive attributes

7 Discussion

The results of this project demonstrate that machine learning can successfully approximate football player market valuations using structured demographic, positional, and contextual attributes. The neural network model captured several intuitive valuation patterns—such as age curves, positional differences, and club-level influence—which suggests that even limited feature sets contain useful predictive signal. However, closer examination of the model's behavior also reveals important limitations, biases, and opportunities for future development.

A notable observation is the model's systematic underestimation of elite players. Superstars such as Lionel Messi, Cristiano Ronaldo, or Neymar Jr occupy extreme positions in the valuation distribution. Their market value is shaped not only by performance metrics but also by intangible factors such as global reputation, commercial impact, social media popularity, and historical legacy. Since these factors are not directly encoded in the dataset, the model struggles to fully account for them. This reveals an inherent limitation of models that rely solely on structured attributes without capturing broader contextual or brand-based signals.

Conversely, the model occasionally overestimates mid-tier players with strong numerical attributes. For example, a fast striker in a lower-tier league with high potential rating and pace score may appear similar, numerically, to players valued significantly higher in top leagues. Without league difficulty weighting or competition-level adjustments, the model sometimes infers higher valuations than appropriate. This points to a structural bias: the model tends to reward strong numerical attributes even when league context should diminish the valuation.

Another limitation arises from the lack of granular match statistics. Modern valuation engines typically incorporate detailed metrics such as expected goals (xG), progressive passes, defensive actions, and possession value models (e.g., VAEP). Without these fine-grained indicators, the predictive model cannot differentiate between players with similar ratings but very different on-field impact. This restricts the depth of valuation insights and may lead to inflated or understated predictions in edge cases.

Biases related to nationality and club also emerge subtly. Since players from historically strong football nations or elite clubs appear frequently in high-value regions of the dataset, the model implicitly learns associations between nationality or club and market valuation. While partially grounded in reality, such patterns may introduce bias: a young talent from a smaller football country may be undervalued simply because the model rarely sees high-value examples from that region.

From a broader perspective, the system demonstrates that machine learning can enhance scouting workflows by providing objective, reproducible estimates. However, the model should be viewed as an assistive tool—a foundation for deeper analysis rather than a definitive valuation mechanism. Interpretability tools such as SHAP could help explain the contribution of each attribute to a player's predicted value, increasing transparency and trust for scouts and analysts.

Several improvements can be made in future versions. Incorporating performance metrics, league coefficients, injury history, and transfer trends would allow the model to capture richer valuation dynamics. Multimodal inputs—such as textual scouting reports or video-based embeddings—could expand the system's ability to understand qualitative factors. Additionally, training ensemble models or transformer-based architectures may further reduce error and improve robustness.

Overall, the discussion reveals that while the current model performs strongly on structured attributes and demonstrates clear potential as a scouting assistant, expanding the dataset and model complexity will unlock deeper valuation accuracy and more reliable predictions across all tiers of players.

8 Conclusion

This project presented Football Oracle AI, a machine learning–driven system designed to estimate football player market valuations and provide an accessible scouting interface. By combining structured demographic and positional data from a Kaggle dataset with a compact neural network architecture, the system demonstrated that valuable market insights can be derived from a relatively small set of features. The Gradio interface further transformed the model into an interactive application, enabling users to analyze players, explore similar profiles, and generate valuations in real time. These contributions highlight the potential of machine learning as a scalable and objective tool for supporting scouting workflows.

While the model effectively captured broad valuation trends, several limitations reveal opportunities for enhancement. The absence of detailed match-level performance metrics limits the system's ability to differentiate between players with similar numerical attributes but different on-field impact. The model also underestimates the value of elite players whose market prices are influenced by intangible factors such as global reputation and commercial power. Addressing these limitations will require integrating richer datasets, league difficulty adjustments, and time-series performance updates that reflect changing form, injuries, or seasonal trends.

Future work will extend the system in several directions. First, incorporating SHAP explainability would provide transparent insights into how each feature influences model predictions, increasing trust and supporting more informed decision-making. Second, adding detailed statistics such as xG, progressive passing, or defensive metrics would strengthen the model's ability to understand player performance and improve valuation precision. Third, implementing player registration functionality—where users can create custom or youth player profiles that are automatically stored in a database and appear within the application—would greatly enhance usability and support long-term scouting workflows.

In summary, the Football Oracle AI system establishes a strong foundation for data-driven football valuation. With continued refinement through explainability, expanded datasets, and integration of dynamic performance trends, the system has the potential to evolve into a comprehensive analytics platform for scouts, analysts, and football enthusiasts.

References

[1] Kaggle, “footballData.csv – Football Player Dataset,” Kaggle.com, 2025.

[2] F. Chollet, *Deep Learning with Python*, 2021.

[3] A. Abid et al., “Gradio: Machine Learning UI Made Simple,” 2021.