



Advanced Sports Analytics

Erasmus+ Blended Intensive Program (BIP)

Analysis of Euro Football Data

Faculty of Statistics

Master's Degree Program in Data Science

Supervisor:

Prof. Dr. Andreas Groll
Technical University of Dortmund

Student:

Bashirul Alam
Matriculation No.: 265109

Academic Year: 2024/2025

Contents

Introduction	2
Explanatory Data Analysis	3
Description of Dataset	3
Data Distribution	4
Relationship between goal difference and market value	5
Relationship between goal difference and CL players	6
Models	7
Poisson model	7
Regression Tree	10
conditional inference model	11
Conclusion	13

Introduction

This report investigates the factors influencing team performance in the UEFA European Championship, with a focus on predicting the number of goals scored by each team in a match and translating these predictions into match outcome probabilities. Using data from several editions of the tournament, we apply both statistical and machine learning models to capture patterns in team performance.

We begin by preparing and exploring the Euro dataset to ensure data quality and consistency across tournaments. Our first modeling approach uses **Poisson regression**, which is well-suited for handling count data such as goals. To account for potential nonlinear relationships and interactions between predictors, we extend the analysis with **regression trees** and **conditional inference forests**.

Once expected goals for home and away teams are predicted, we employ the **Skellam distribution** to derive probabilities for win, draw, and loss outcomes. Model performance is then evaluated using the **Rank Probability Score (RPS)**, which allows for a fair comparison between competing approaches.

This analysis not only identifies key variables—such as GDP, market value, FIFA ranking, UEFA coefficients, and Champions League experience—that contribute to football outcomes, but also compares the predictive strength of different modeling techniques. The ultimate aim is to highlight both the interpretability of statistical models and the flexibility of machine learning methods in understanding and forecasting football results.

Explanatory Data Analysis

Description of Dataset

The UEFA Euro data are separated into a training set and a test set. The training data cover the tournaments from 2008 to 2020, while the test data contain matches from the 2024 European Championship.

Each observation corresponds to a single match, with the outcome represented by the goal difference between the two teams.

The dataset includes the following predictors:

- **GroupStage**: Indicator variable (1 if the match was played in the group stage, 0 if in the knockout stage).
- **GDP**: Gross Domestic Product per capita of the respective country.
- **MarketValue**: Cumulative market value of the squad.
- **FifaRank**: FIFA ranking of the team before the tournament.
- **UefaPoints**: UEFA coefficient points of the team before the tournament.
- **CLPlayers**: Number of players in the squad who participated in the UEFA Champions League semifinal prior to the tournament.

##	Goals	Team	Opponent	Year	GroupStage	GDP	MarketValue
## 1	0	Switzerland	Czech Republic	2008	1	1.1550862	-0.3133334
## 2	1	Czech Republic	Switzerland	2008	1	-1.1550862	0.3133334
## 3	2	Portugal	Turkey	2008	1	0.8342002	0.9367693
## 4	0	Turkey	Portugal	2008	1	-0.8342002	-0.9367693
## 5	1	Czech Republic	Portugal	2008	1	-0.1346162	-0.8883897
## 6	3	Portugal	Czech Republic	2008	1	0.1346162	0.8883897
## FifaRank UefaPoints CLPlayers							
## 1	42	-10700	-1				
## 2	-42	10700	1				
## 3	-16	10042	5				
## 4	16	-10042	-5				
## 5	-3	-9258	-4				
## 6	3	9258	4				

Data Distribution

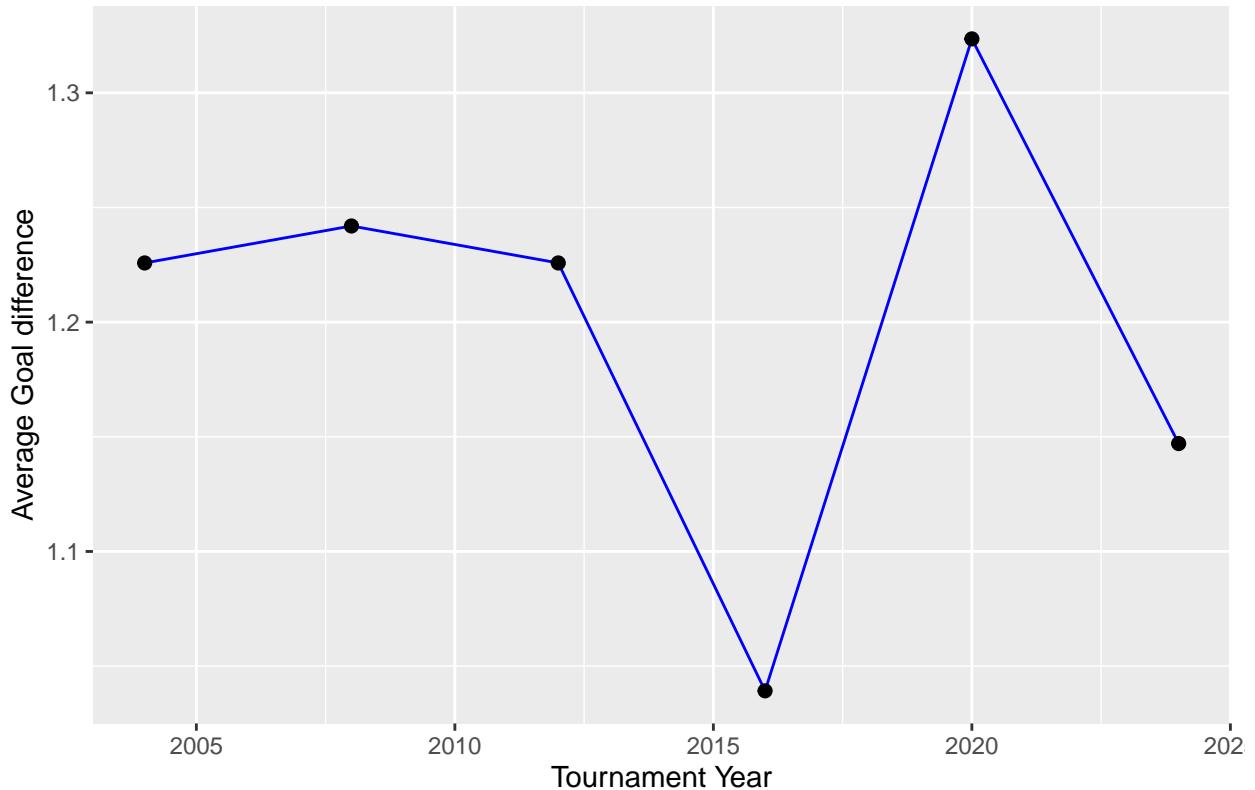
In the data distribution, we first see the distribution of goals difference in different tournaments.

Table 1: Average Goal Difference per Tournament

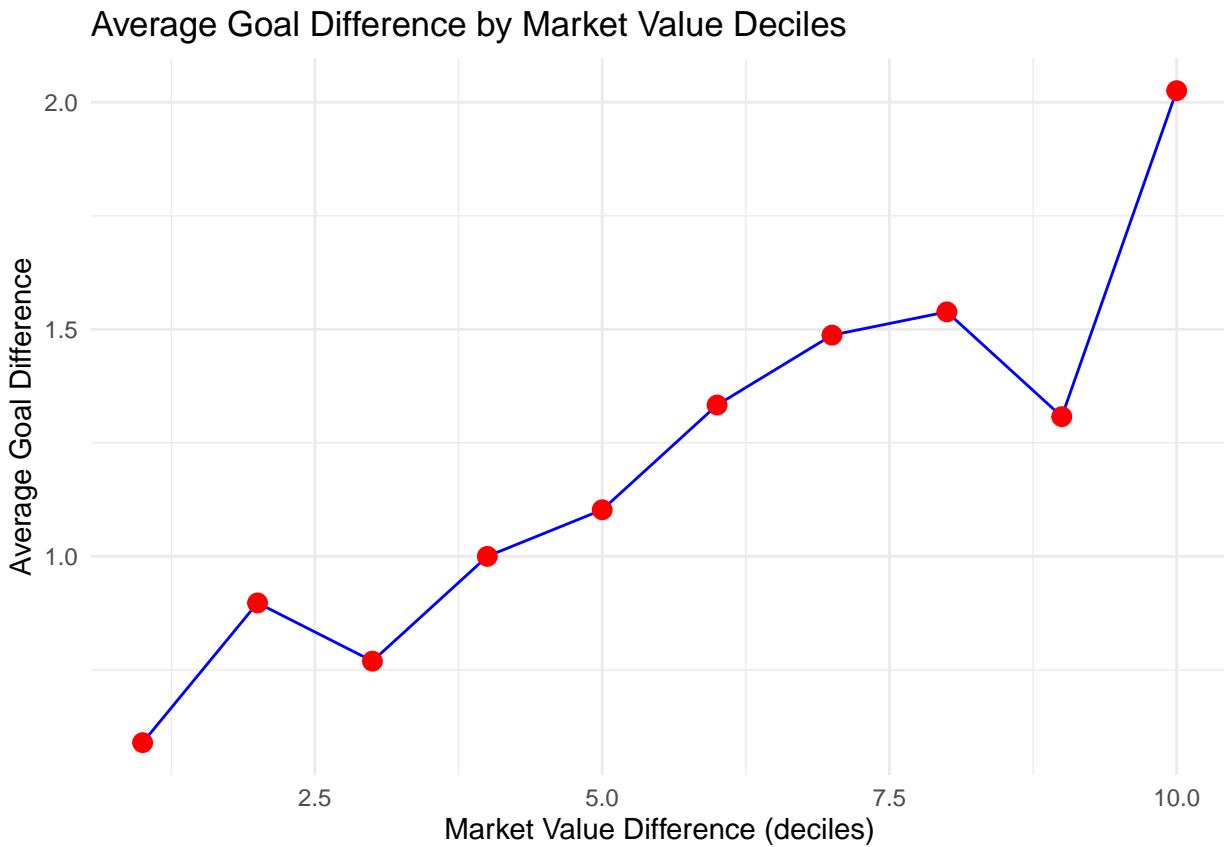
Year	avg_goals
2004	1.225807
2008	1.241936
2012	1.225807
2016	1.039216
2020	1.323529
2024	1.147059

The average goal difference across tournaments stays fairly stable (1.1–1.3), reflecting the overall competitiveness of the Euros. Euro 2016 had the lowest average (1.04), while Euro 2020 peaked at 1.32, with Euro 2024 returning closer to the long-term mean.

Average Goal Difference per Tournament

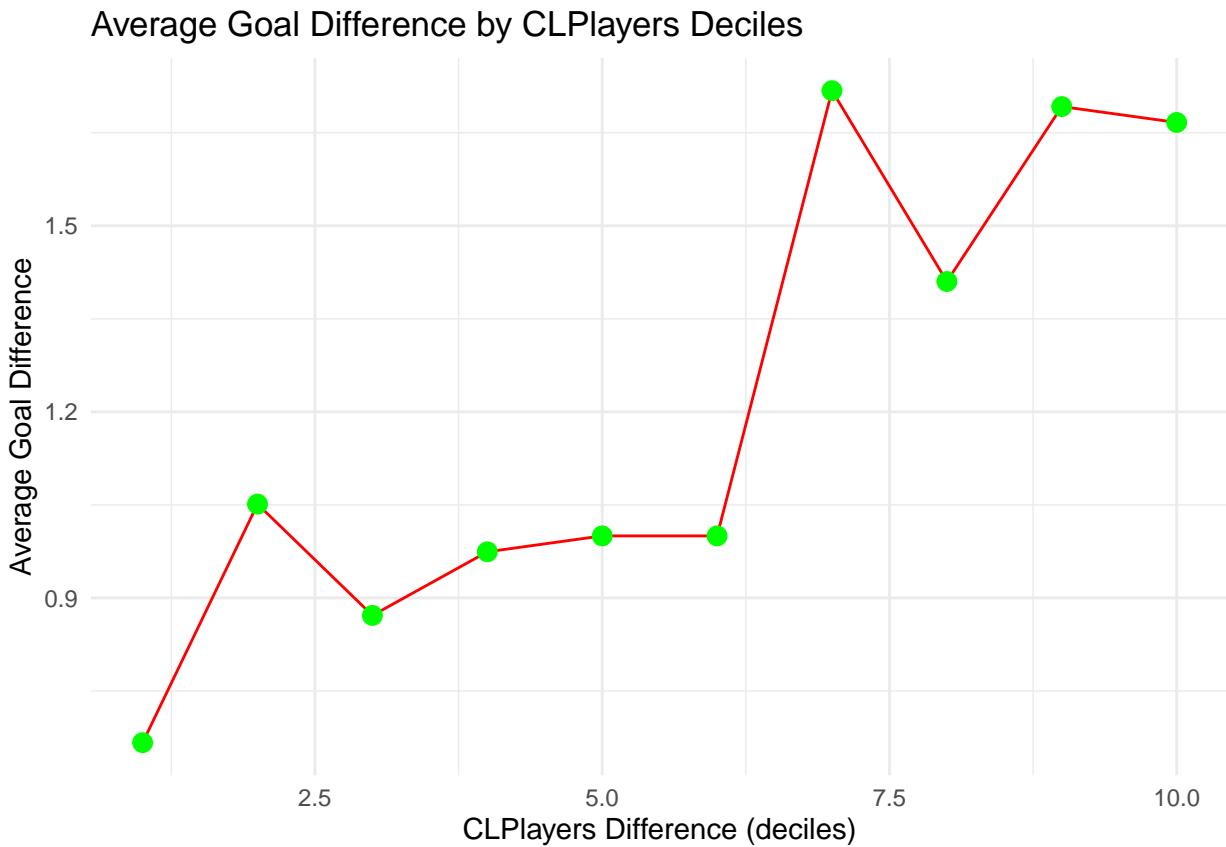


Relationship between goal difference and market value



The line graph shows that teams with higher market value differences generally achieve larger positive goal differences, indicating that financially stronger squads tend to outperform their opponents. However, the trend is not perfectly linear, suggesting that while market value matters, it is not the only factor determining match outcomes.

Relationship between goal difference and CL players



The plot shows that teams with more Champions League-experienced players generally achieve larger goal differences. This effect becomes especially strong from the 6th decile onward, where the advantage translates into significantly higher winning margins.

Models

Poisson model

The Poisson regression model is a classical statistical approach for modeling count data, where the outcome variable represents the number of occurrences of an event within a fixed unit of time, space, or context. In our case, the event of interest is the goal difference in UEFA Euro matches. This model belongs to the family of generalized linear models (GLMs) and assumes that the response variable follows a Poisson distribution.

The defining feature of the Poisson model is its ability to relate the expected count of events to a set of explanatory variables through a logarithmic link function. This ensures that predicted values remain non-negative, as is required for count data. The model is widely used in fields such as epidemiology, economics, and sports analytics, particularly when outcomes are discrete and non-negative.

For match i with predictors $x_{i1}, x_{i2}, \dots, x_{ip}$:

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i > 0$$

The expected goal difference λ_i is linked to the predictors through the log link:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

where:

- Y_i is the observed goal difference in match i ,
- $\lambda_i = \mathbb{E}[Y_i]$ is the expected goal difference,
- $\beta_0, \beta_1, \dots, \beta_p$ are the model coefficients to be estimated.

We fit the poisson model with all the features .

```
##  
## Call:  
## glm(formula = Goals ~ GDP + MarketValue + FifaRank + UefaPoints +  
##       CLPlayers, family = poisson(link = "log"), data = train)  
##  
## Coefficients:  
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 1.244e-01 4.898e-02 2.539   0.0111 *  
## GDP         1.003e-01 4.413e-02 2.273   0.0231 *  
## MarketValue 1.734e-01 8.323e-02 2.083   0.0372 *  
## FifaRank    -2.491e-03 2.977e-03 -0.837   0.4028  
## UefaPoints  -7.803e-07 2.337e-06 -0.334   0.7385  
## CLPlayers   2.716e-02 1.687e-02 1.610   0.1074  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for poisson family taken to be 1)  
##  
## Null deviance: 470.63  on 389  degrees of freedom  
## Residual deviance: 413.11  on 384  degrees of freedom  
## AIC: 1066.8  
##  
## Number of Fisher Scoring iterations: 5
```

Here we found that, GDP and Market Value are significant positive predictors of goals—higher values are associated with more expected goals. FIFA Rank and UEFA coefficient points are not significant, and Champions League players shows only a weak, non-significant effect. The model reduces deviance from 470.6 to 413.1 (12% explained) with AIC 1066.8, indicating a modest fit typical for goal-count data. Overall, financial strength appears more informative than ranking metrics in this specification.

Now, we predict train and test data using the fitted model. To see how this model perform in train and test data.

Dataset	MAE
Training Data	0.838
Testing Data	0.732

Here, we found that our **training** error is larger than the **testing** error, which is a bit unusual. We will drop the insignificant features, refit the model, and then calculate the training and testing errors for the new model.

```
##  
## Call:  
## glm(formula = Goals ~ GDP + MarketValue, family = poisson(link = "log"),  
##       data = train)  
##  
## Coefficients:  
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept)  0.12813   0.04879  2.626  0.00863 **  
## GDP         0.09914   0.04351  2.279  0.02269 *  
## MarketValue 0.24730   0.04151  5.957 2.56e-09 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for poisson family taken to be 1)  
##  
## Null deviance: 470.63  on 389  degrees of freedom  
## Residual deviance: 416.22  on 387  degrees of freedom  
## AIC: 1063.9  
##  
## Number of Fisher Scoring iterations: 5
```

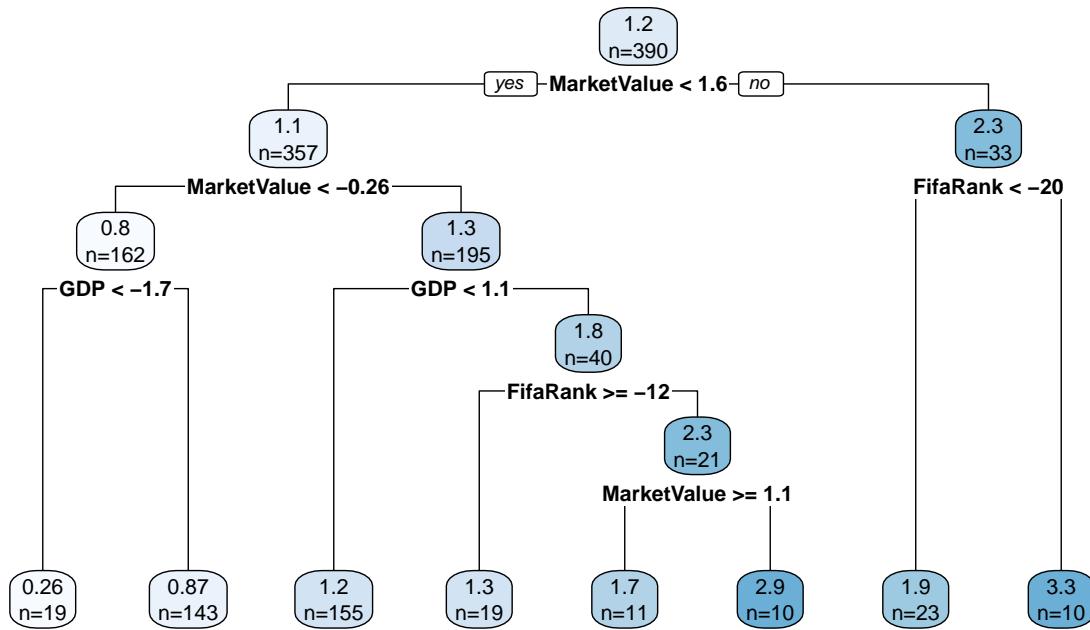
Our new model perform a little better.

Dataset	MAE
Training Data	0.841
Testing Data	0.731

Our testing error is still lower than the training error. Now, we will fit regression tree to find out it it perform better than poisson model.

Regression Tree

Regression Tree for Euro Goals



We find out, the MarketValue is the most important feature, indicating it contributes the

most to the model. FifaRank is the next most influential, while UEFA Points, CLPlayers, and GDP provide smaller incremental value.

Table 4: Importance of the features

	x
MarketValue	84.983628
FifaRank	45.240977
UefaPoints	26.910075
CLPlayers	25.569565
GDP	25.174972
GroupStage	8.864792

Now we will use the regression tree to predict for the train and test data set. Then we will calculate the train and test error.

Dataset	MAE	RMSE
Training Data	0.779	0.971
Testing Data	0.784	1.055

conditional inference model

Now, we will train a conditional inference model to our data set.

Table 6: Importance of the features

	x
GroupStage	-0.0030628
GDP	0.0262250
MarketValue	0.1480652
FifaRank	0.0186563
UefaPoints	0.0101267
CLPlayers	0.0330327

The conditional inference model flags MarketValue as the dominant driver, with smaller but meaningful contributions from CLPlayers and GDP. FIFA Rank and UEFA Points add limited signal, while GroupStage shows slightly negative importance, suggesting no real predictive value.

Now , we will combine the training ans testing data set. And run k fold cross validation on both regression tree model and conditional inference model. Then we will compare the model performance using rank probability score.

Table 7: Rank Probability Score

Model	RPS
Cforest	0.1926733
Tree	0.2011447

Lower RPS is better, so the cforest model (0.193) outperforms the single tree (0.201) on probabilistic accuracy for the ordered outcomes. The gap is modest (~ 0.0085 , 4% improvement), suggesting more stable, better-calibrated forecasts from cforest.

Conclusion

This report set out to explain and predict team performance at the UEFA European Championship by modeling goals and translating them into match-outcome probabilities. Exploratory analysis showed a stable average goal-difference across tournaments and clear associations between financial strength (market value) and performance, with Champions League experience adding signal at higher deciles.

Across models, financial variables dominated: in Poisson regression, **GDP** and **Market Value** were the only consistently significant predictors of goals, and a reduced model with just those two achieved a lower AIC with only a small loss in deviance fit. Regression trees and conditional inference forests confirmed **Market Value** as the most influential feature.

In out-of-sample evaluation, errors for tree models were similar between train and test, indicating reasonable generalization, while rank probability score favored the **cforest** over a single tree (0.193 vs 0.201), suggesting better calibrated probabilistic forecasts. Overall, financial depth appears more informative than ranking-based metrics for explaining goal outcomes, though there remains headroom for improvement via richer match context (home/away, opponent strength), variance-robust count models, and hierarchical structures to capture team effects in future work.