Leveraging Machine Learning and Data Science for Competitive Advantage: Estimating Football Player Market Values

Author: Kavi Patak Supervisor: James Garza

College Dublin
Computing • IT • Business

Introduction

The role of analytics in football has evolved over the past decade, and will continue to do so. Technological developments continue to improve the volume and quality of data available to the world's leading clubs, as well as the ability to derive insight from them. The opportunity exists for clubs of all sizes to use analytics to build a sustainable competitive advantage, something which will be most evident in the area where they invest most: the transfer market.

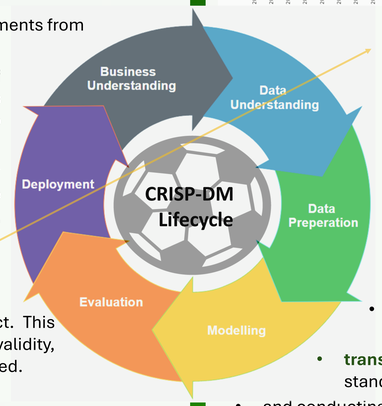
The Methodology...

1. Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective.

Football is a highly lucrative sport that depends heavily on the services of football players. It is therefore no surprise that the majority of a club's expenditure is on player transfers and wages, with smaller clubs unable to compete financially with their larger counterparts.

- This project aims to level the playing field, at least financially in the transfer market, by developing a machine learning (ML) model to estimate the market value of football players based on various features such as age, goals, assists and other contributing factors.



2. Data Understanding

Consolidating and Characterising the Data

In this phase, we gather and understand the data required for the project. This includes identifying data sources, collating data, and exploring its structure, validity, and content. A large dataset Is sourced from Kaggle, and its suitability confirmed.

- Data: "Football Data from Transfermarkt" provided by Kaggle
- Available at: https://www.kaggle.com/datasets/davidcariboo/player-scores/versions/284
- 9 CSV Files:

appearances_df :	(1485697, 13)
club_games_df :	(128586, 11)
clubs_df:	(426, 16)
competitions_df:	(43, 10)
game_events_df:	(652010, 10)
game_lineups_df :	(86822, 9)
games_df :	(64293, 23)
player_valuations_df :	(440663, 9)
players_df:	(30302, 23)

Merged Dataset

```
<class 'pandas.core.frame.DataFrame'>
Index: 30130 entries, 0 to 30301
Data columns (total 79 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   player_id                             30130 non-null  int64
 1   first_name                            30130 non-null  object
 2   last_name                             30130 non-null  object
 3   name                                  30130 non-null  object
 4   last_season                           30130 non-null  int64
 5   current_club_id                       30130 non-null  int64
 6   player_code                           30130 non-null  object
 7   country_of_birth                      27326 non-null  object
 8   city_of_birth                         29861 non-null  object
 9   country_of_citizenship                30130 non-null  object
 10  date_of_birth                         30130 non-null  datetime64[ns]
 11  sub_position                          30130 non-null  object
 12  position                              30130 non-null  object
 13  foot                                  27964 non-null  object
 14  height_in_cm                          30130 non-null  float64
 15  market_value_in_eur                   30130 non-null  float64
 16  highest_market_value_in_eur           30130 non-null  float64
 17  contract_expiration_date              30130 non-null  datetime64[ns]
 18  image_url                             30130 non-null  object
 19  url                                   30130 non-null  object
 20  current_club_domestic_competition_id  30130 non-null  object
 21  current_club_name                     30130 non-null  object
 22  age                                   30130 non-null  int64
 23  remaining_contract_days               18799 non-null  float64
 24  age_group                             30130 non-null  object
 25  games_total                           30130 non-null  int64
 26  goals_total                           30130 non-null  int64
 27  assists_total                         30130 non-null  int64
 28  minutes_played_total                  30130 non-null  int64
 29  goals_for_total                       30130 non-null  int64
 30  goals_against_total                   30130 non-null  int64
 31  clean_sheet_total                     30130 non-null  int64
 32  yellow_cards_total                    30130 non-null  int64
 33  red_cards_total                       30130 non-null  int64
 34  games_2019                            30130 non-null  float64
 35  goals_2019                            30130 non-null  float64
 36  assists_2019                          30130 non-null  float64
 37  minutes_played_2019                   30130 non-null  float64
 38  goals_for_2019                        30130 non-null  float64
 39  goals_against_2019                    30130 non-null  float64
 40  clean_sheet_2019                      30130 non-null  float64
 41  yellow_cards_2019                     30130 non-null  float64
 42  red_cards_2019                        30130 non-null  float64
 43  games_2020                            30130 non-null  float64
 44  goals_2020                            30130 non-null  float64
 45  assists_2020                          30130 non-null  float64
 46  minutes_played_2020                   30130 non-null  float64
 47  goals_for_2020                        30130 non-null  float64
 48  goals_against_2020                    30130 non-null  float64
 49  clean_sheet_2020                      30130 non-null  float64
 50  yellow_cards_2020                     30130 non-null  float64
 51  red_cards_2020                        30130 non-null  float64
 52  games_2021                            30130 non-null  float64
 53  goals_2021                            30130 non-null  float64
 54  assists_2021                          30130 non-null  float64
 55  minutes_played_2021                   30130 non-null  float64
 56  goals_for_2021                        30130 non-null  float64
 57  goals_against_2021                    30130 non-null  float64
 58  clean_sheet_2021                      30130 non-null  float64
 59  yellow_cards_2021                     30130 non-null  float64
 60  red_cards_2021                        30130 non-null  float64
 61  games_2022                            30130 non-null  float64
 62  goals_2022                            30130 non-null  float64
 63  assists_2022                          30130 non-null  int32
 64  minutes_played_2022                   30130 non-null  float64
 65  goals_for_2022                        30130 non-null  float64
 66  goals_against_2022                    30130 non-null  float64
 67  clean_sheet_2022                      30130 non-null  int32
 68  yellow_cards_2022                     30130 non-null  int32
 69  red_cards_2022                        30130 non-null  int32
 70  games_2023                            30130 non-null  float64
 71  goals_2023                            30130 non-null  float64
 72  assists_2023                          30130 non-null  float64
 73  minutes_played_2023                   30130 non-null  float64
 74  goals_for_2023                        30130 non-null  float64
 75  goals_against_2023                    30130 non-null  float64
 76  clean_sheet_2023                      30130 non-null  float64
 77  yellow_cards_2023                     30130 non-null  float64
 78  red_cards_2023                        30130 non-null  float64
dtypes: datetime64[ns](1), float64(44), int32(6), int64(12), object(15)
memory usage: 18.7+ MB
```

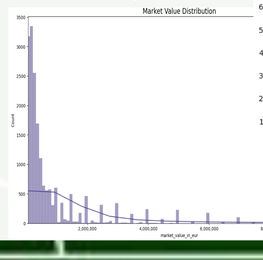
Exploratory Data Analytics (EDA)

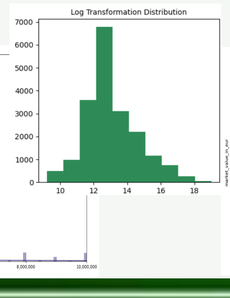
EDA is a process of examining the available data to discover patterns of correlation, identify trends, and gain insights of interest. It aids in spotting anomalies, testing hypotheses and clarifying any assumptions.

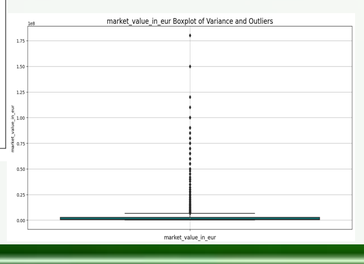
We find that the target variable data, 'Market Value', is right skewed, and requires a log transformation for normal distribution. The data also contains multiple outliers, identifiable through dots in the box plot below.

Through EDA we discover similar patterns and irregularities within the independent variables used for predicting. These findings will influence preprocessing decisions and techniques applied in Machine Learning.

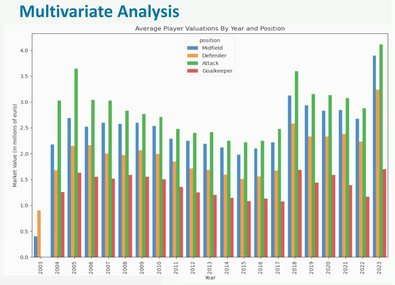
Univariate Analysis

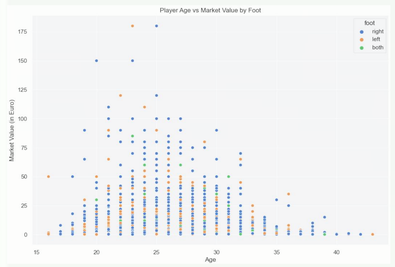






Multivariate Analysis





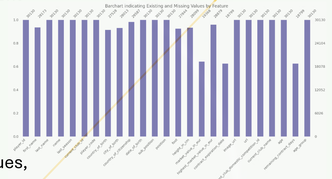
3. Data Preparation

Preprocessing

Data preparation and preprocessing is a crucial step in any Data Science project's lifecycle. It involves several operations and transformations being applied to raw data to make it suitable for analysis and modelling.

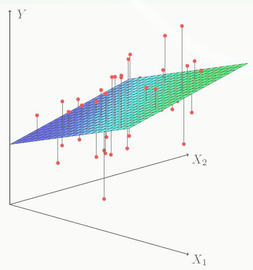
The aim is to enhance the quality of the data by:

- addressing **missing** or inconsistent values,
- handling various data types including **type casting**,
- number formatting, and **label encoding** categorical variables,
- **transforming** or scaling the data using standardisation or **normalisation** techniques,
- and conducting **feature engineering**.
- Multiple datasets are created based on different preprocessing and imputation techniques.



4. Modelling

In this phase, we build and evaluate models to address the project objectives. This involves selecting appropriate modelling techniques, applying appropriate train : test splits, building different models, and adjusting their parameters through hyperparameter tuning before cross-validating (CV) the results.

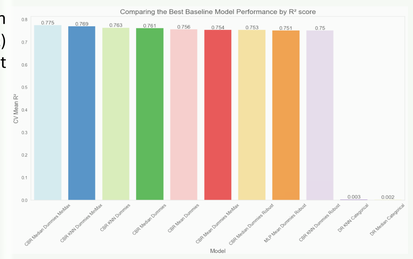


- The objective is to predict the market value of a player, which is continuous data, using multiple predictors, making this **a multiple regression problem**.
- Furthermore, we are using labelled data making this a supervised ML application.
- Simple Linear Regression is a statistical model which estimates the linear relationship between two variables.
- Multiple Regression is the analysis between the relationships between a single dependent variable and several independent variables.

Over 25 different baseline models are tested on default parameters, including Kernal Ridge (KR) and Multi-Layer Perceptrons with the most promising selected for optimisation.

Models selected for optimisation:

- Kernel Ridge Regressor (KN)
- Random Forest Regressor (RFR)
- Gradient Boosting Regressor (GBR)
- Categorical Boost Regressor (CBR)
- Light Gradient Boosting Machine (LGBM)



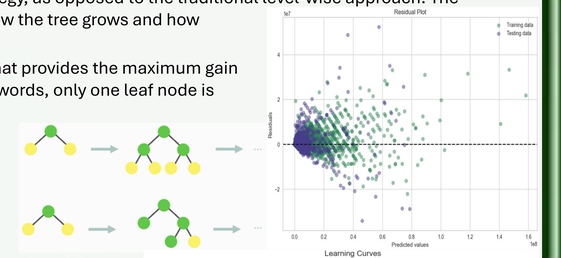
The best model developed for predicting a football player's market value with the available data is LightGBM. LightGBM is a gradient-boosting framework that uses tree-based learning algorithms.

It adopts a leaf-wise growth strategy, as opposed to the traditional level-wise approach. The fundamental difference lies in how the tree grows and how branches expand.

The tree grows by adding a leaf that provides the maximum gain at each expansion step. In other words, only one leaf node is added at each expansion step.

Traditional Models

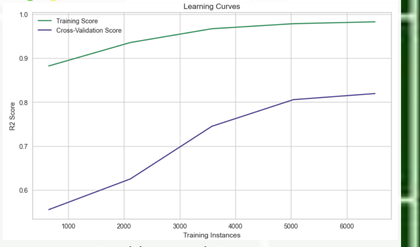
Light GBM



5. Evaluation

Evaluating The Model

- From the residuals plot we find that the model appears to predict well for lower market values, densely located around the origin.
- Promisingly, there appears to be a normal distribution of error predictions about the origin line.
- The learning curves indicate an overfitting model, determinable by the lower test scores (purple line) relative to the high training scores.
- Promisingly, the learning curves suggest further improvements with more data.

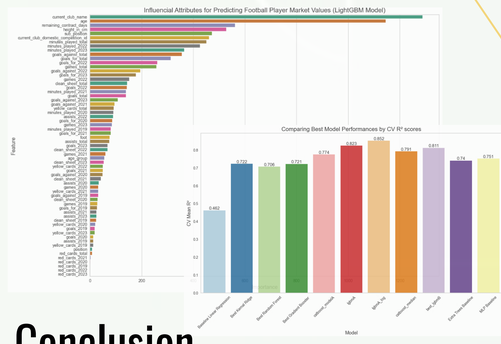


Evaluating Results

After building models, they need to be evaluated to determine their performance and effectiveness.

For regression problems, Sklean provides multiple metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) for more advanced metrics, with R^2 being more intuitive, and the preferred choice.

The model achieves a CV Mean R^2 score of 0.823 which increases to **0.852** when optimised and trained on the log transformation of the target variable.

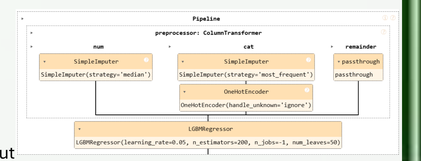


- In comparison to the baseline LR model, we have achieved an **84.5% improvement.**

6. Deployment

The Deployment phase within the CRISP-DM methodology involves implementing the solutions derived from modelling into practical use.

This includes integrating them with existing systems, providing documentation, and offering support to end-users to derive business value. Although perhaps premature, a pipeline of the model has been generated.



Conclusion

Player scouting and market value analysis are key areas where ML can provide valuable knowledge and insights to the football industry. This capstone project aims to contribute to this, by empowering clubs and stakeholders with data-driven decision-making capabilities, for a competitive advantage.

At the end of the process, the results are positive but the project still provides plenty of scope for improvement.

The exact degree of success may only be determined by domain experts and per application.

Technologies...

Jupyter Notebook, scikit learn, NumPy, seaborn, PyCaret, pandas, matplotlib, GitHub, SHAP

References...

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining . Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, (Vol. 1, pp. 29-39).
Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Beijing: O'reilly.
Technology, T. (2023). Light GBM Light and Powerful Gradient Boost Algorithm. [online] Medium.
Available at: https://medium.com/@turkishtechnology/light-gbm-light-and-powerful-gradient-boost-algorithm-eaa1e804eca8.

SCAN ME