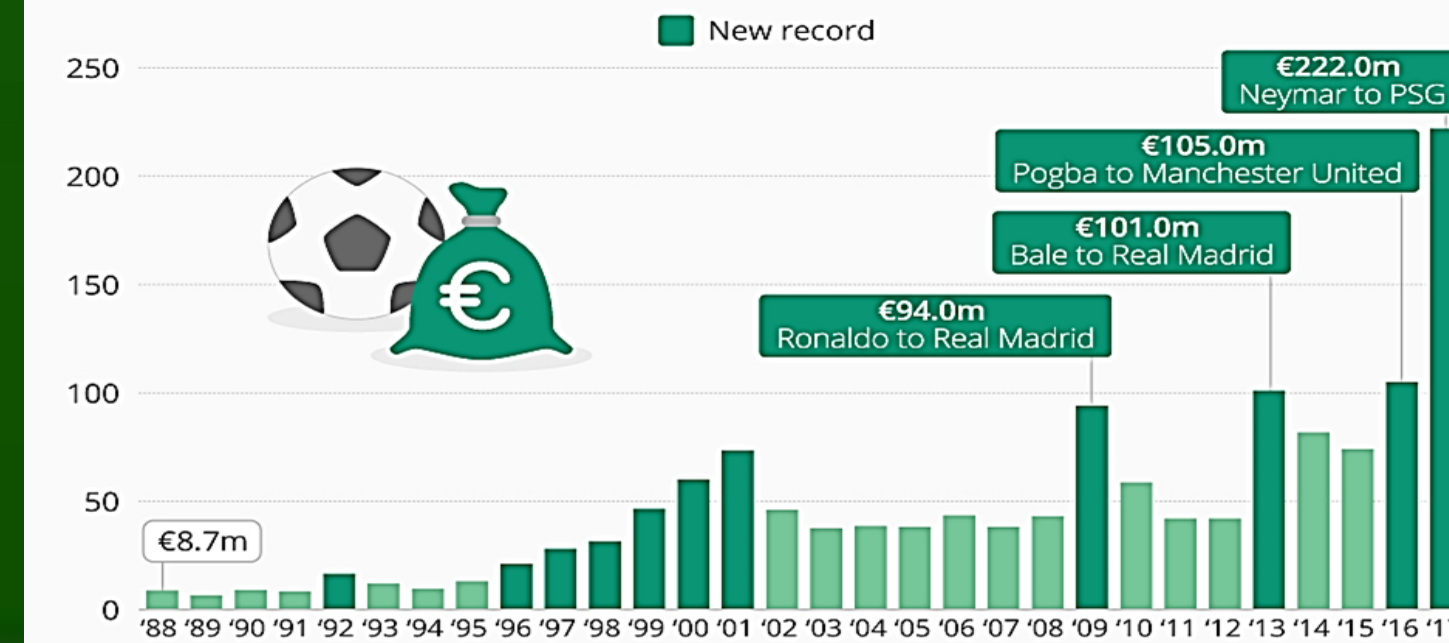


# Leveraging Machine Learning and Data Science for Competitive Advantage: Estimating Football Player Market Values

Author: Kavi Patak Supervisor: James Garza



## 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/davidc-ariboo/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)

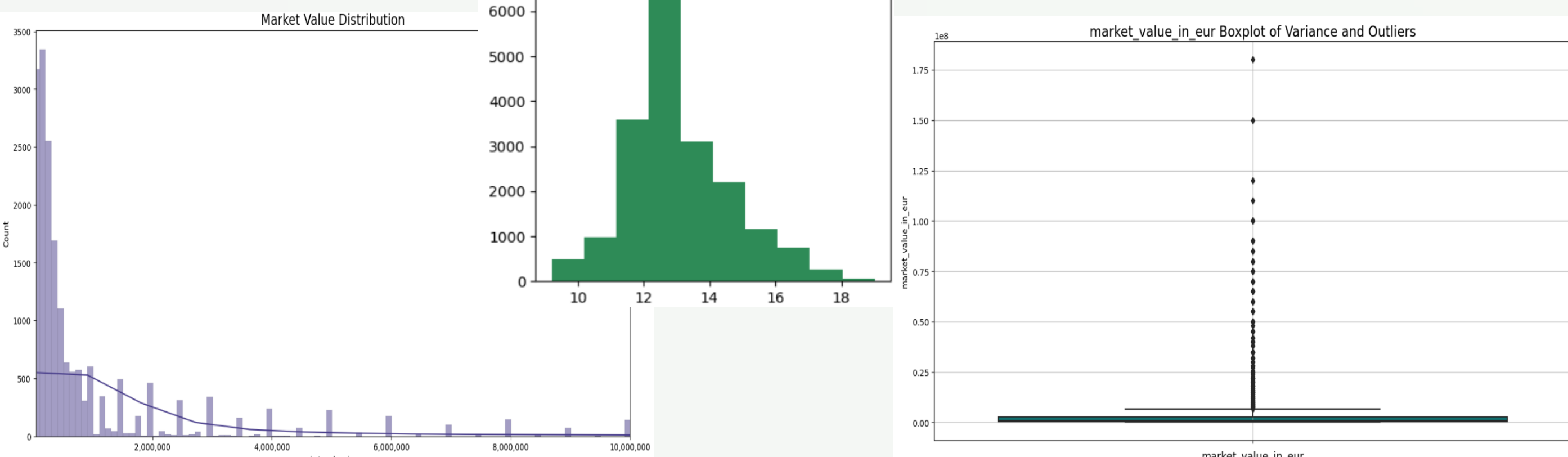
#### Exploratory Data Analytics (EDA)

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.

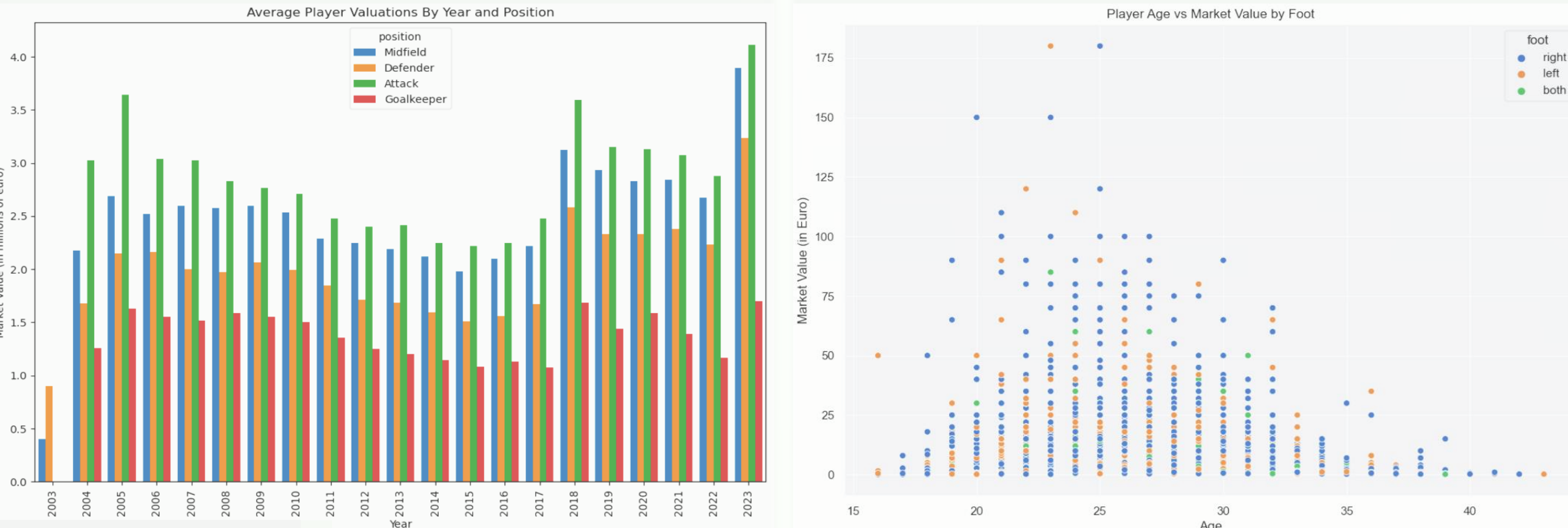
We find that the target variable data, 'Market Value', is right skewed, and requires a log transformation for normal distribution. The data also contains multiple outliers, identifiable through dots in the box plot below.

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.

#### Bivariate Analysis



#### Multivariate Analysis



### 3. Data Preparation

#### Preprocessing

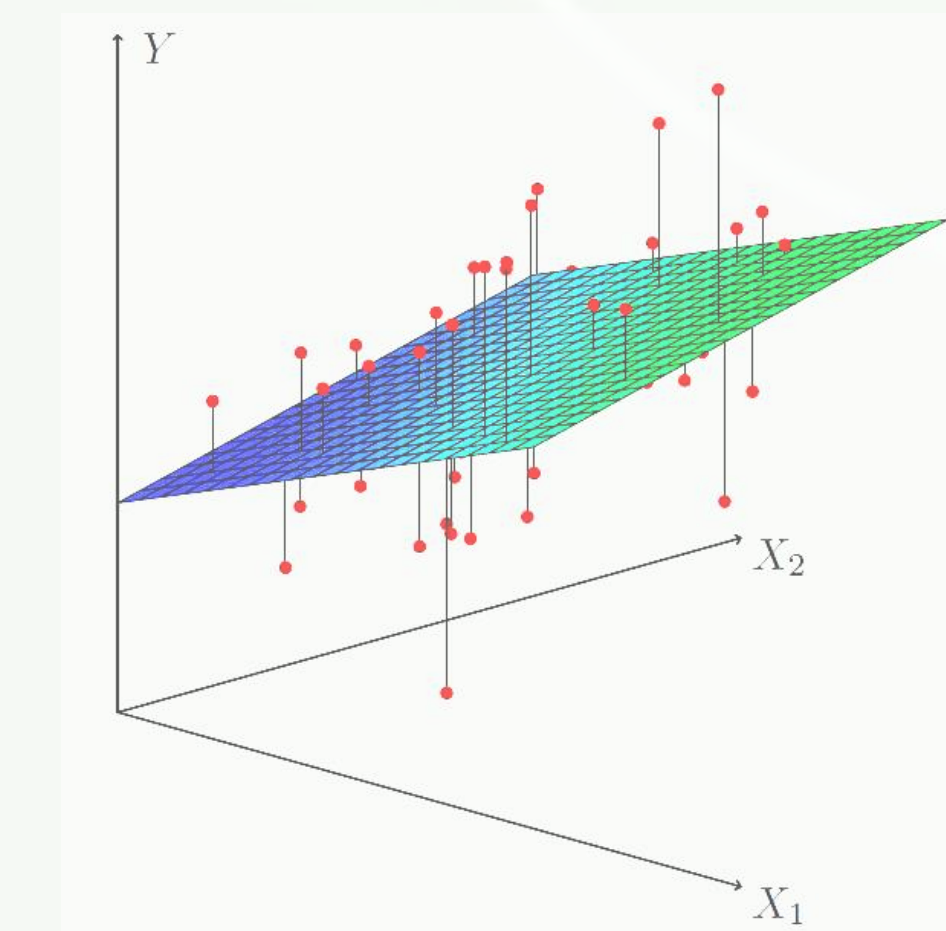
Data preparation and preprocessing is a crucial step in any Data Science project's lifecycle. It involves several operations and transformations being applied to raw data to make it suitable for analysis and modelling.

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.

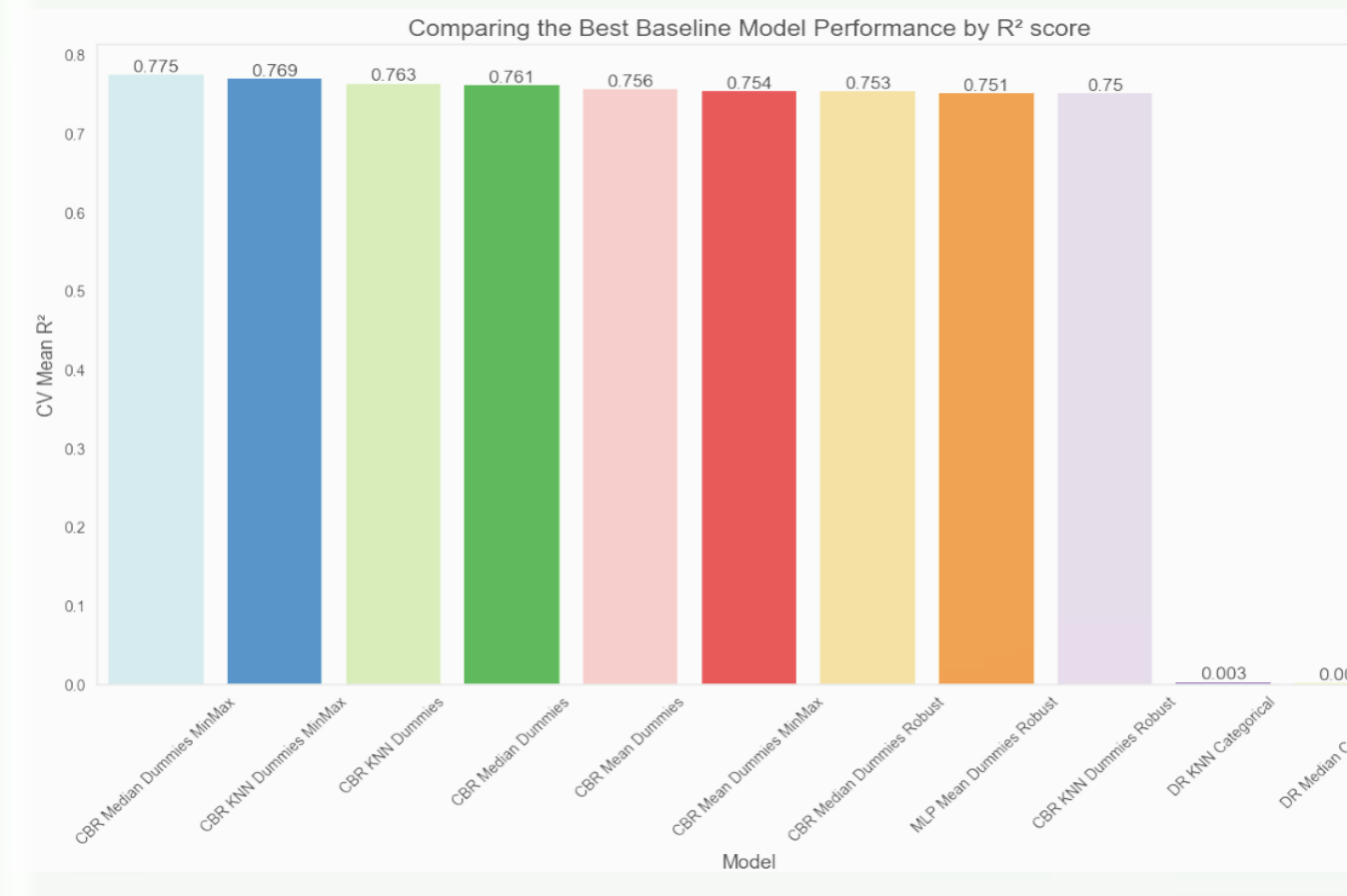


- The objective is to predict the market value of a player, which is continuous data, using multiple predictors, making this a multiple regression problem.
- Furthermore, we are using labelled data making this a supervised ML application.
- Simple Linear Regression is a statistical model which estimates 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 baseline models are tested on default parameters with the most promising selected for optimisation.

#### Models selected for optimisation:

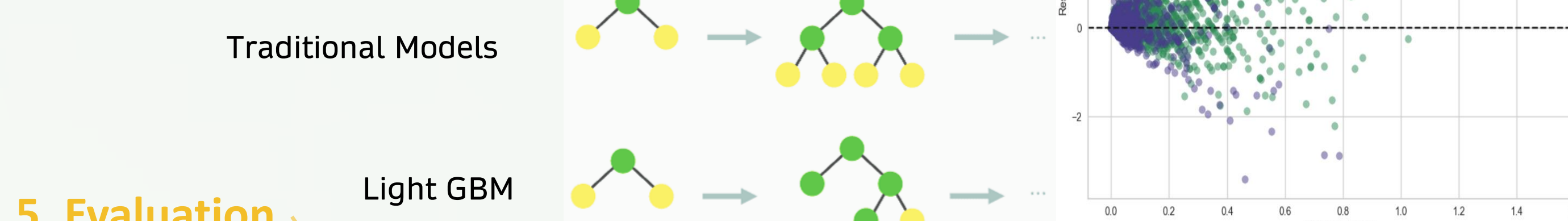
- Kernel Ridge Regressor (KN)
- Random Forest Regressor (RFR)
- Gradient Boosting Regressor (GBR)
- Categorical Boost Regressor (CBR)
- Light Gradient Boosting Machine (LGBM)



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.

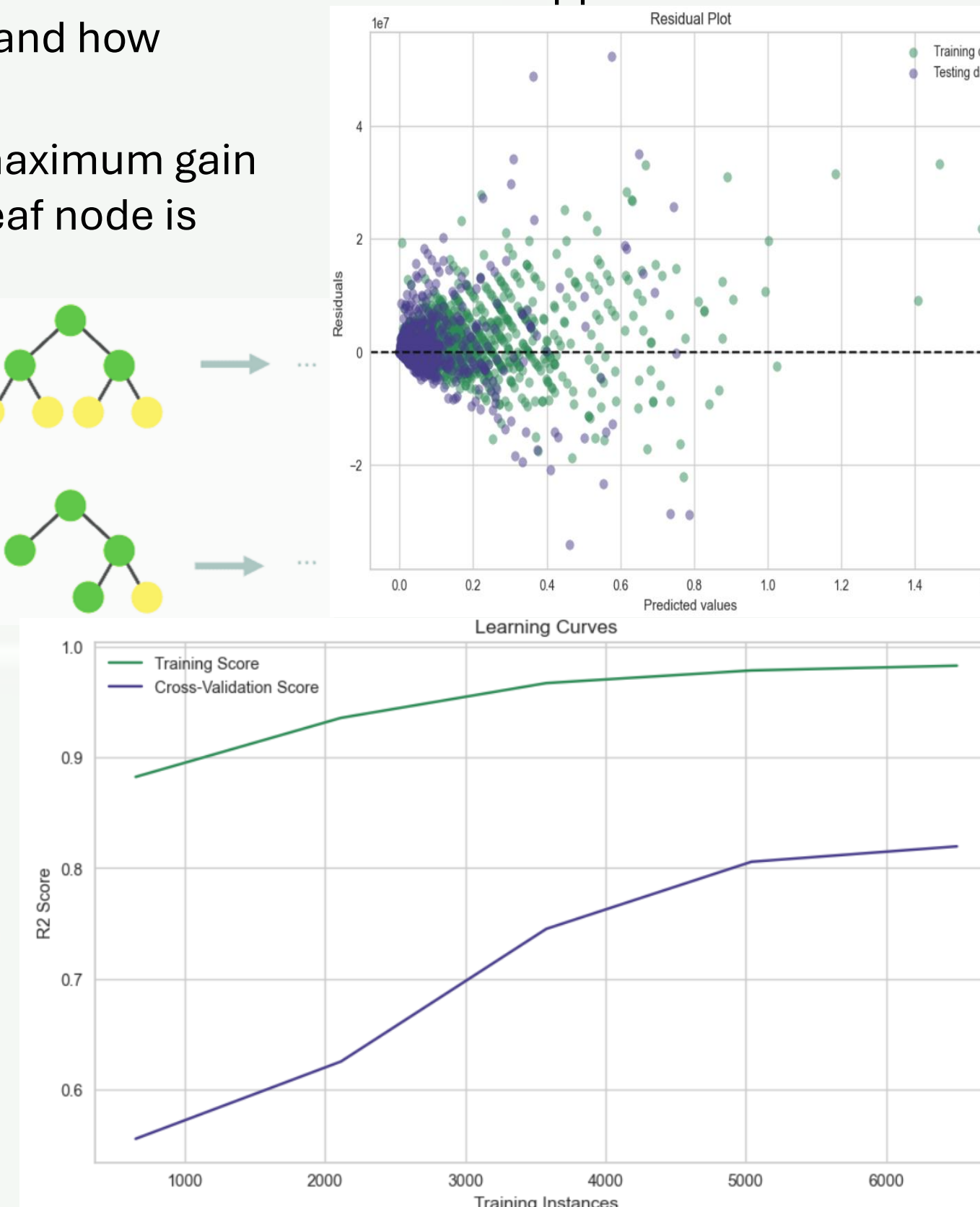
The tree grows by adding a leaf that provides the maximum gain at each expansion step. In other words, only one leaf node is added at each expansion step.



### 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.

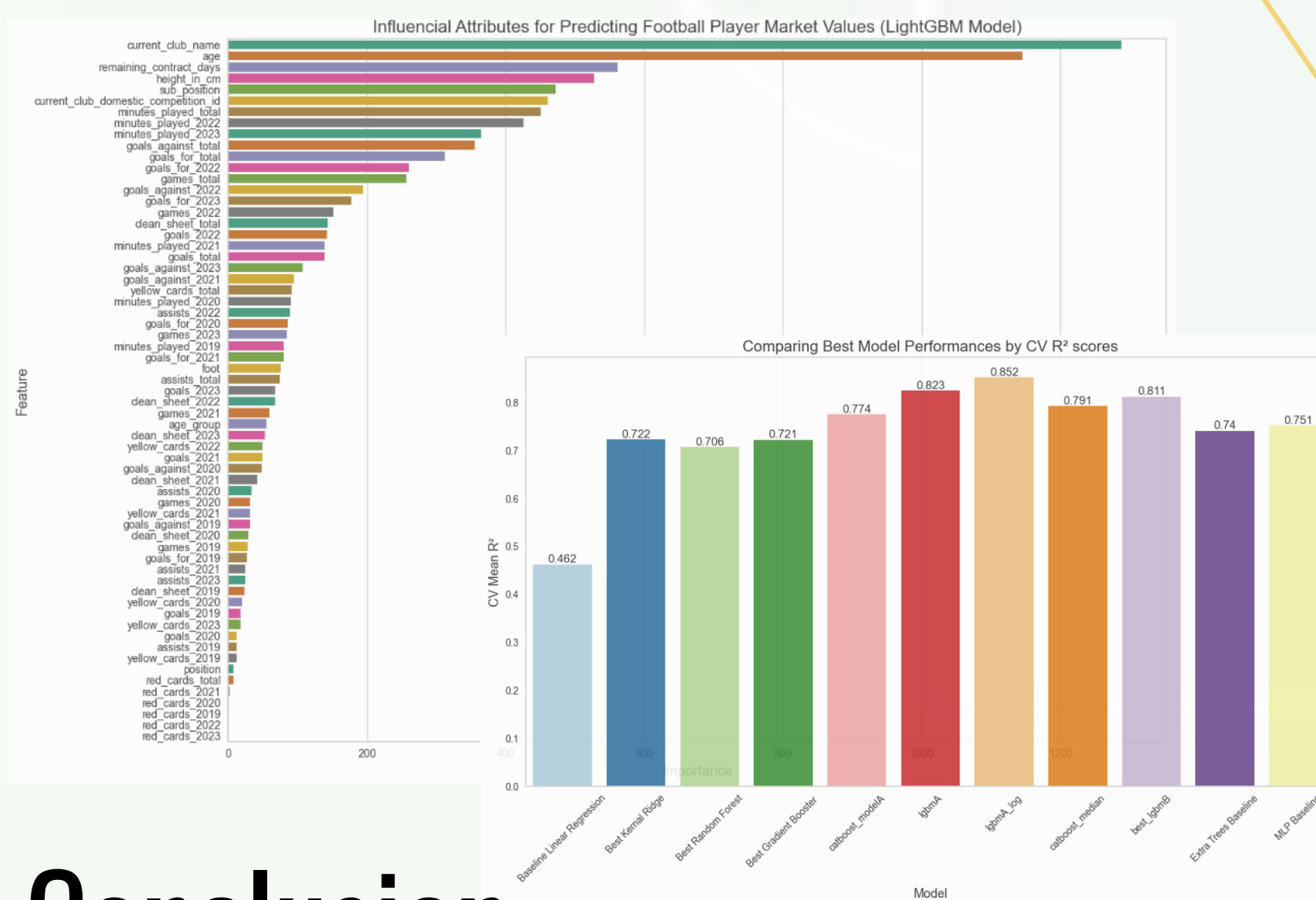


#### Evaluating Results

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.

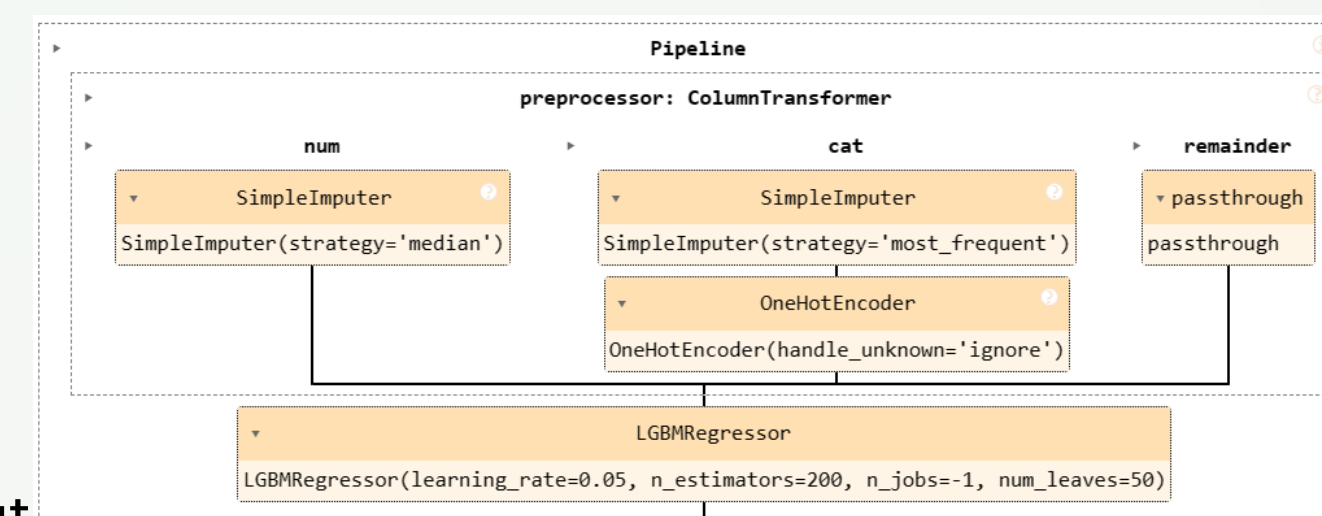
The model achieves a CV Mean  $R^2$  score of 0.823 which increases to **0.852** when optimised and trained on the log transformation of the target variable.



- 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.



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.

## 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/@turkishtechology/light-gbm-light-and-powerful-gradient-boost-algorithm-eaa1e804eca8>.

