DETERMINING SOCCER PLAYER VALUE

Team 2: Abhinav Garg, Isaac Margulies, Jessie Gondo, John Salloum, Lu Li,
Na Chuan Huang, Rachita Iyengar, Xiaomeng Huang

Neymar €50m

Pogba €---

Coutinho €9m

*All numbers as of December 3rd, 2023

AGENDA

1

Problem

2

Dataset and Features

3

Testing Models

4

Conclusion

The Problem

Optimizing soccer player valuation through statistical analysis to enhance decision-making in the transfer market.

WHY?

- Negotiation Leverage
- Informed decisions
- Financial Considerations
- Financial Fair Play & Transparency

HOW?

- Use player statistics and find importances from their value for that year
- With the importances we then predict values(€) for new years

This model will help us predict player value for future seasons given their demographic information and game statistics.

Dataset Overview

Transfer Market Value:

- Dataset from Kaggle
- Our dataset includes players from the top 5 leagues of world soccer

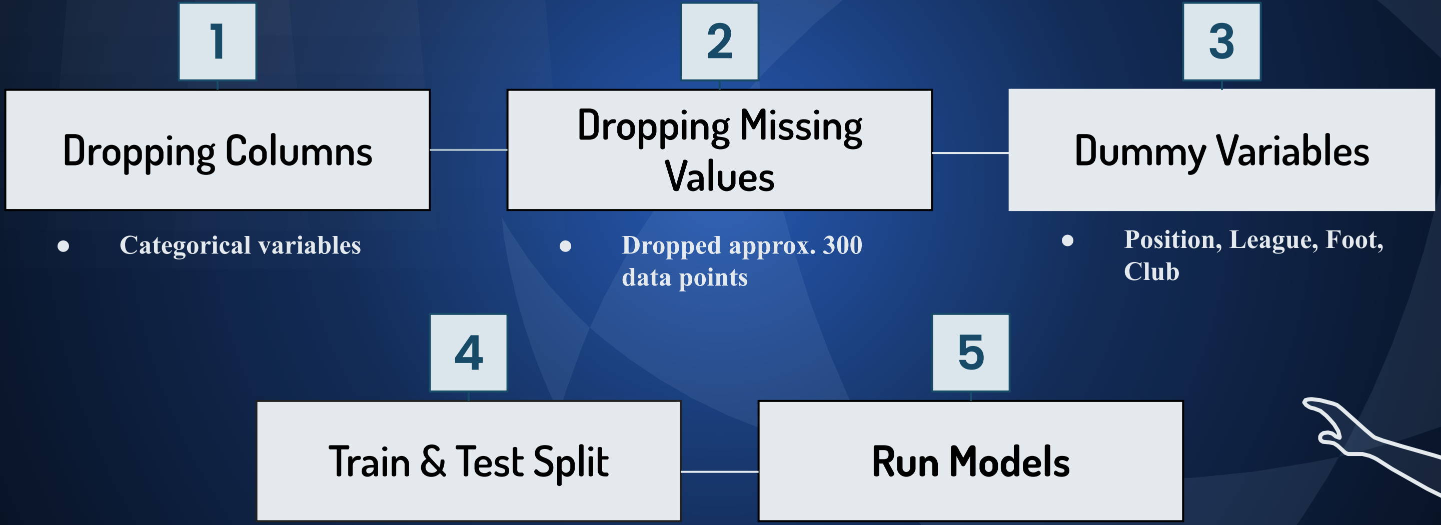
Value measured in Euros

- Outcome variable: Transfer Market player value for 2019-2020 season

How We Use it:

- The dataset included a wide variety of different statistics of a player for 2019-2020 season
- Using that, we would make predictions on their associated value

Dataset Pre-processing



1 Dropping Columns

- Categorical variables

2 Dropping Missing Values

- Dropped approx. 300 data points

3 Dummy Variables

- Position, League, Foot, Club

4 Train & Test Split

- 60% Train
- 40% Test

5 Run Models

Our Predictors

Numerical

Total stats for the entire season

- Total goals
- Assists
- Touches
- Passes, etc,

Informative

Information about plays

- Touches in opposition halves
- Amount of miscontrols

Percentage

Ratios per 90 min of gametime

- Goals/90
- Touches/90

Categorical

Descriptive information about players

- League
- Club
- Position
- Foot

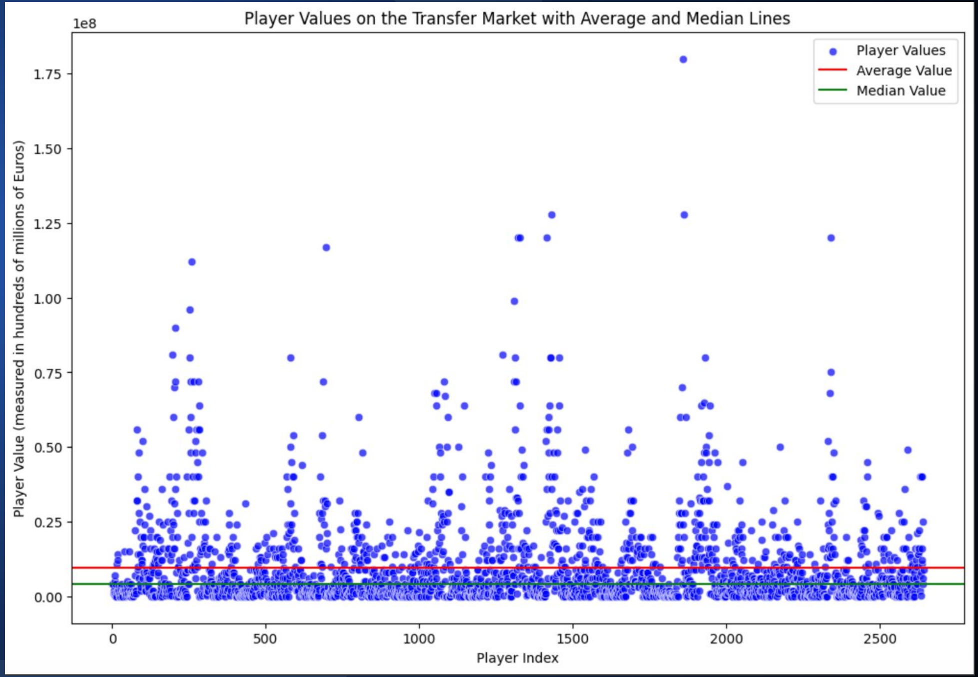
Descriptive Analytics

€9,570,623

Mean TransferMarkt Value

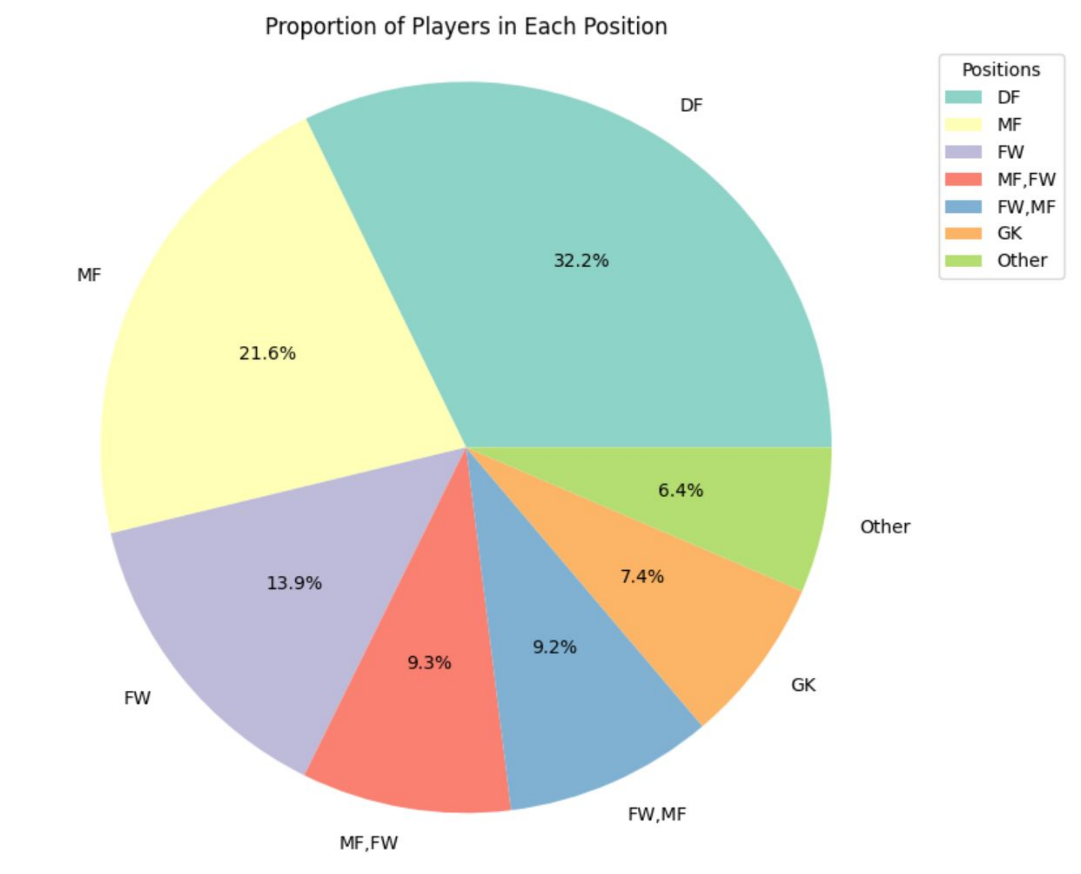
€4,000,000

Median TransferMarkt Value

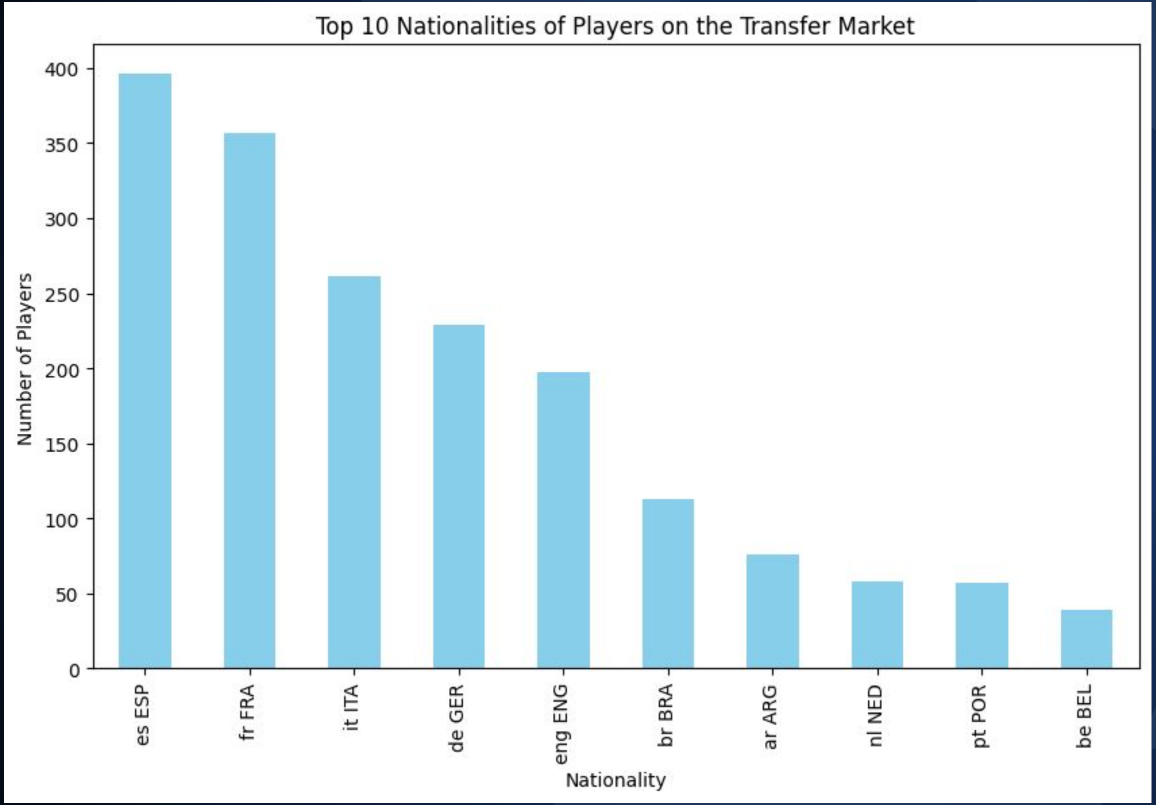


Players by Position

Majority of players fall within 6 main positions



Top 10 Nationalities in Dataset



Although there are 200+ nationalities in the dataset, over 500 are in these 10 countries alone.

Models

- Linear Regression
- k-NN
- Naive Model
- Decision Tree
- Random Forest
- Ridge & Lasso Regression
- Boosting Model

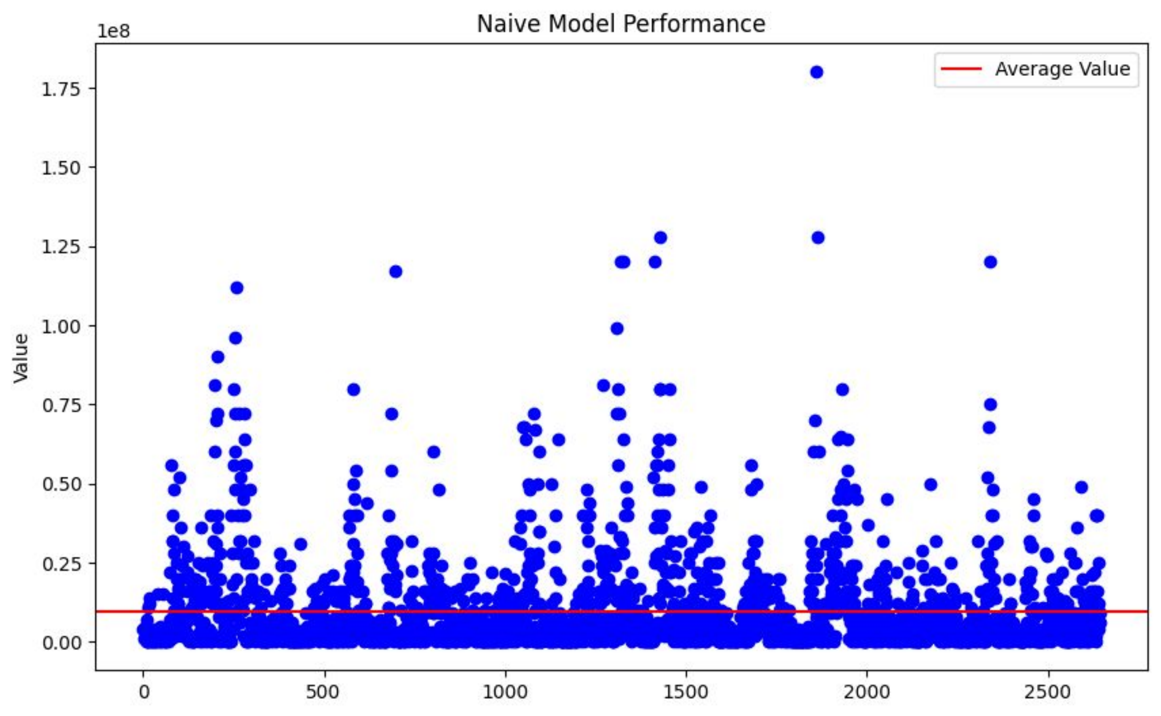
Naive Model

€9,570,623

average player value

€15,442,749

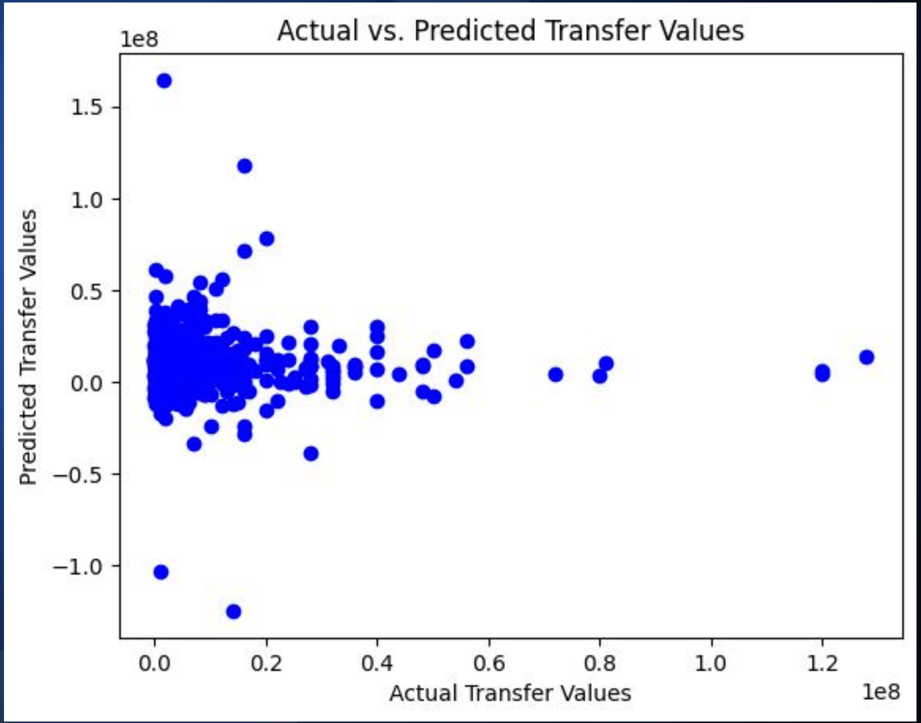
rMSE



Linear Regression

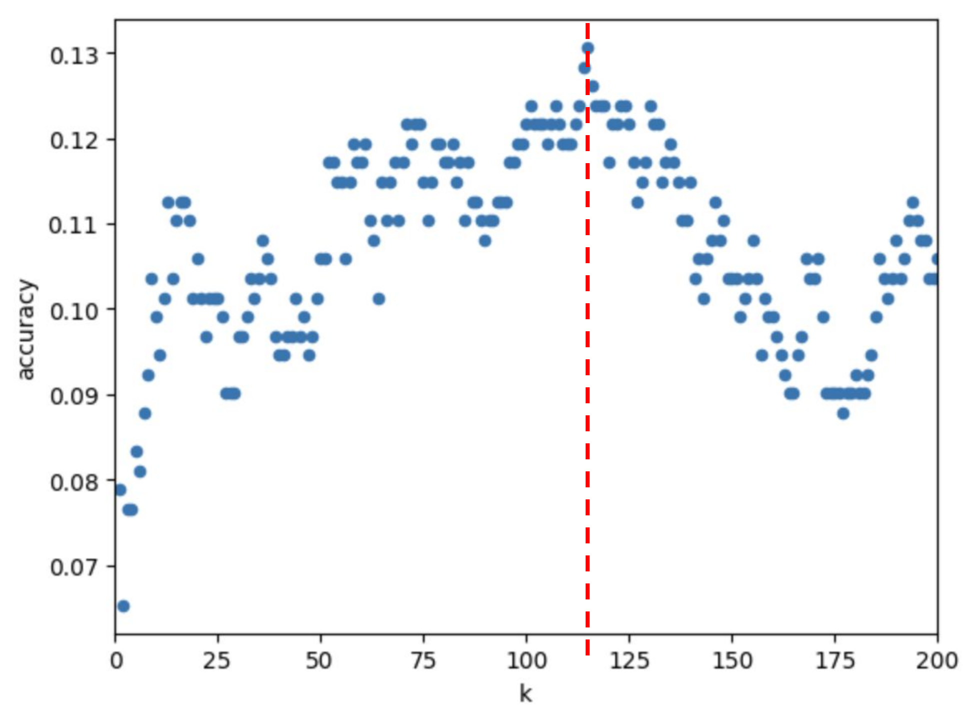
€16,586,771

rMSE



k-NN

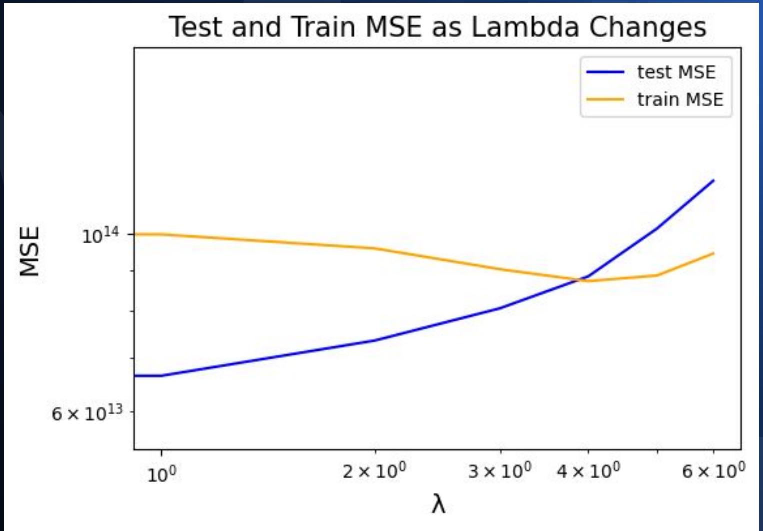
- Iterate from k = 1 to k = 200
- Optimal k value = 56
- RMSE: €18,860,892



Ridge & Lasso

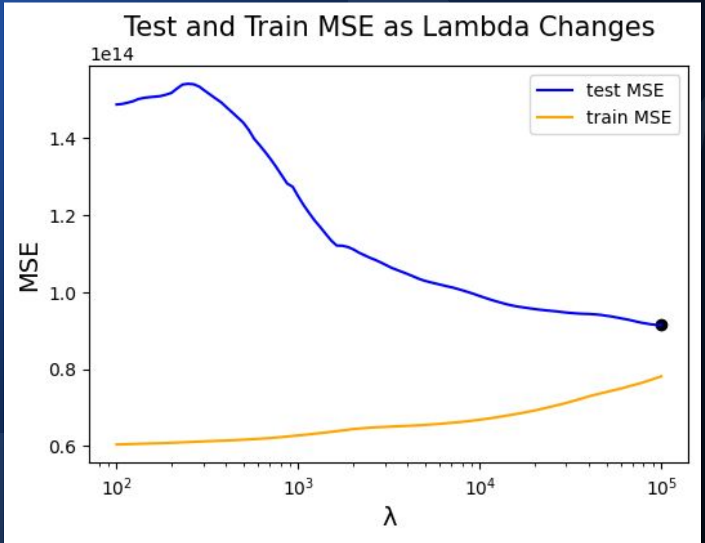
Ridge Regression

rMSE: €10,453,073



Lasso Regression

rMSE: €12,010,986



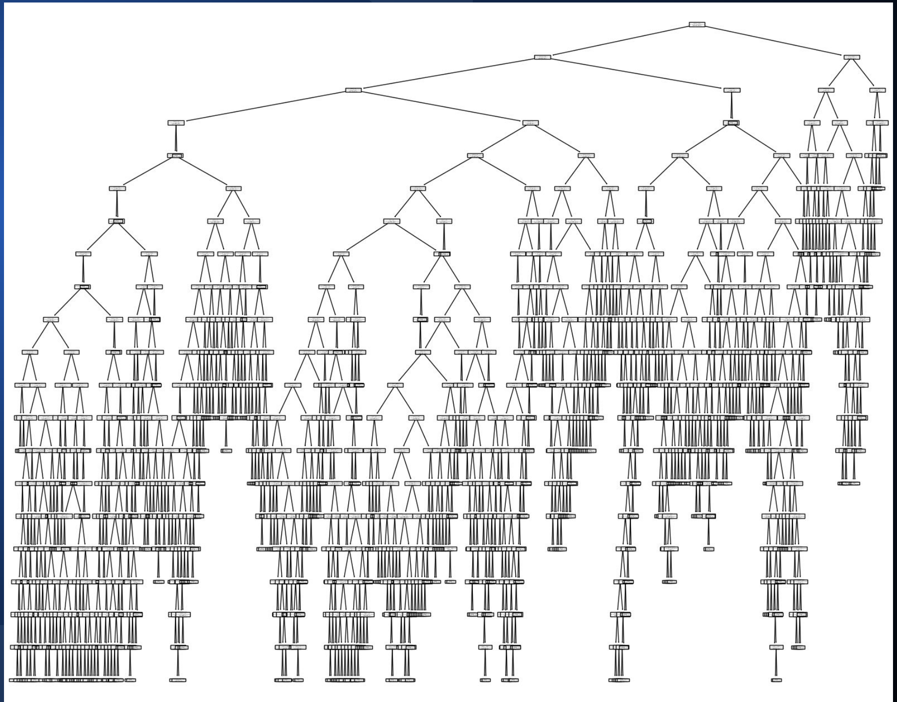
Decision Tree

RMSE

- €11,844,366
- Max Depth =20

Top 3 Features

- Pass Targets
- Diff between Expected and actual Goals within 90 (min)
- Shots on Target

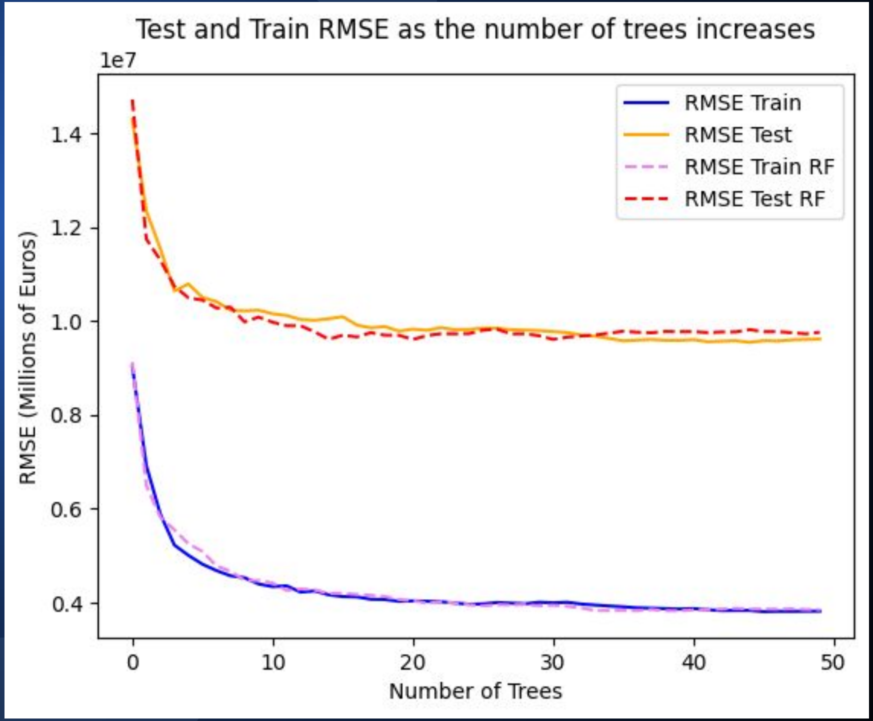


Random Forest

RMSE after Cross Validation

- **€8,933,027**

We used 50 trees and 10 iterations



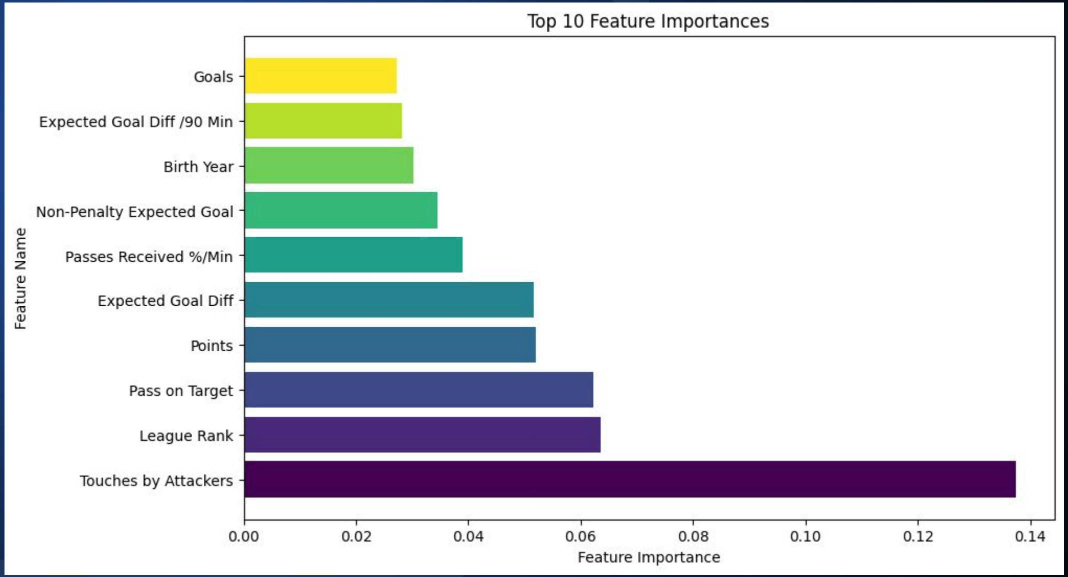
Boosting

RMSE: €10,603,299

Higher RMSE than non-boosted model because of possible

1. **Noise in our dataset**
2. **Computational Power**

The most important features in determining player value are **Touches by Attackers** and **League Rank**



Model Comparison

	RMSE
Naive Rule	€15,442,749
k-NN	€18,860,892
Linear Regression	€16,586,771
Ridge Regression	€11,219,498
Lasso Regression	€12,010,986
Decision Tree	**€11,605,762**
Random Forest	**€8,933,027**
Boosting	**€10,603,299**

Let’s Compare Players!

Phil Foden: €110m

Foden plays for Man City
2023 Stats:
Goals: 11
Assists: 5

Bukayo Saka: €120m

Saka Plays for Arsenal
2023 Stats:
Goals: 14
Assists: 11

Considerations

Data Preprocessing:

With over 400 columns, we had to choose which variables to keep on the basis of relevance

Cross Validation

Given number of variables in our dataset, our Cross Validated RF model took multiple hours to run

Using rMSE

Since our model's outcome is measured in millions of euros, using MSE showed unreadable measures of error

Conclusion

- Our best model is Random Forest with an rMSE of **€8,933,027** compared to our naive rMSE of **€15,442,749**
- If we wanted to predict the price of a new player we would be **about €9 mil off on average**

BACKUP SLIDES

Dataset Overview

395 VARIABLES

- 5 dropped
- 7 Categorical
  - *Player, Nationality, Position, Squad, Position2 , Foot, League,*

1773 OBSERVATIONS

- Approx. 300 Missing Values

FEATURES:

- Age
- Birth year
- Height
- Games
- Games_starts
- Minutes
- Goals/Assists

Challenges

Data Preprocessing:

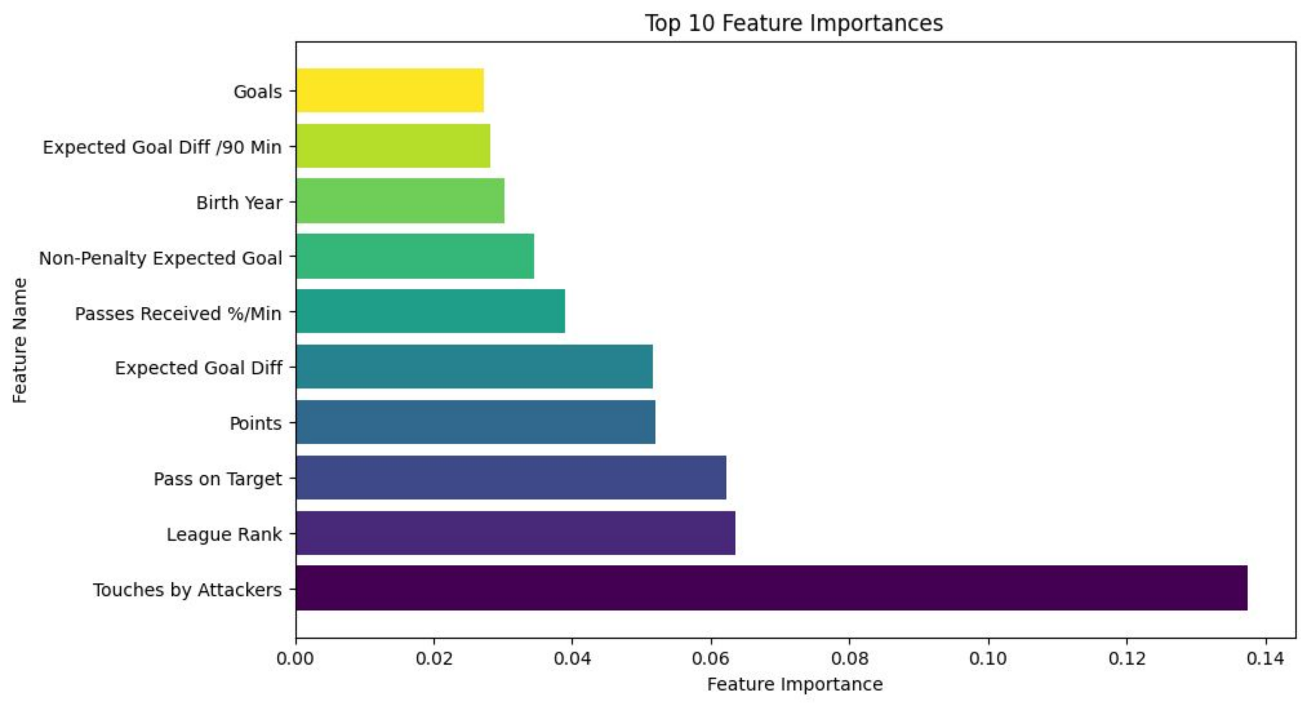
400 columns

Cross Validation

Given number of variables in our dataset, our Cross Validated RF model took multiple hours to run

Using rMSE

Since our model's outcome is measured in millions of dollars, using MSE showed unreadable measures of error



Based on our Boosted Random Forest model, the most important features in determining player value are **Touches by Attackers** and **League Rank**