U
OF
T
University of Toronto
Department of Civil
and Mineral Engineering

TEAM GEOARCH INNOVATORS

Big Project

Presenters:
Fion Ouyang, Mohammed Hossein Khosravi, Nanqiao Du & Nyah Bay

Presentation Date:
December 2024

Presentation For:
CME538H

02

Data + EDA

03

Feature Engineering + Modelling

01

Predicting Soccer Match Outcomes: A Data-Driven Approach

Exploring the World of Soccer, Betting, and Predictive Analytics

Soccer and the World of Sports Betting

- **Soccer's Popularity:** Most popular sport worldwide with over 3.5 billion fans.
- **Sports Betting:** one trillion wagered annually across all sports.
- **Betting Odds:** odds represent the likelihood of outcomes.
- **Losing Streaks:** 90% of betters will lose money in the long run.

Did you know? Leicester City had 5000:1 odds to win the Premier League in 2016 and they did!

Why Predicting Soccer Matters?

- **The Challenge**: Soccer is the hardest sport to predict because of its low scoring nature. External factors like injuries, weather, and referee decisions amplify unpredictability.
- **Betting Implications**: If you can predict outcomes more than **50%** of the time, you can beat sportsbooks.
- **Real-World Relevance**: Beyond betting, accurate predictions can help teams and coaches strategize better. For fans, it enhances the excitement of following games.

CAN DATA SCIENCE UNLOCK THE SECRETS OF SOCCER?

01

MATCH STATISTICS

The Data Behind the Game

01

MATCH STATISTICS

02

PLAYER INFORMATION

03

TEAM DATA

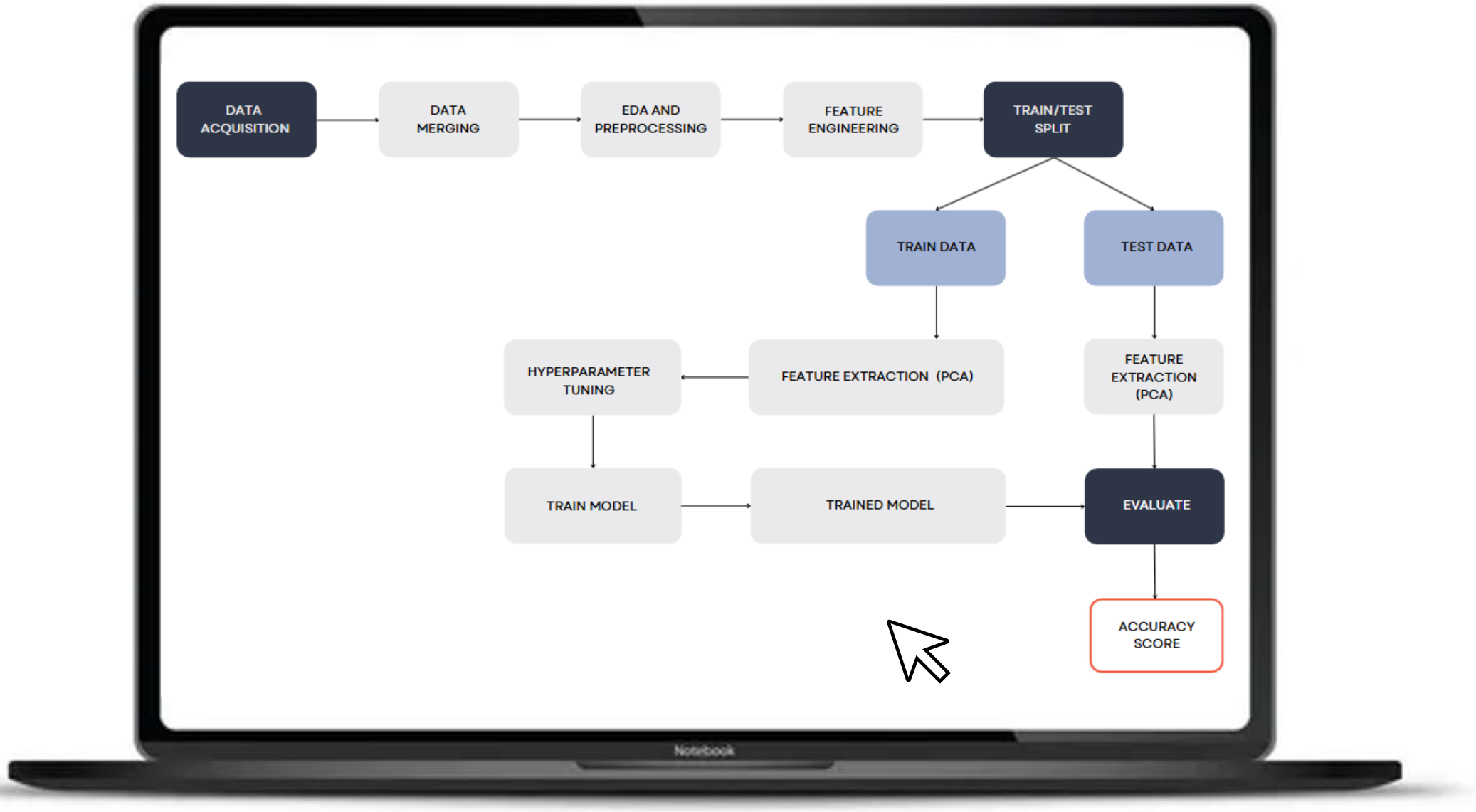
04

LEAGUE & COUNTRY DETAILS

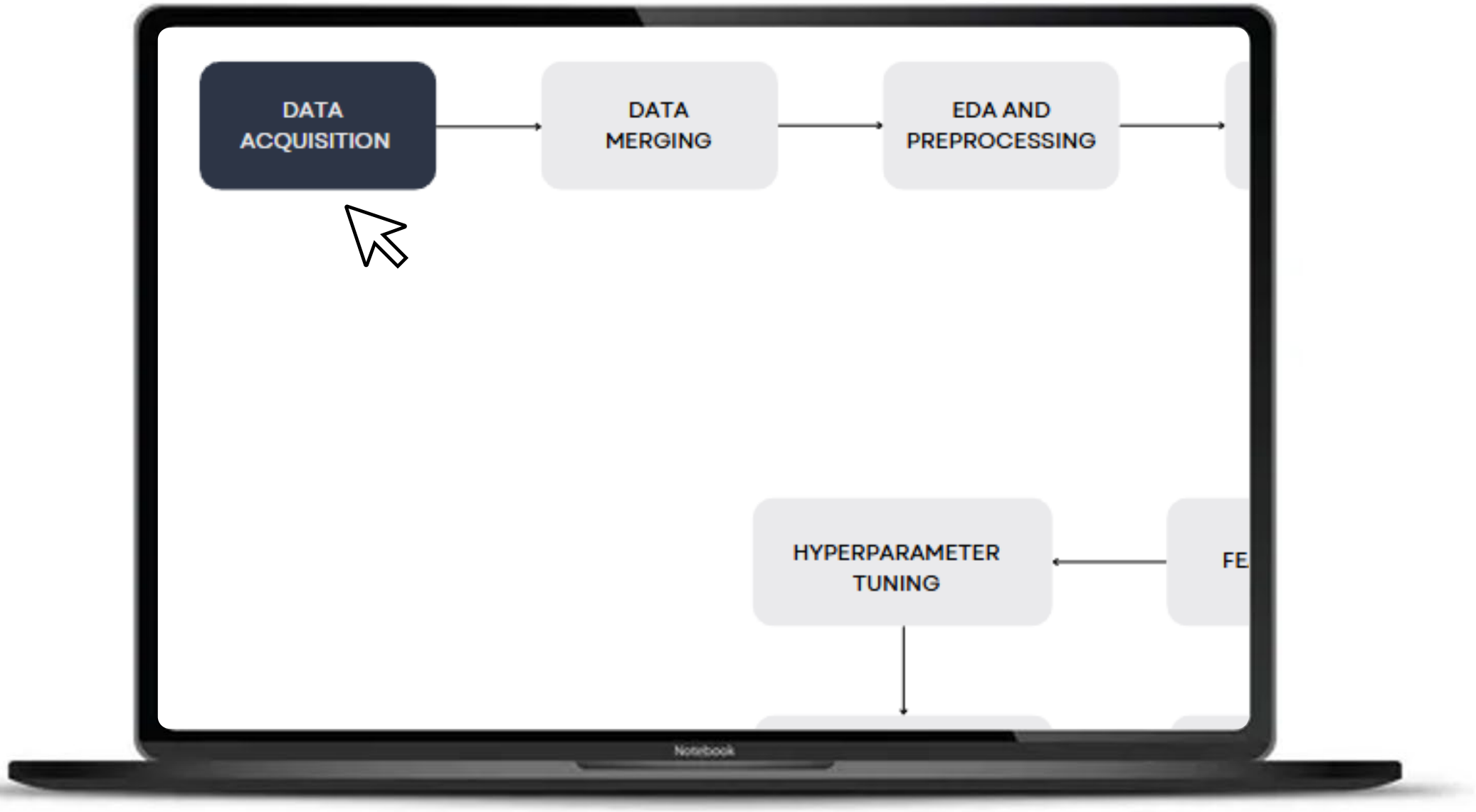
05

BETTING ODDS

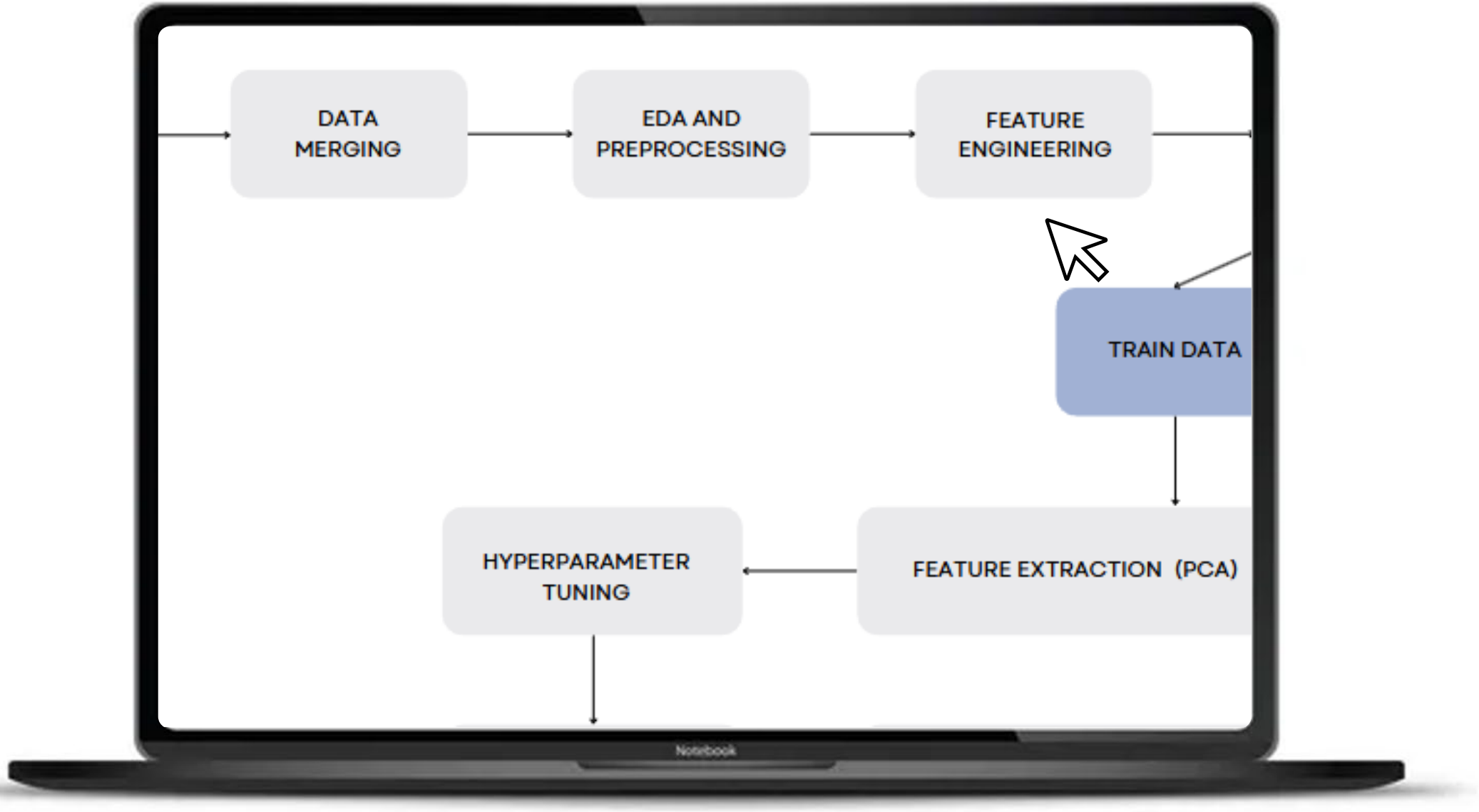
Workflow Diagram



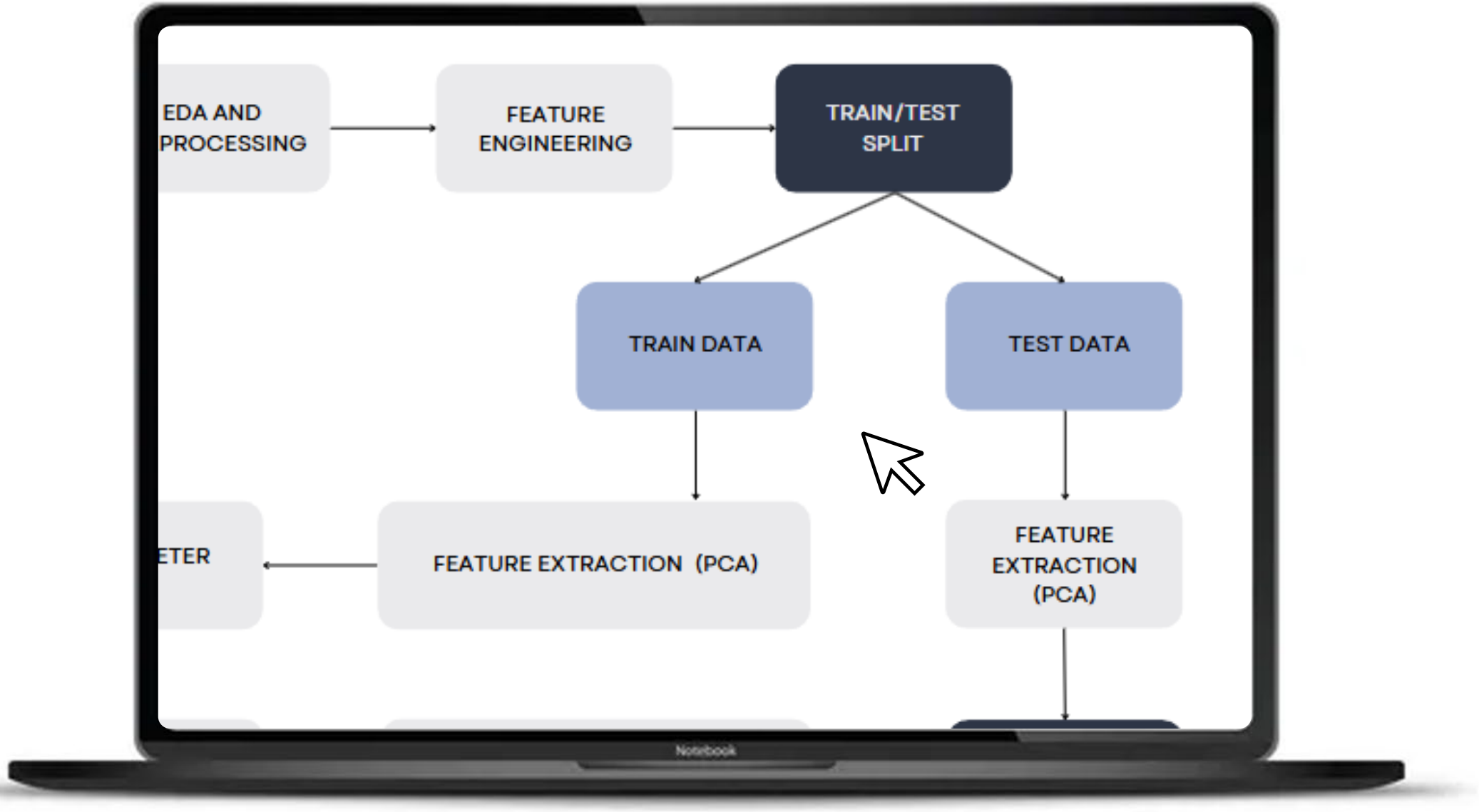
Workflow Diagram



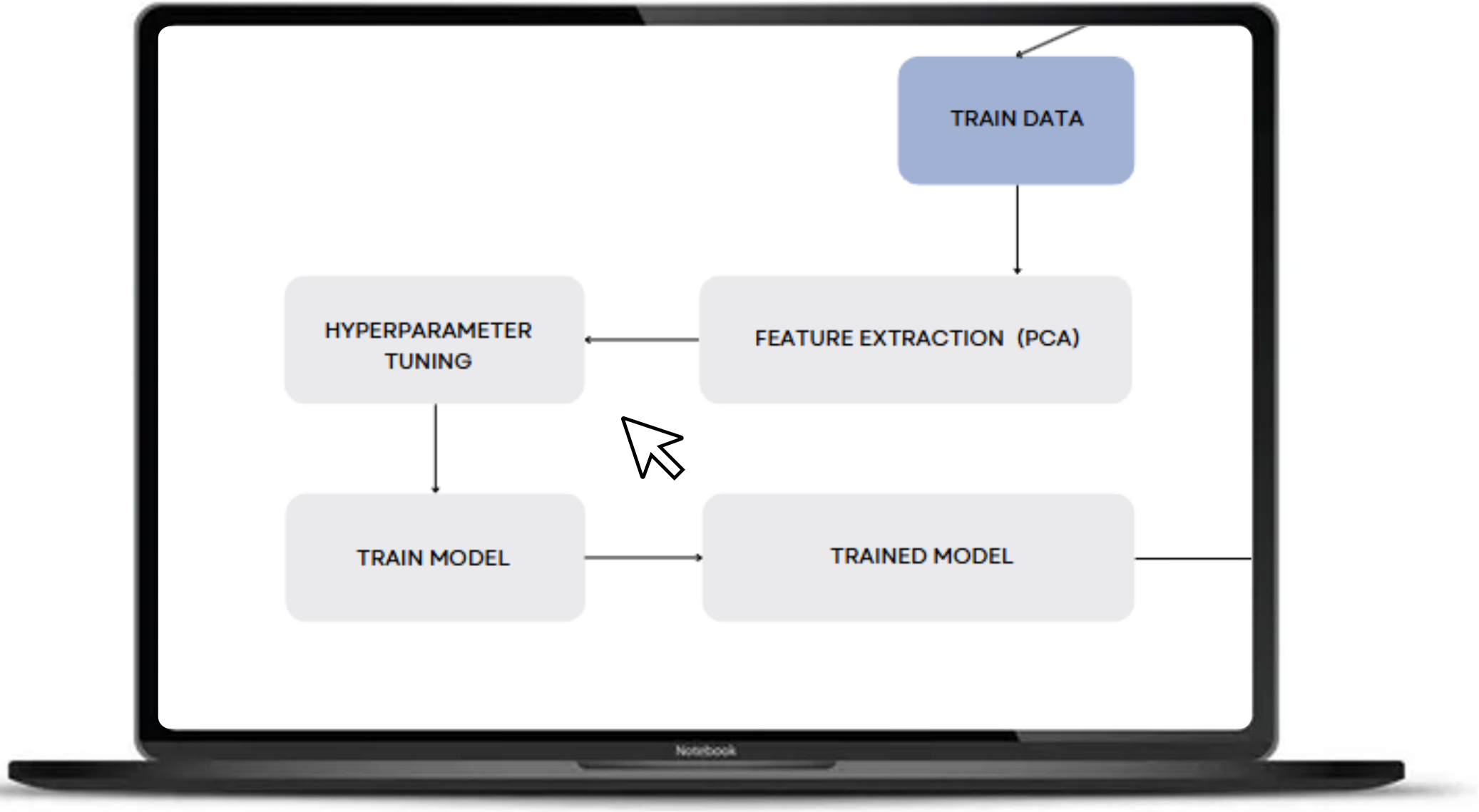
Workflow Diagram



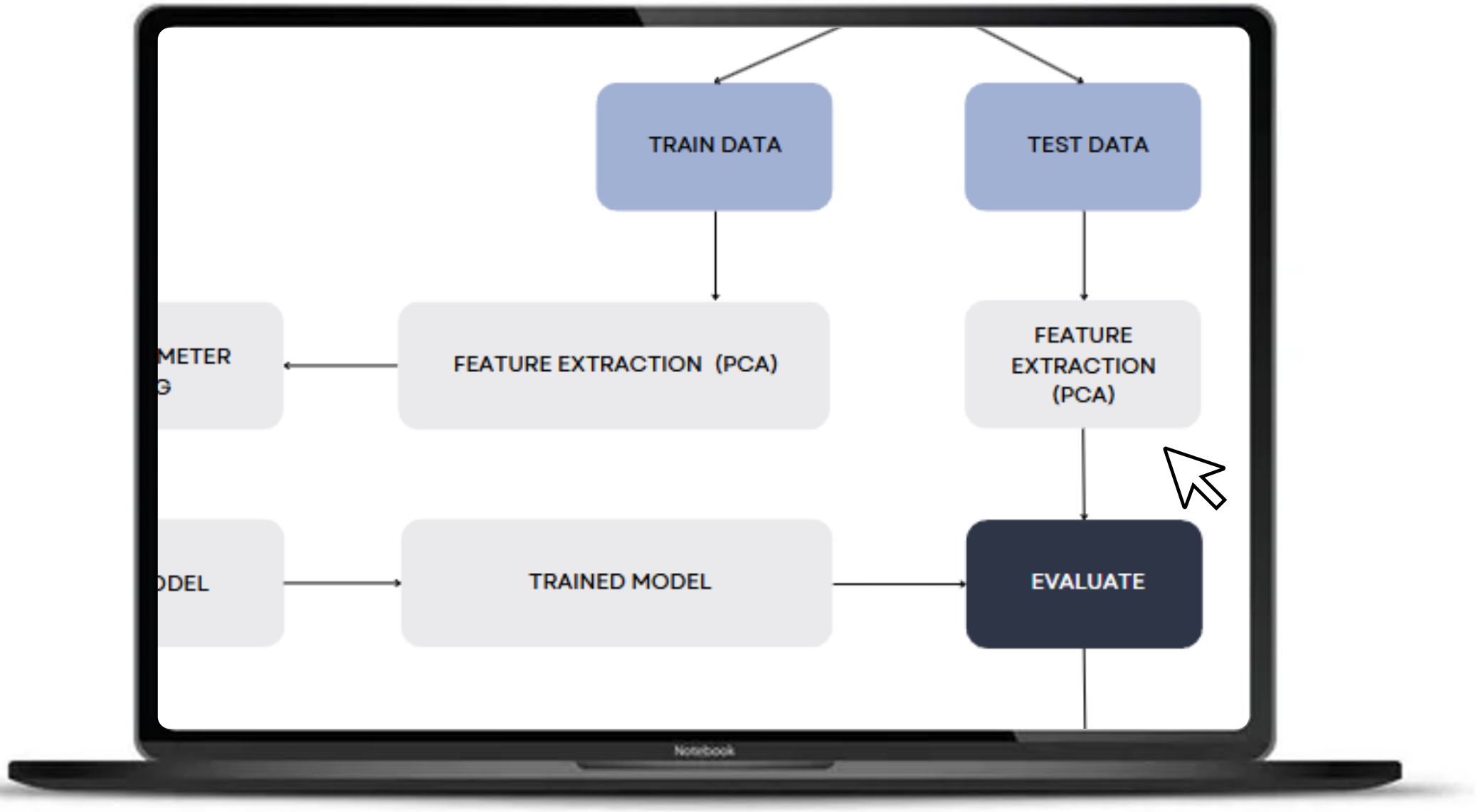
Workflow Diagram



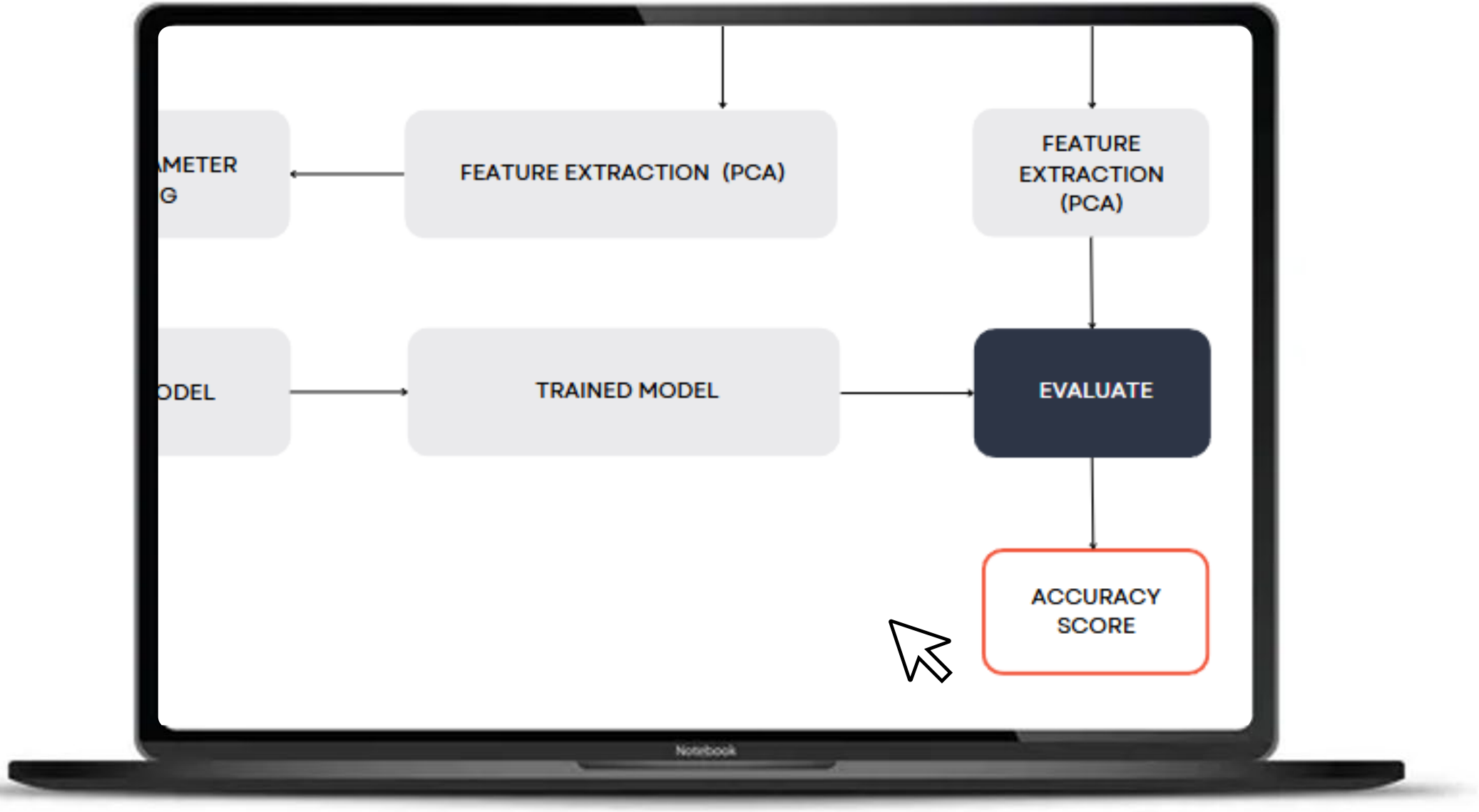
Workflow Diagram



Workflow Diagram



Workflow Diagram



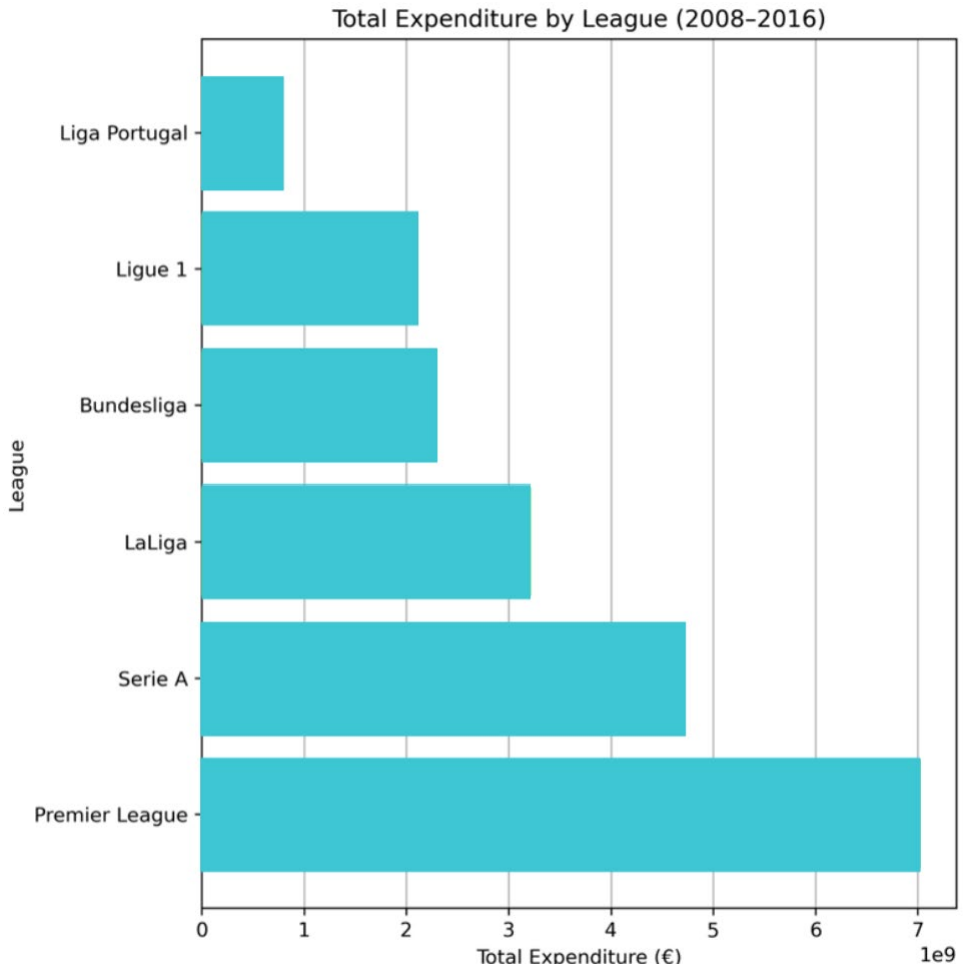
02

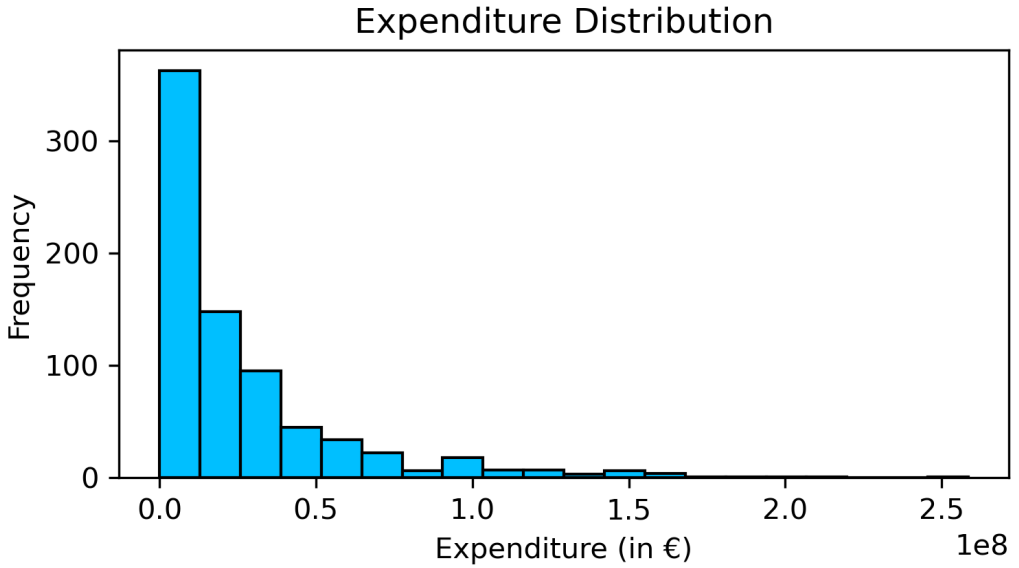
DATA AND EDA

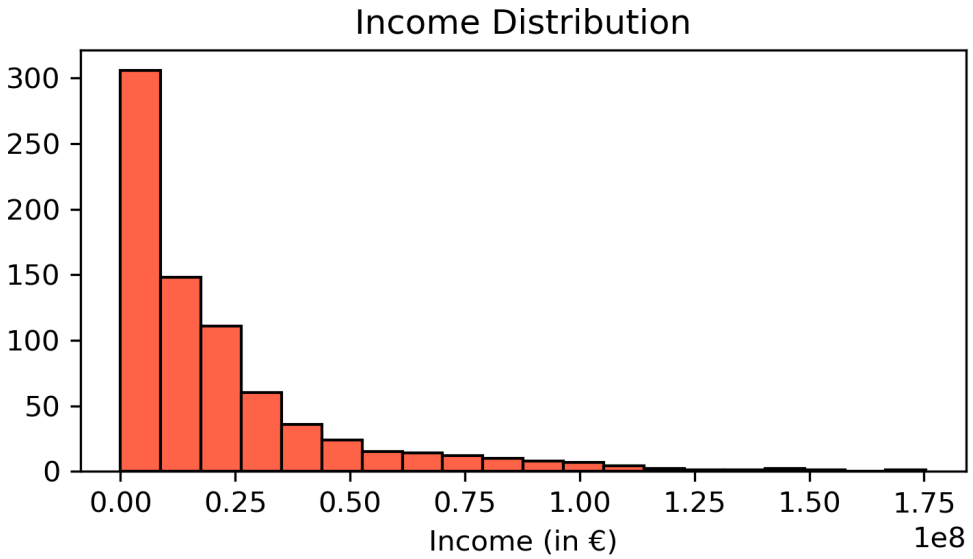
Financial Data

- **Transfermarkt:** a football website for transfers, rumours, market values, and stats
- We used **web scraping** to collect expenditure and income data for 6 European leagues

Expenditure Data (a)

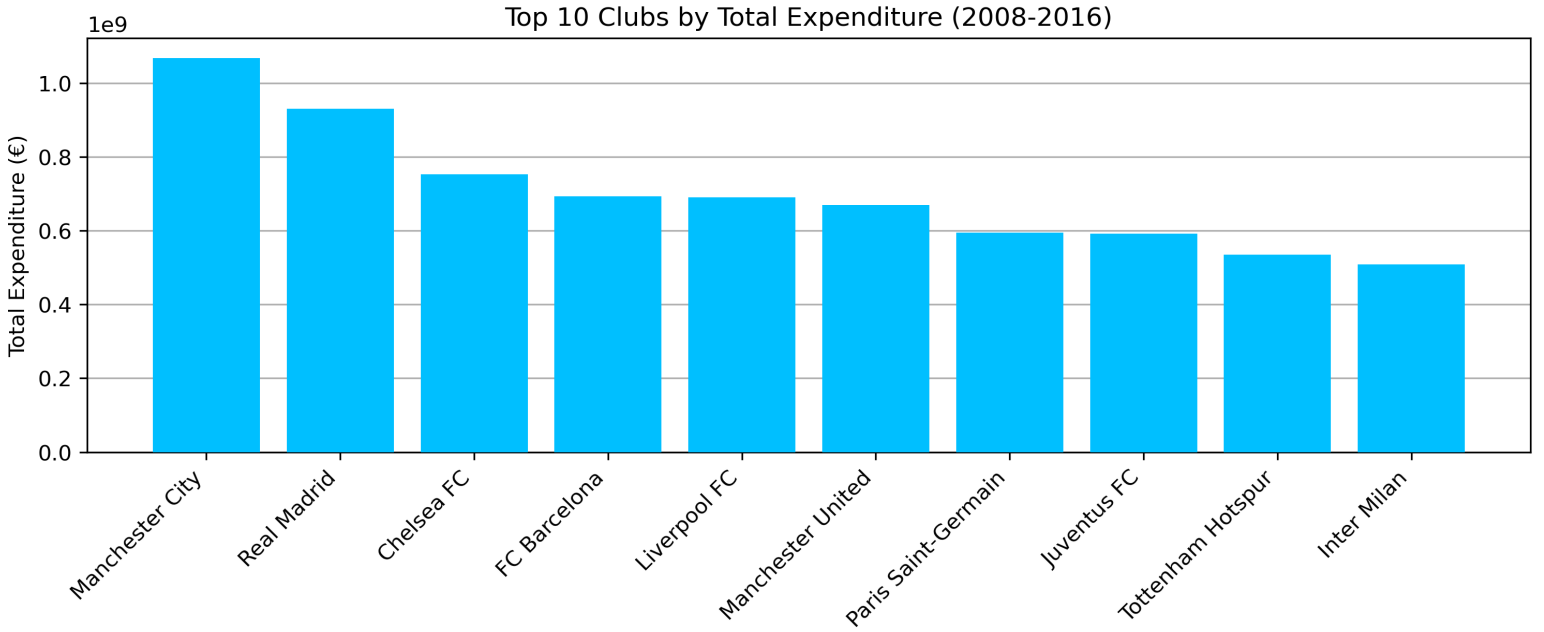




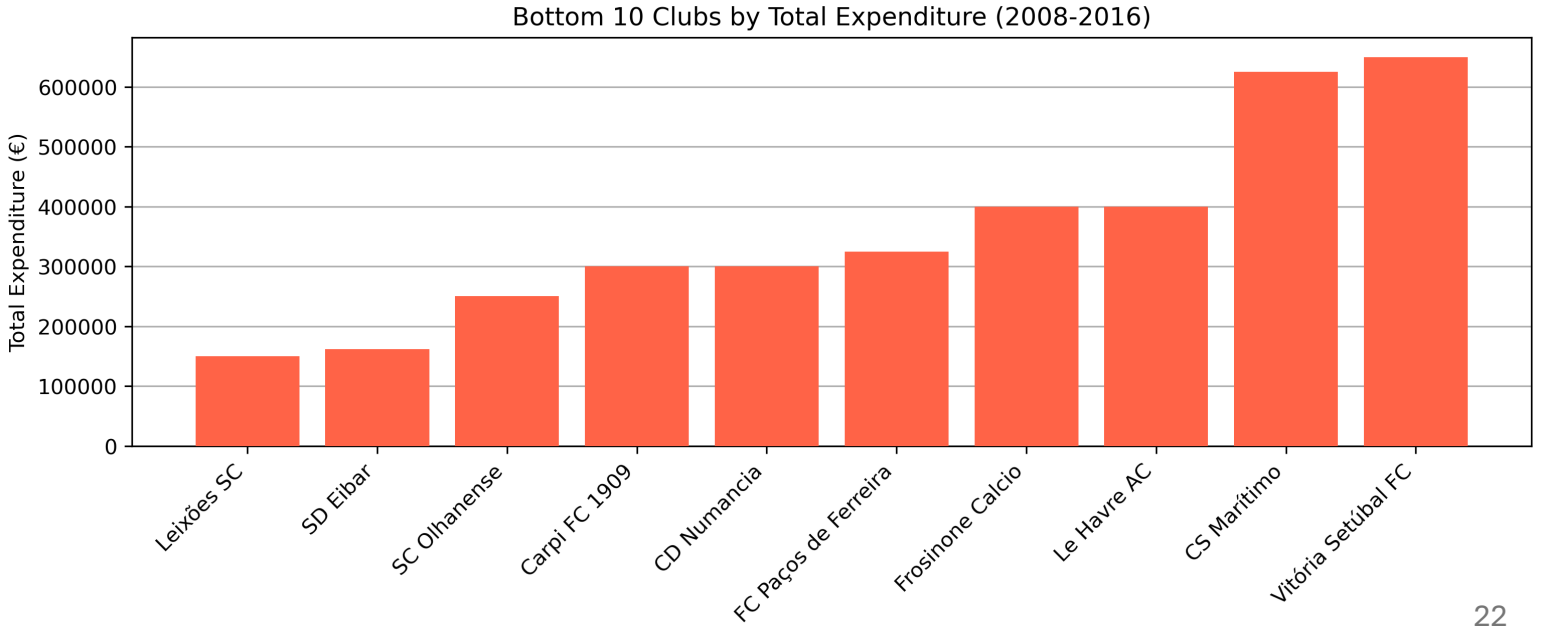


Expenditure Data (b)

TOP 10 TEAMS



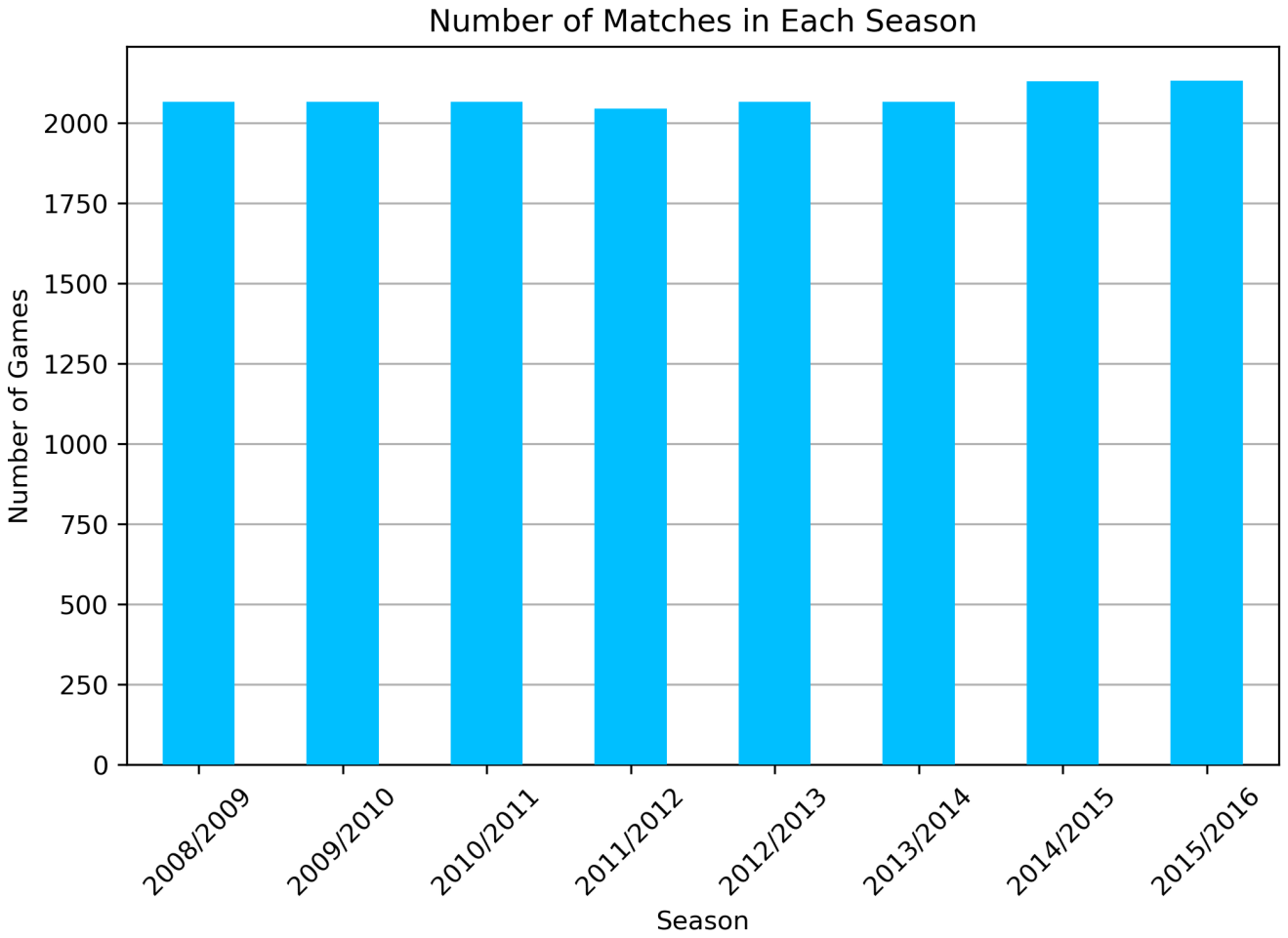
BOTTOM 10 TEAMS



Football Data

- SQLite database from **Kaggle**
- +25,000 matches
- +10,000 players
- Highest leagues of 11 European countries
- 2008 to 2016
- Player and Team attributes

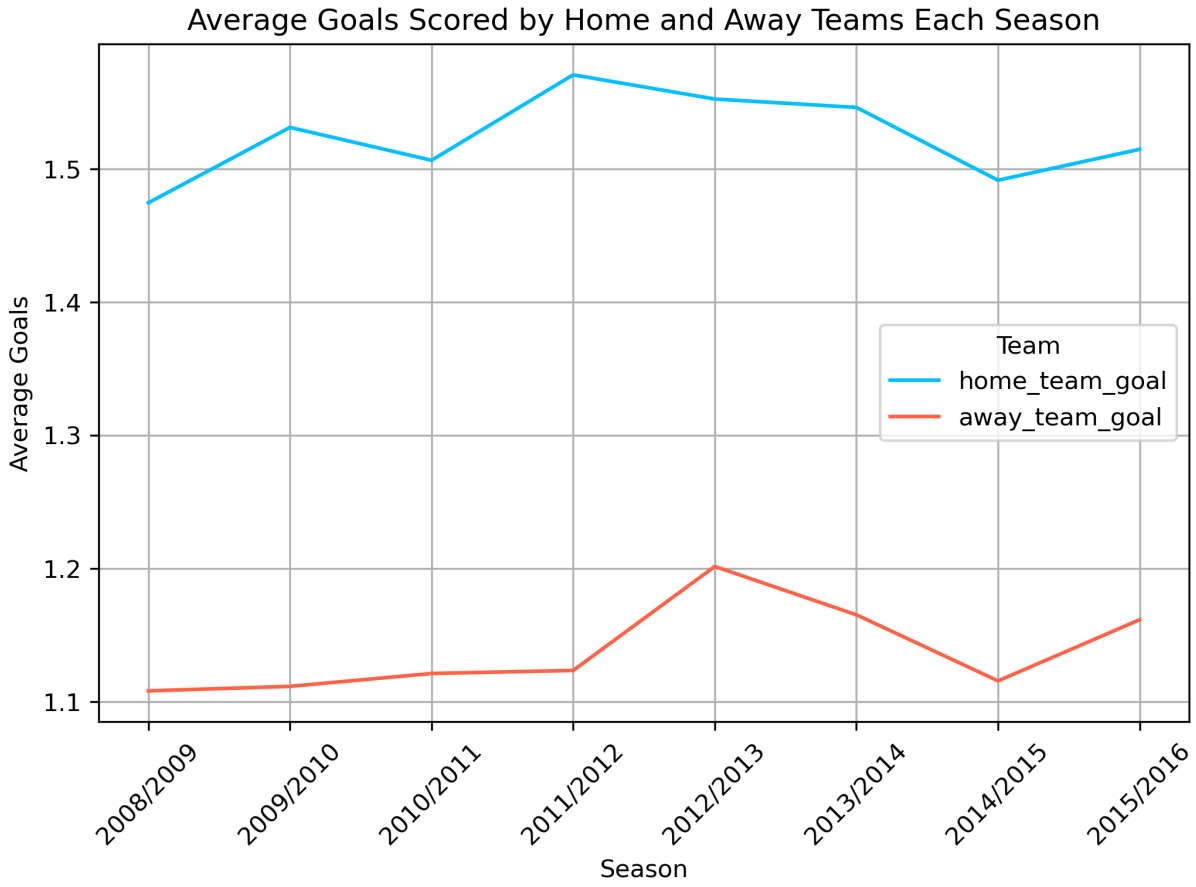
16,637 matches after filtering for the **6 leagues** with expenditure/income data



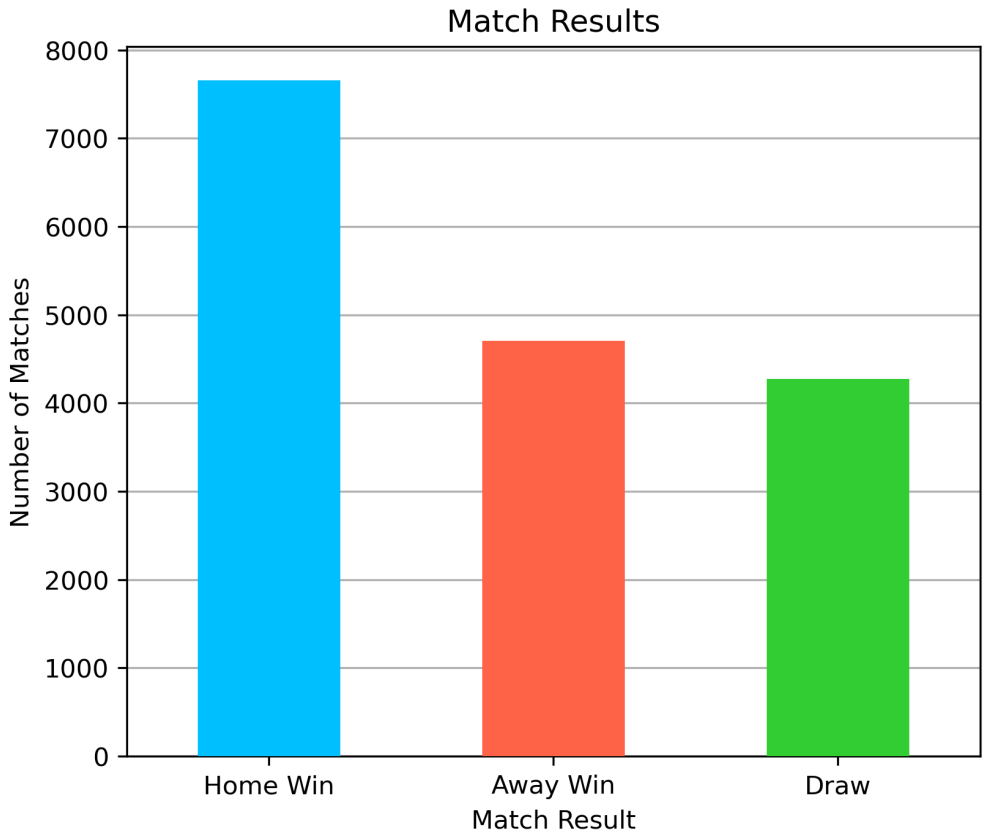
Match Data

- Team and player IDs
- Squad formations as X,Y coordinates
- Betting odds
- Match outcome
- In-game events (goals, fouls, cards, etc.)

a. Goals

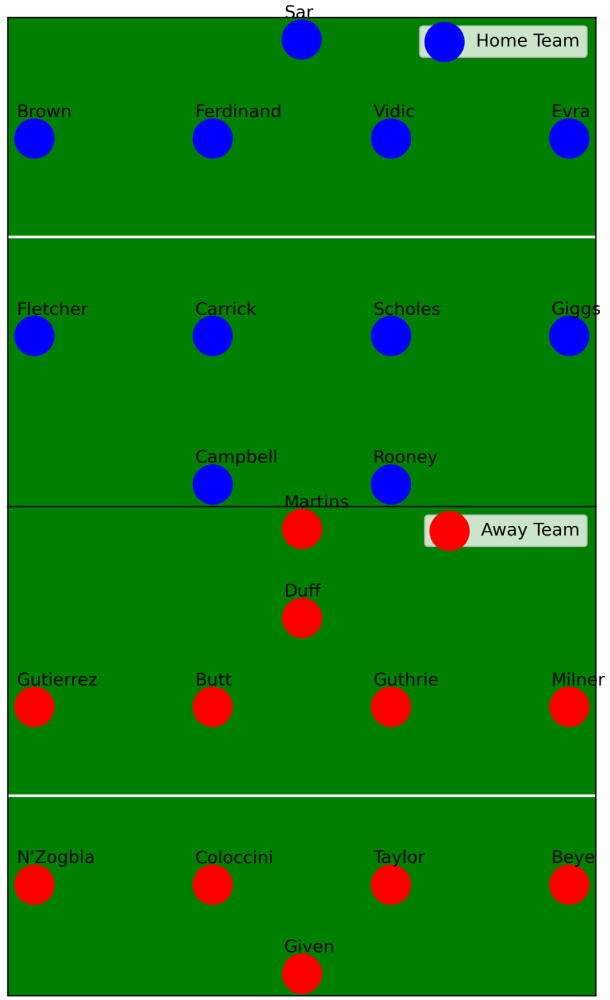


b. Results

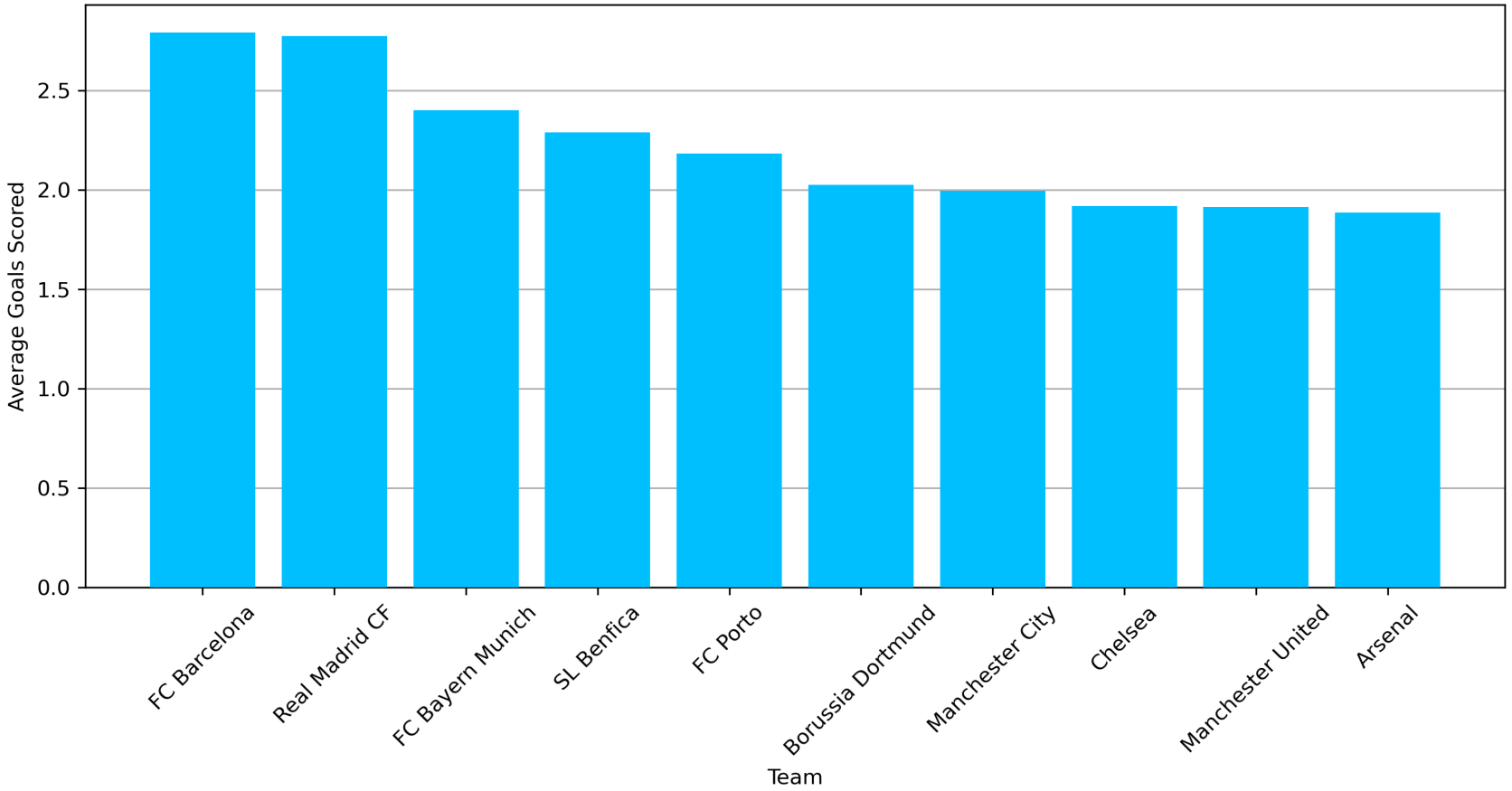


Match Data

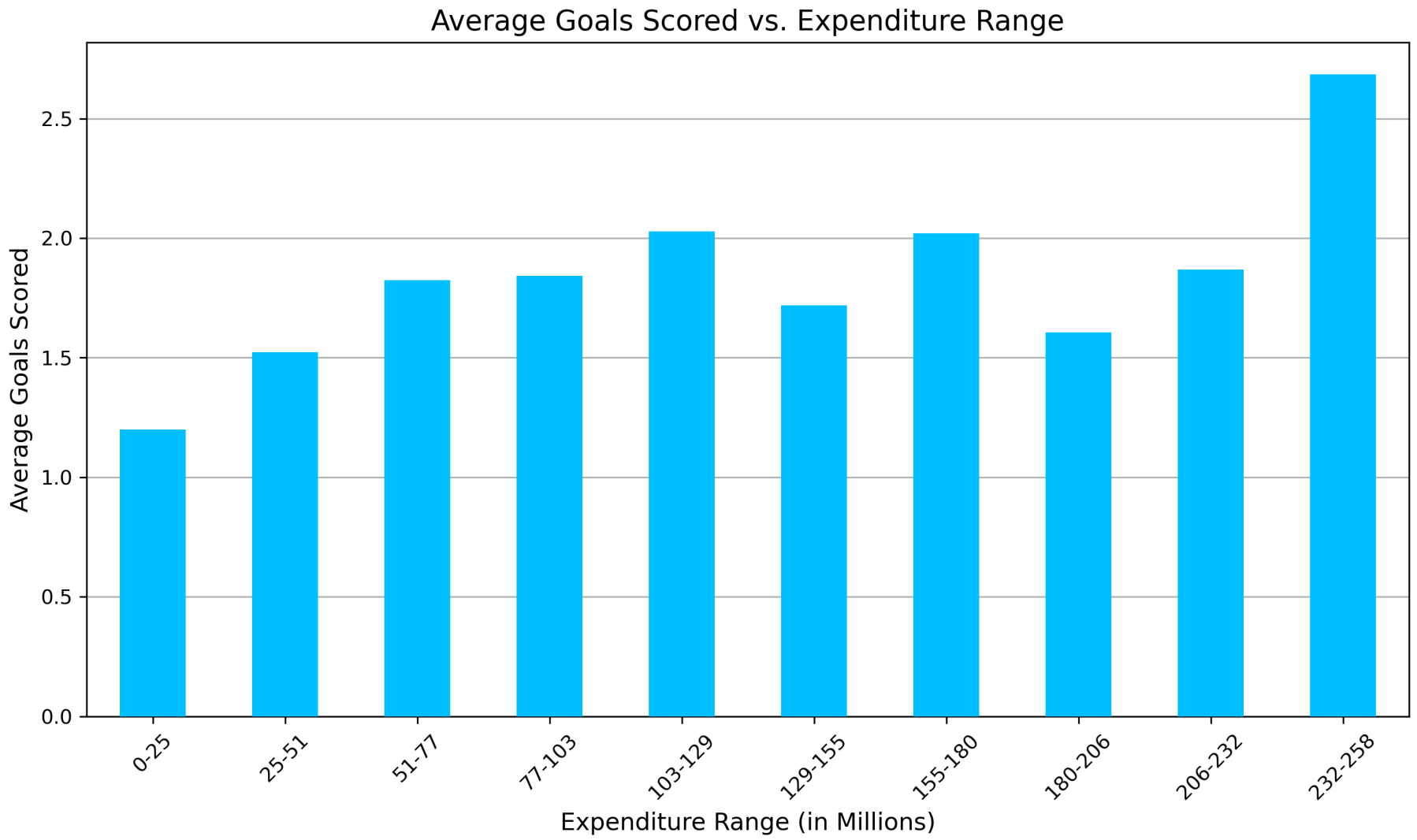
a. Squad Formation



b. Top 10 Teams by Average Goals Scored Per Match



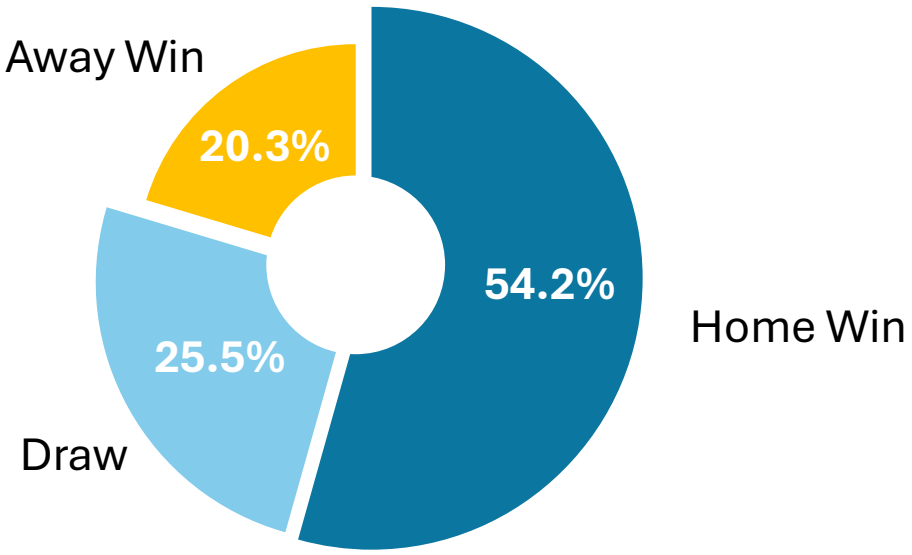
Goals vs. Expenditure Range



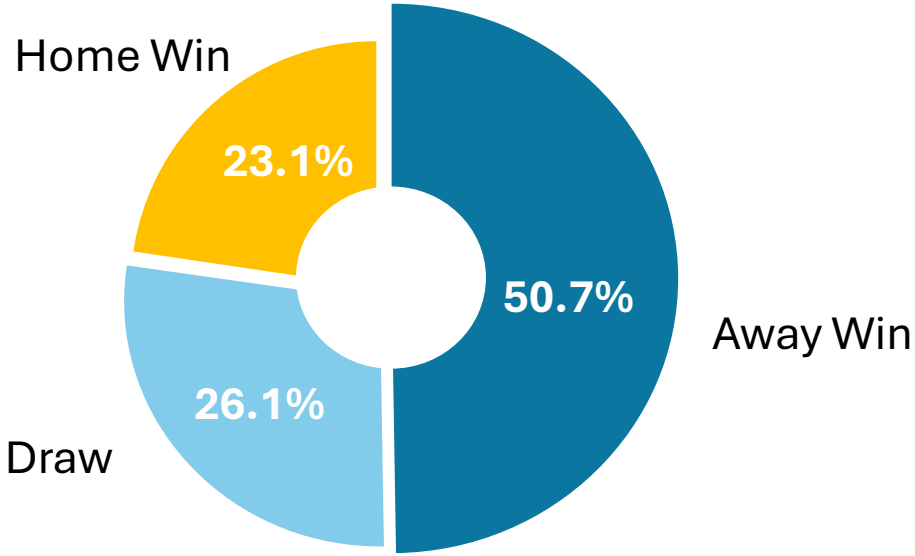
Match Data: Home and Away Favourites

- Betting Odds for *Bet365, BWin, Pinnacle Sports, William Hill, Stan James, Victor Chandler, Gamebookers, BetSafe*.
- **Odds:** reflect the probability of an event occurring. The higher the probability, the lower the odds.
- **Home Favourites:** home win has the lowest betting odds.
- **Away Favourites:** away win has the lowest betting odds.

a. Home Favourites (Bet365 Odds)



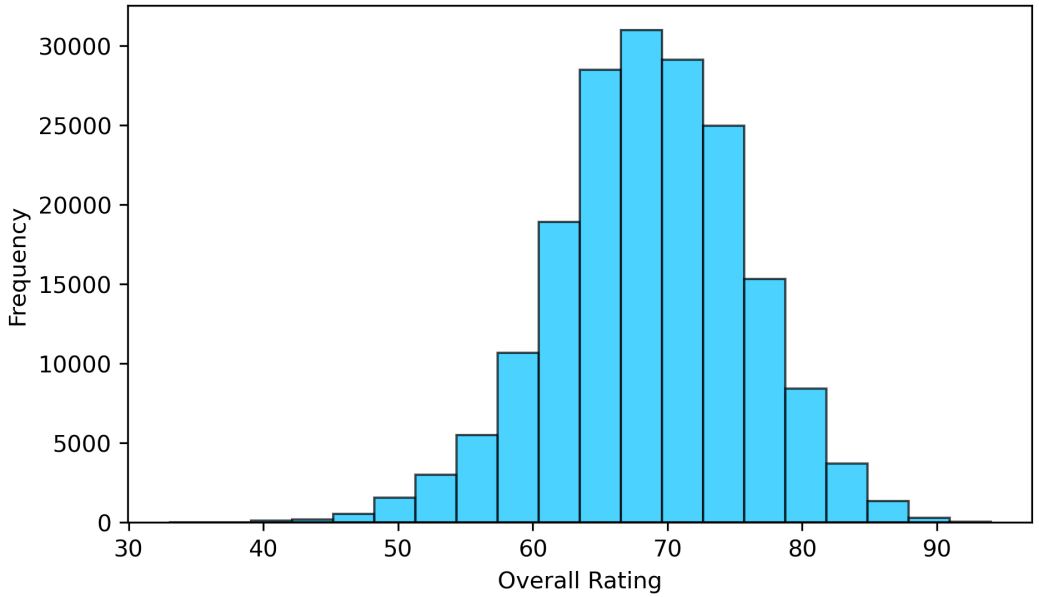
b. Away Favourites (Bet365 Odds)



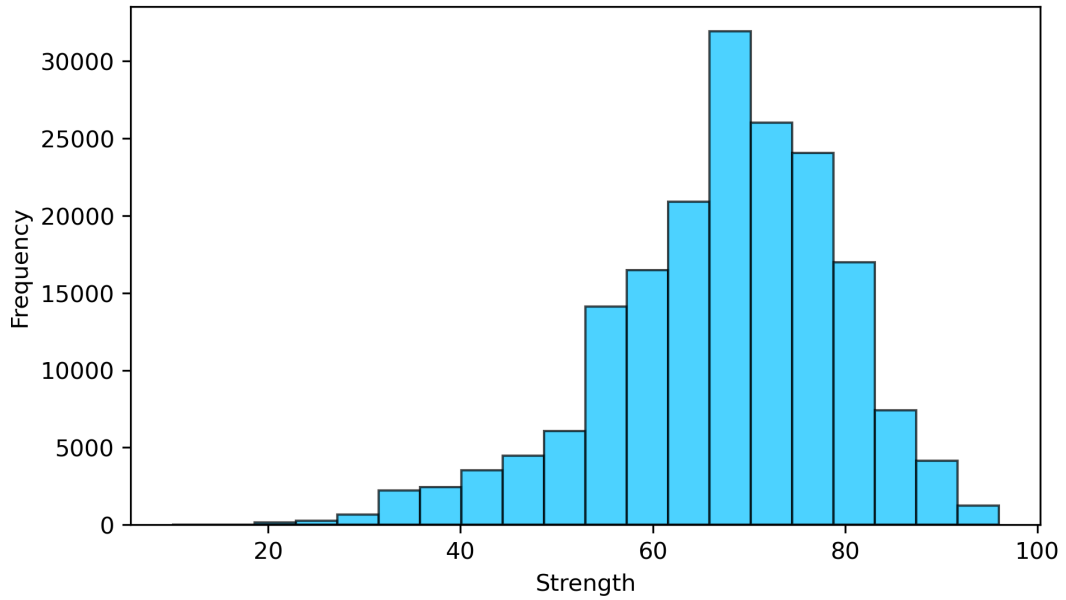
Player Attributes

- 37 columns of player attributes
- **Overall Rating and Potential:** overall rating, potential
- **Technical Skills:** crossing, finishing, dribbling, ball control, etc.
- **Physical Attributes:** acceleration, sprint speed, agility, strength, etc.
- **Mental Attributes:** aggression, positioning, vision, composure, etc.
- **Goalkeeping Skills:** diving, handling, kicking, positioning, reflexes.

a. Overall Rating



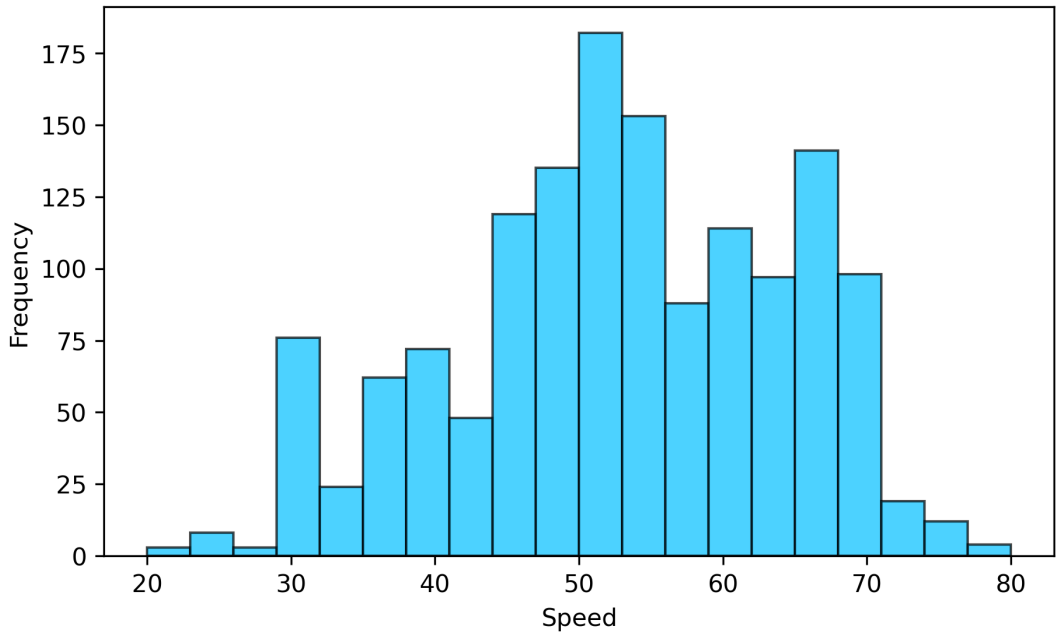
b. Strength



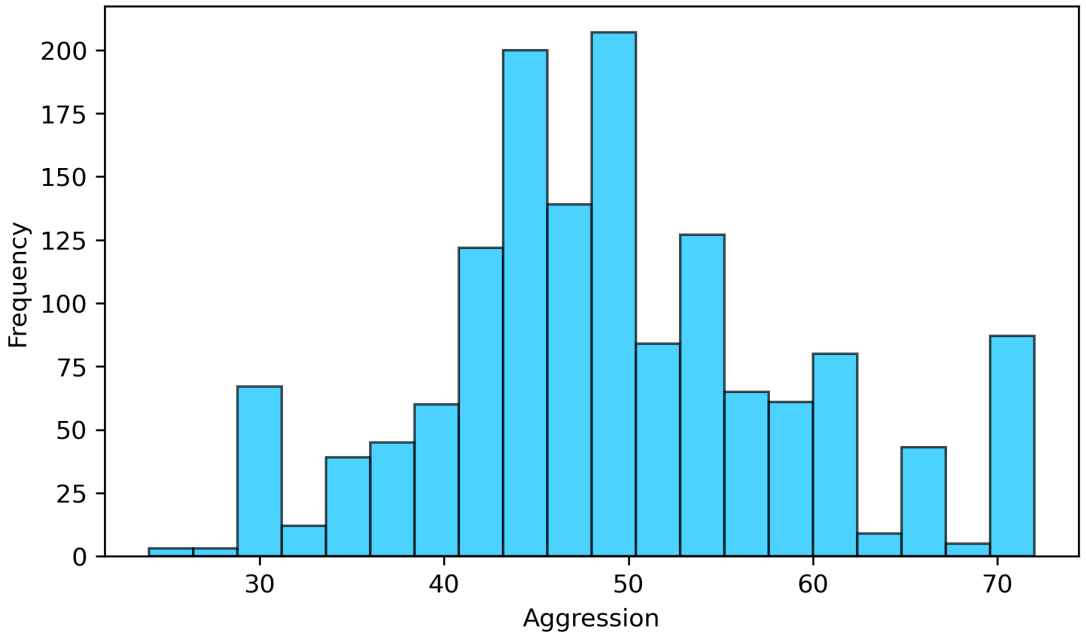
Team Attributes

- 21 columns of team attributes.
- **Build-Up Play**: Speed, Dribbling, Passing, and Positioning.
- **Chance Creation**: Passing, Crossing, Shooting, and Positioning.
- **Defensive Tactics**: Pressure, Aggression, Team Width, and Defender Line.

a. Build- Up Play Speed



b. Defence Aggression



03

Feature Engineering

Data Cleaning

Data Merging

Merging Team, Player, and Match dataset.

Data Merging for Expenditure

Expenditures are from web scraping, only for 6 leagues, and only in 2008-2016.

Data Conversion

String to float/int/datetime

Match Outcome Calculation

From home team goals – away team goals

Unreasonable Data Removing

Team formation, player attributes, etc.

Handle Missing Values + Remove Duplicate Rows.

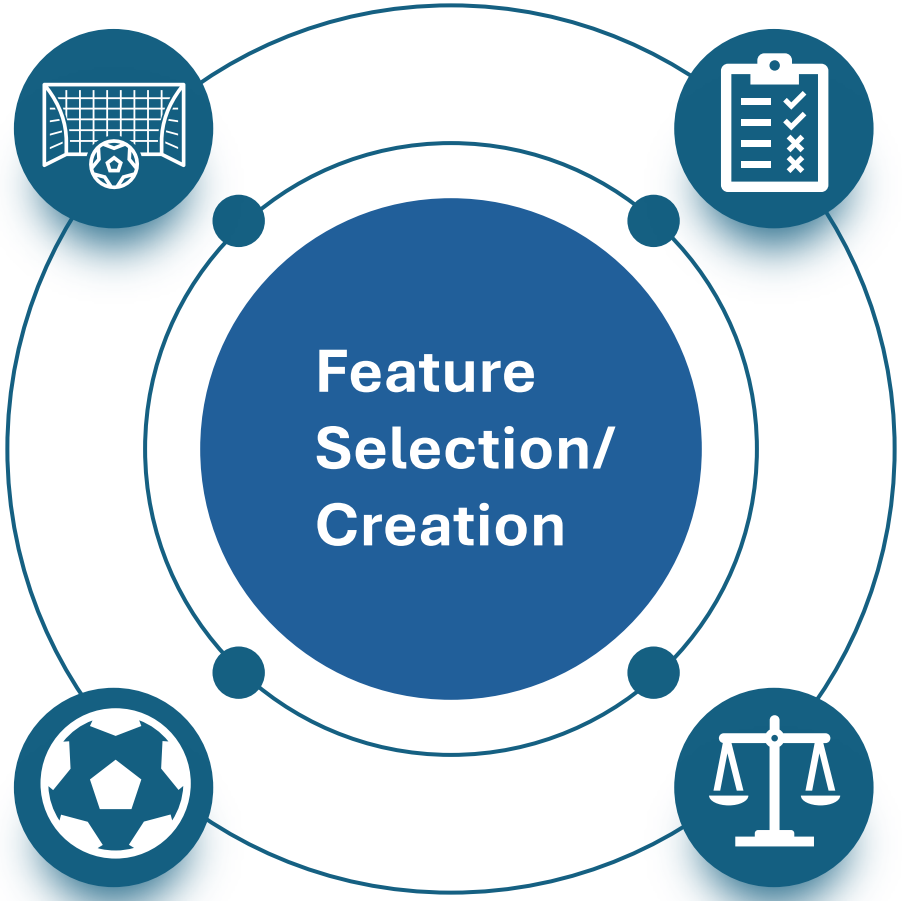
Encoding

Encoding categorical data by label/one-hot encoding

Feature Selection/ Creation

Drop in-game Features
Red/yellow cards, shots, goals, etc.

Goalkeeper Features
Home/away goalkeeper skills ratio

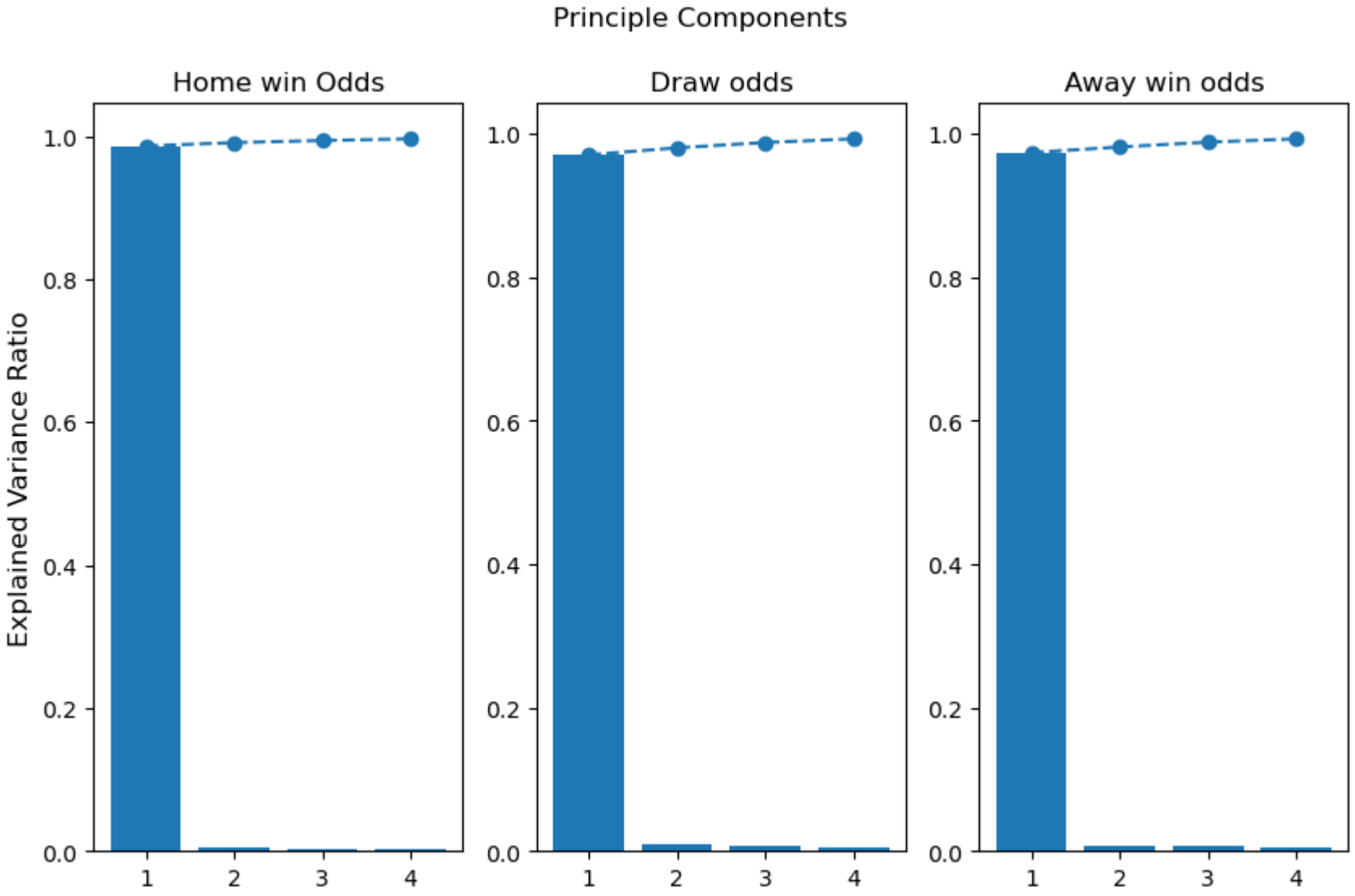


Previous Performance
Results of last 3 games, ranking before each game, etc.

Team Overall Attributes Ratio
Home/away player (except goalkeeper) average skills ratio, team attributes ratio, etc.

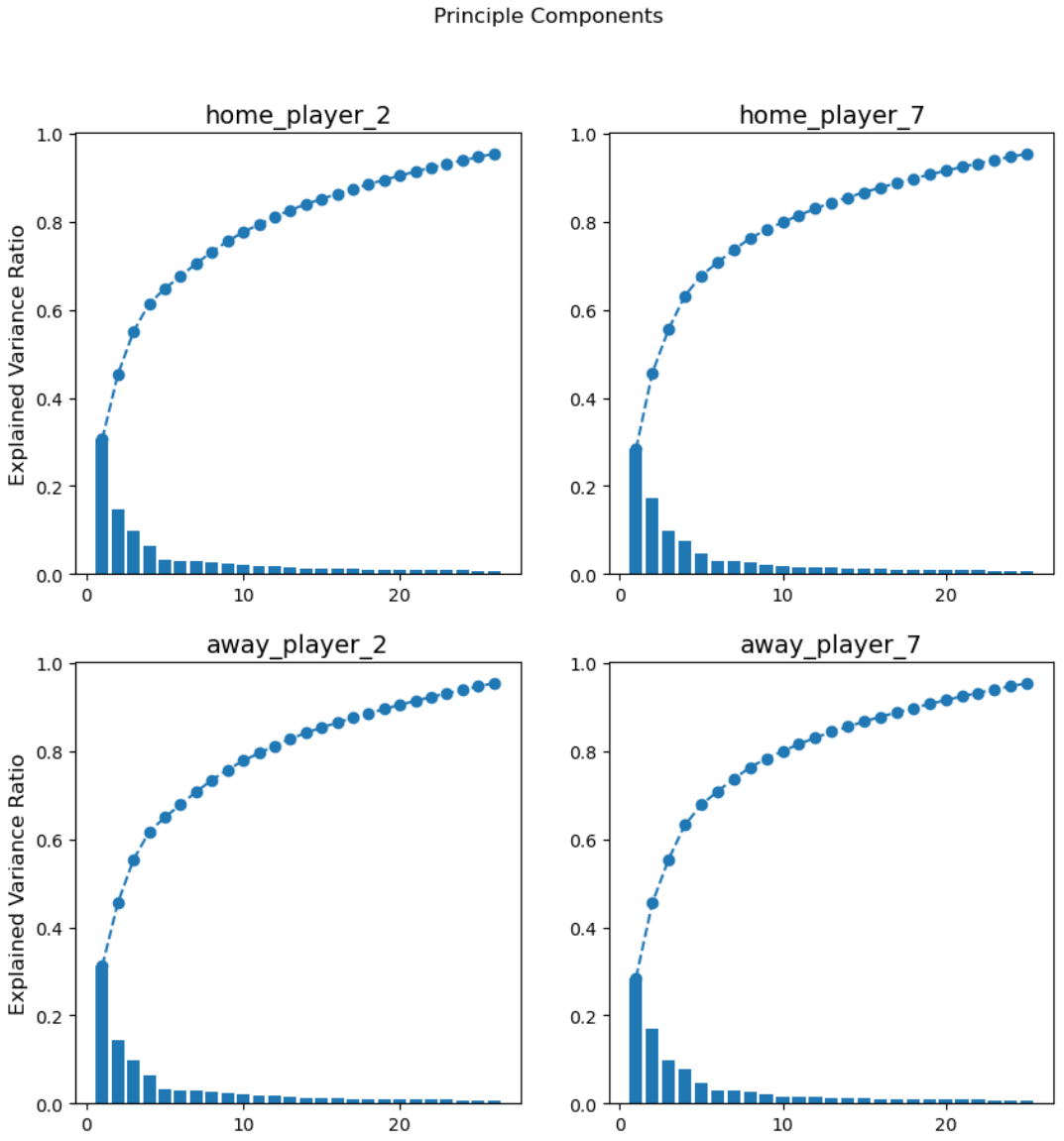
Principle Component Analysis: Betting Odds

- **Betting odds from 5 Institution:** Win/Draw/Lose
- **PCA analysis**: dimensional reduction by choose # of components that can explain 90% of the original data.
- **Results:** 15-d data are compressed to only 3-d.



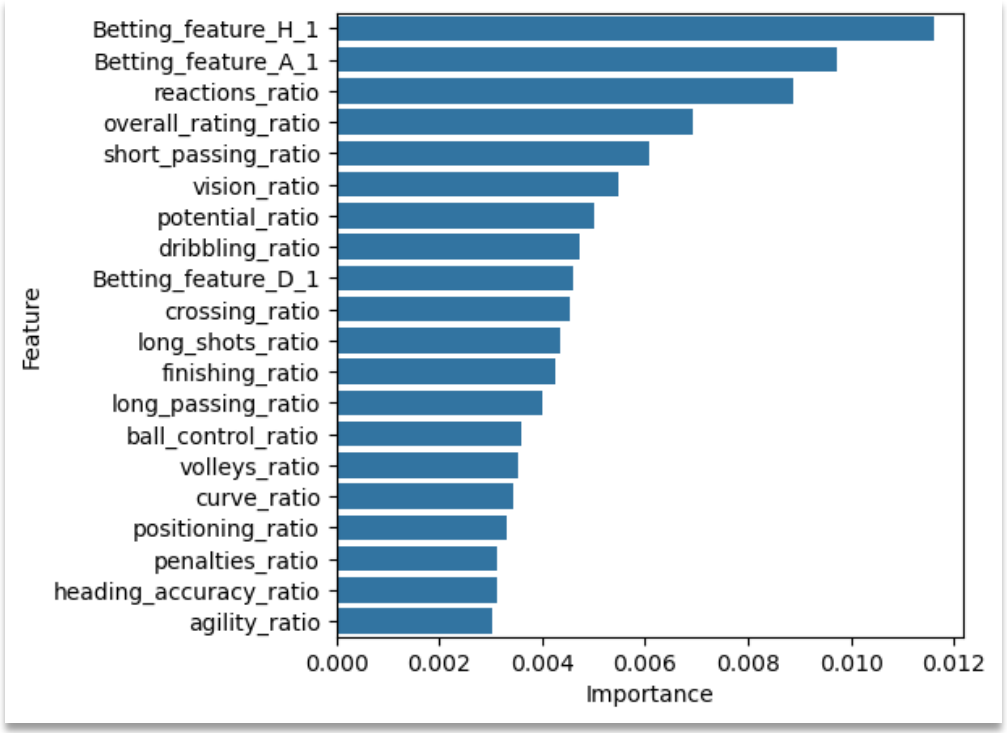
Principle Component Analysis: Players

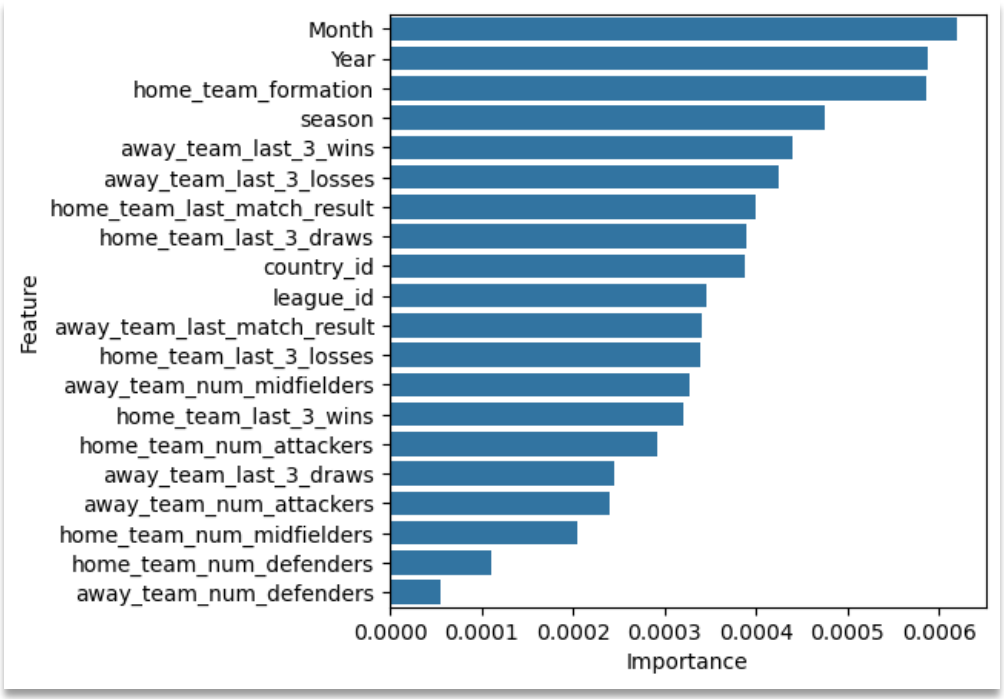
- **Each player for each match has 35 skill features**
- **PCA analysis**: dimensional reduction by choose # of components that can explain 90% of the original data.
- **Results:** 350-d data are compressed to only 200-d.



Feature Importance

- Run a simple Random Forest model and print feature importance.
- Drop the last 15 features.
- 598 features remain.

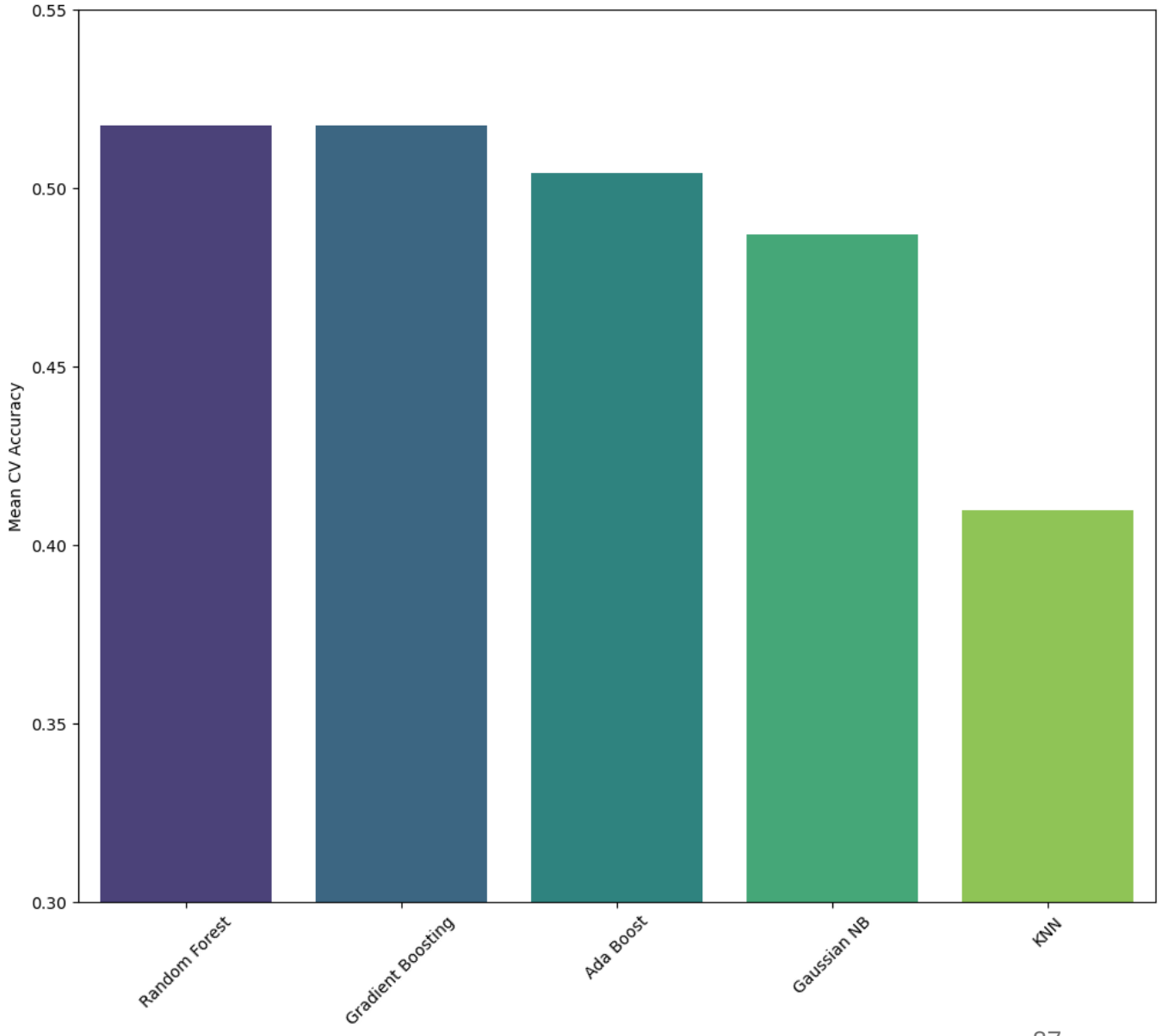




04

Results and Conclusion

Comparisons of Different Classifiers



Hyperparameter Optimization

Using OneHotEncodder...

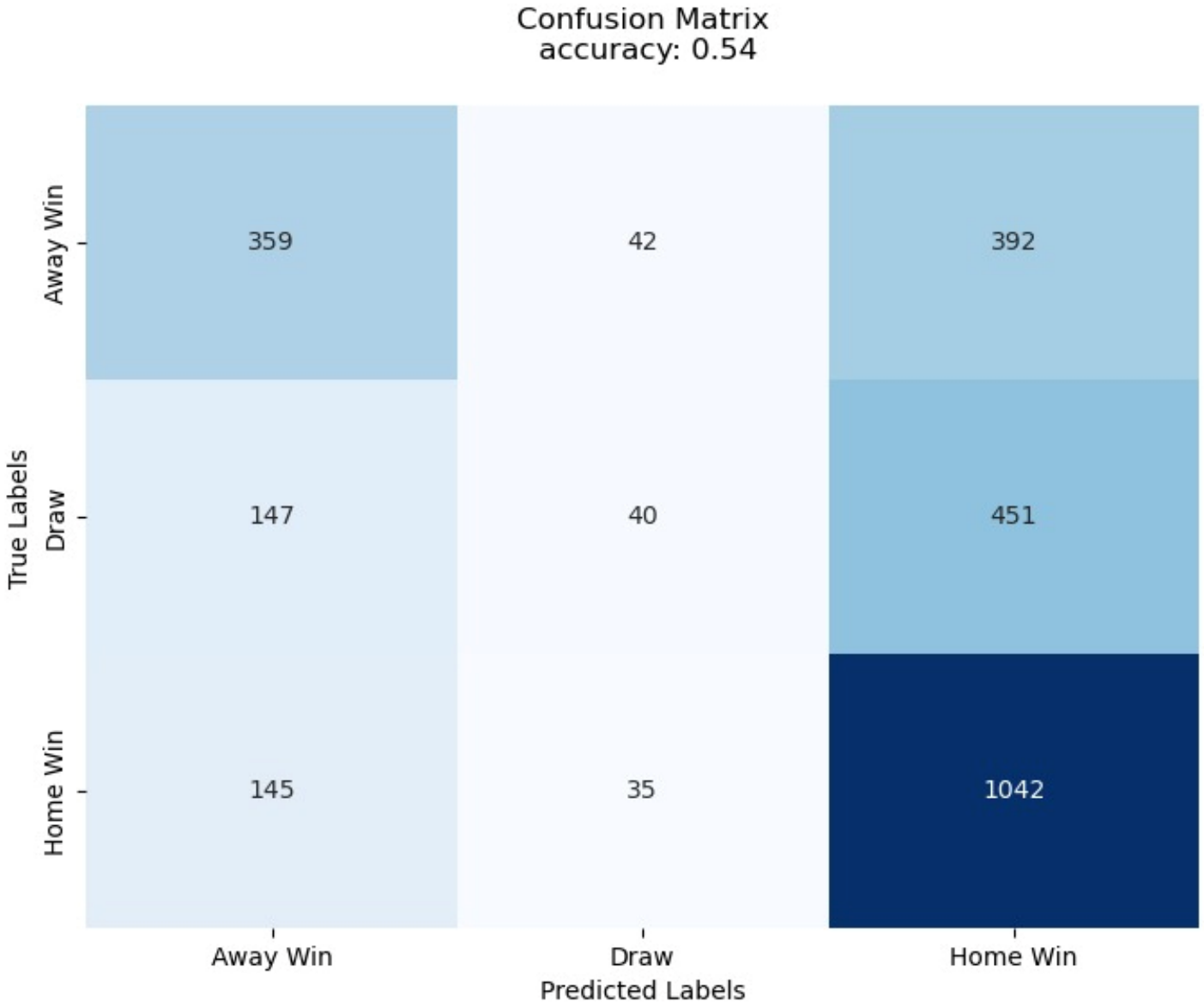
Hyperparameter optimization is **not possible** due to the high dimension of the data because of the creation of new columns from categorical data

VS

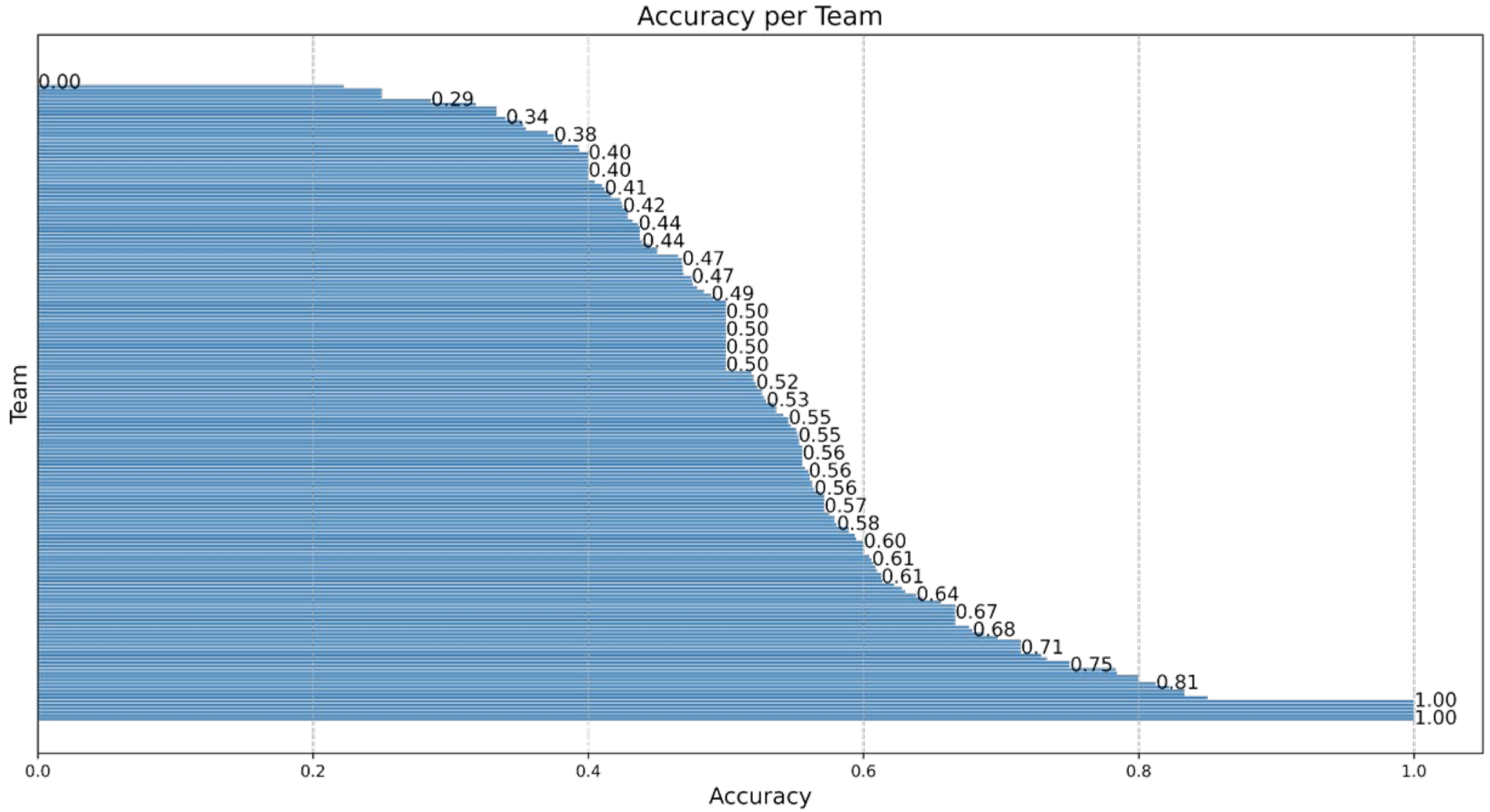
Using LabelEncodder...

We **can apply** hyperparameter optimization and we did that.

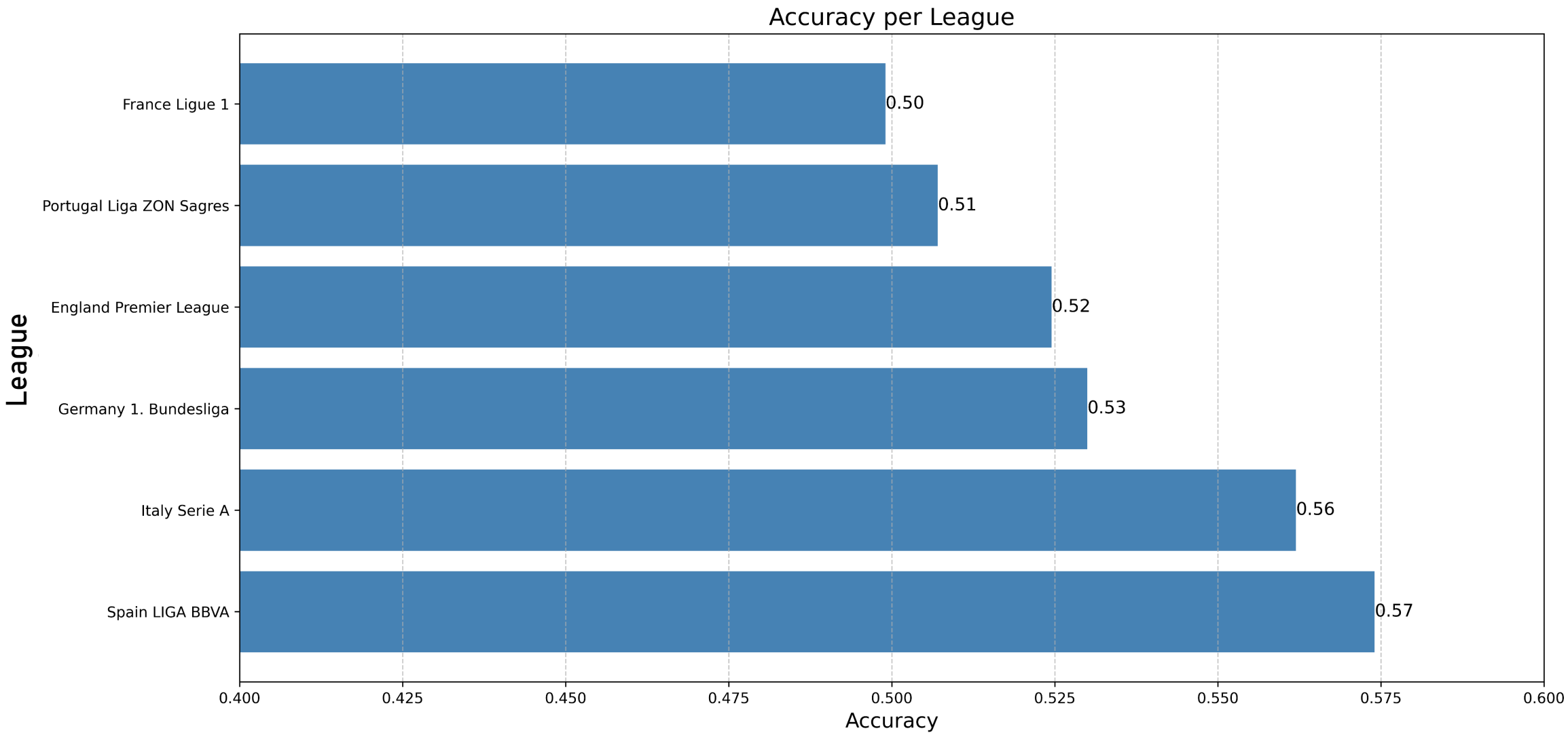
Final Results



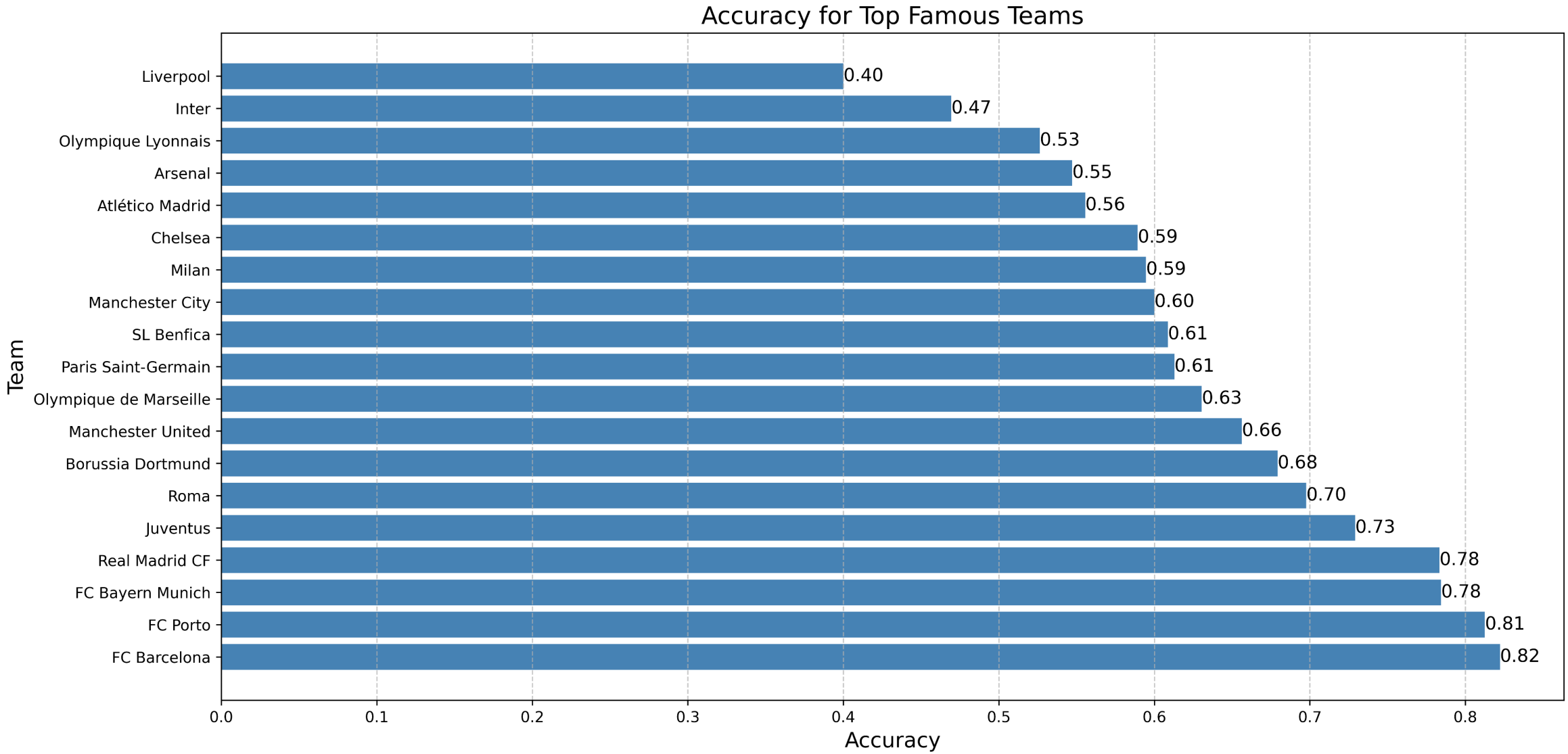
Model Accuracy For Each Team



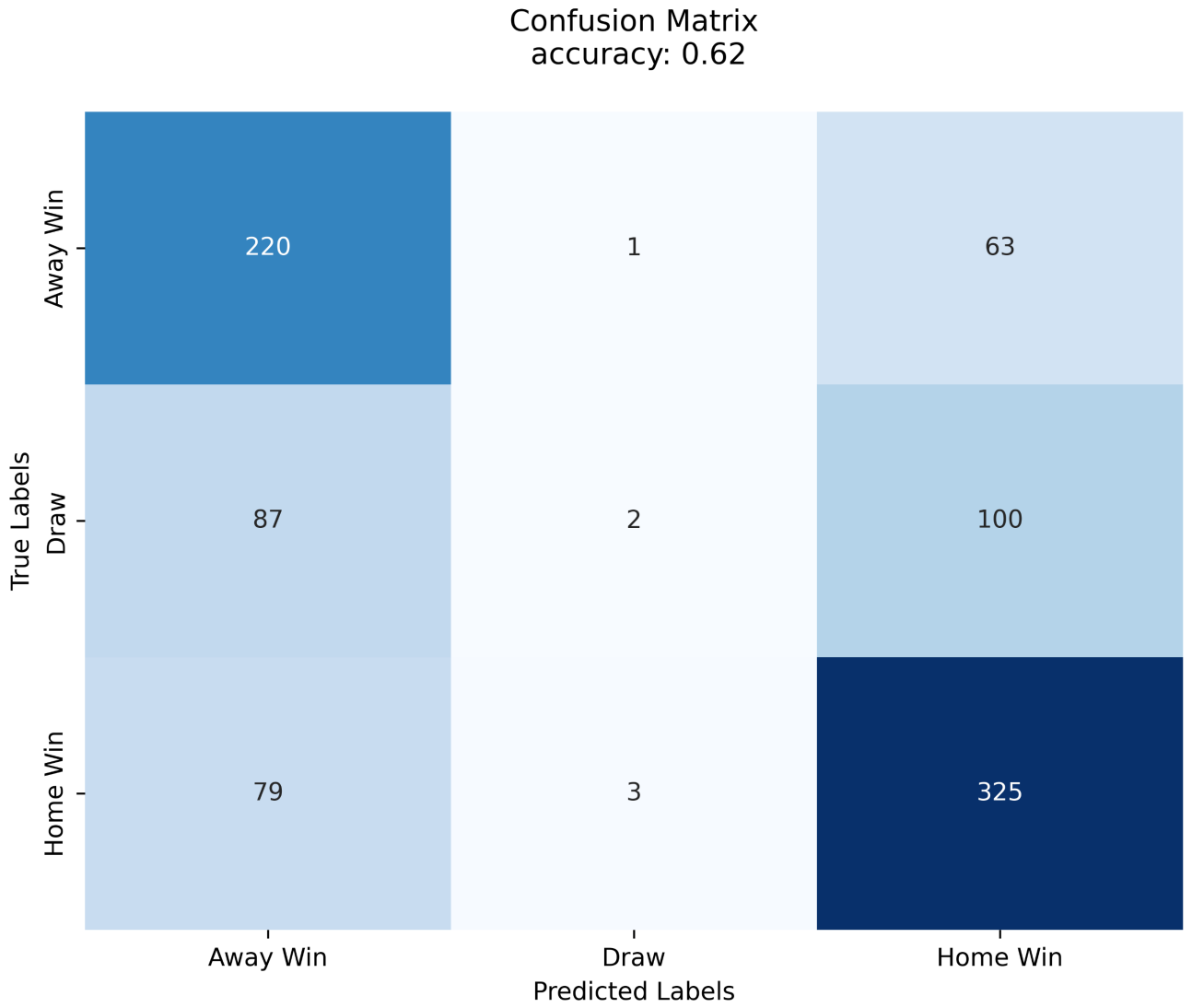
Highly Predictable League



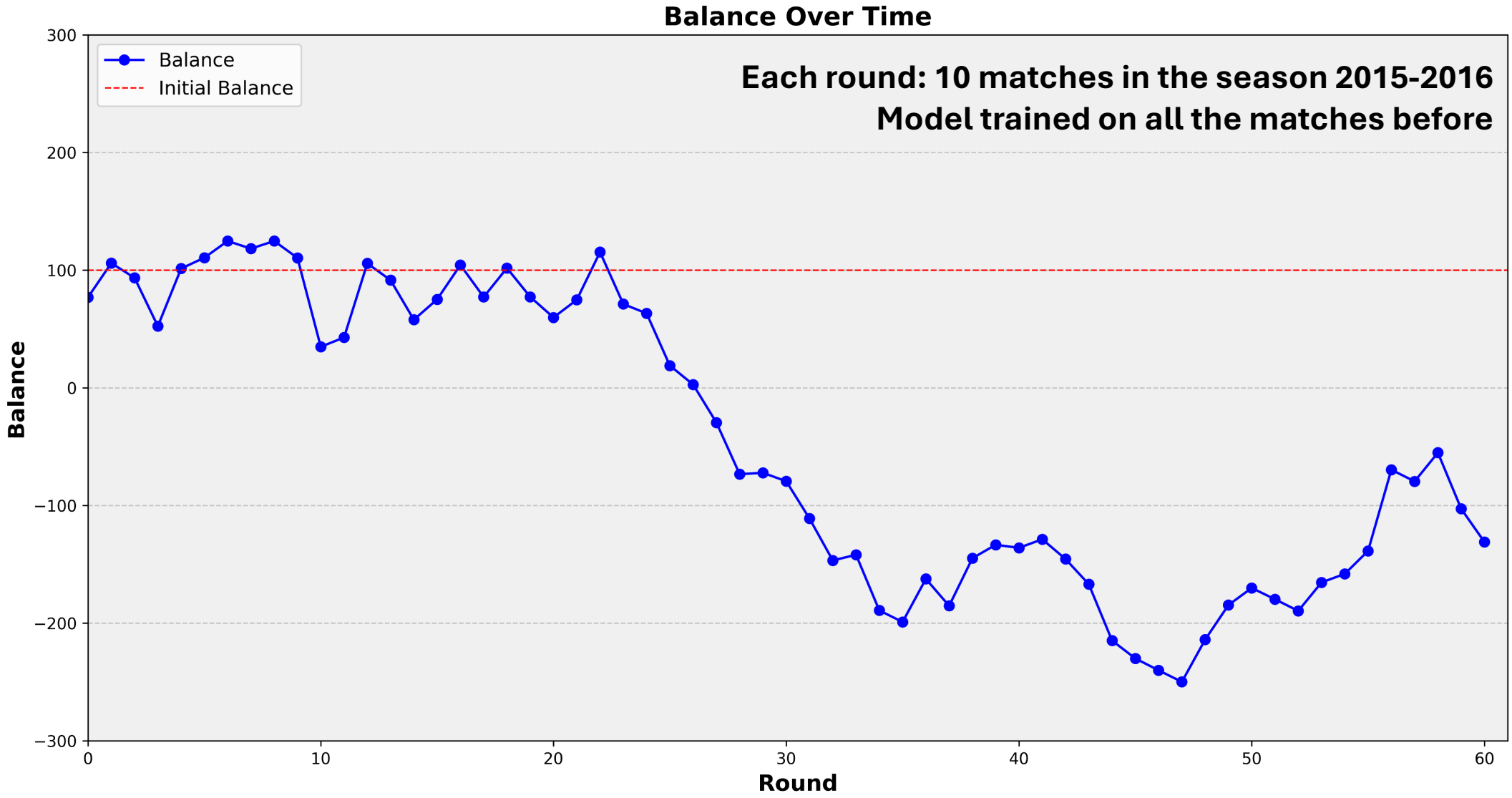
Model Accuracy For Well-Known Teams



Model Training on Famous Teams' Data



Model application in the prediction of last season's famous teams matches



Accuracy of Related Works

55 %

Match Outcome Predictions By Airback

53 %

Match Outcome Prediction By Pinnacle

52 %

Match Outcome Predictions By Samuel Benichou

48 %

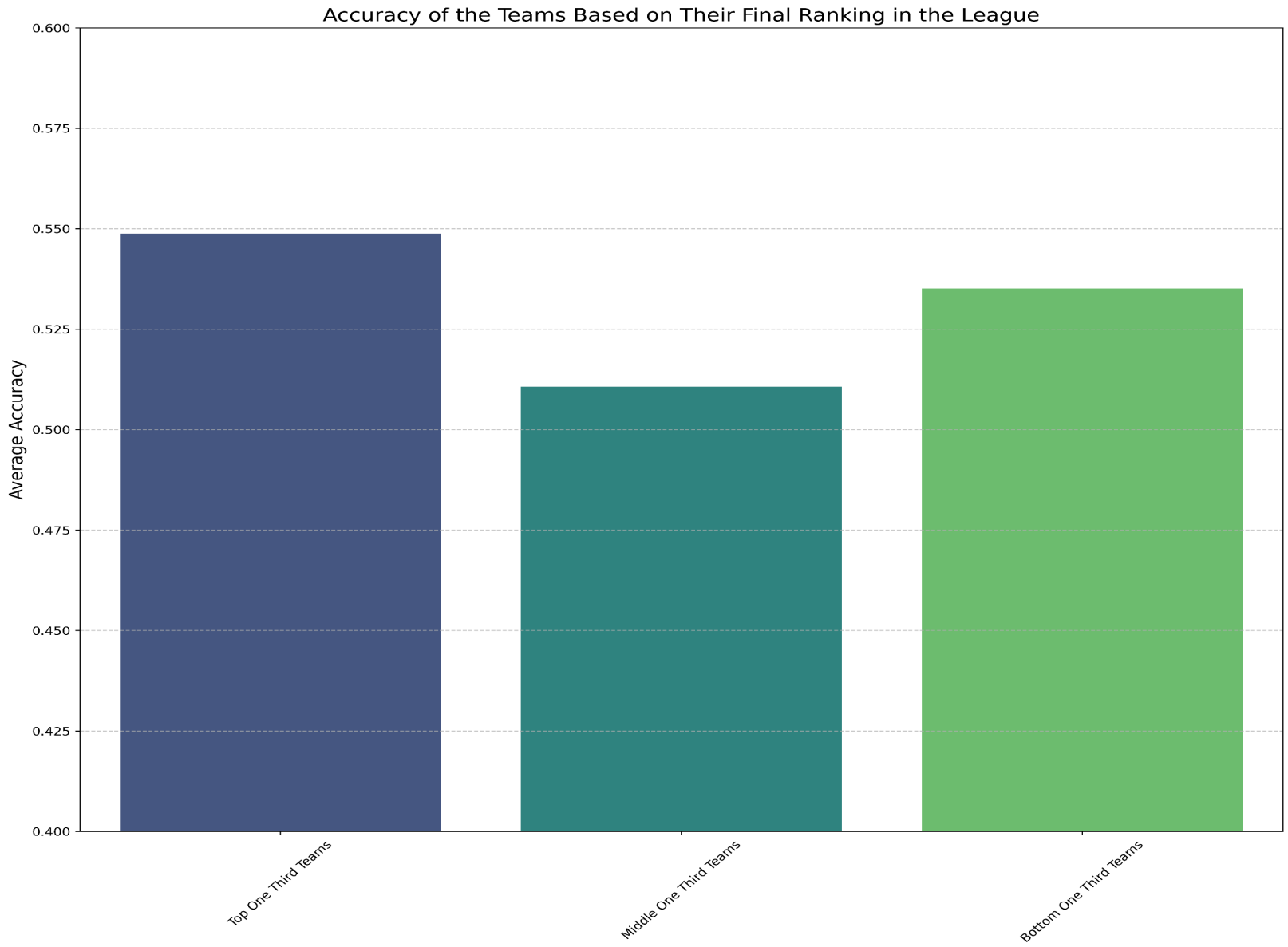
Football Predictions By MSOCZI

U OF T
University of Toronto
Department of Civil
and Mineral Engineering

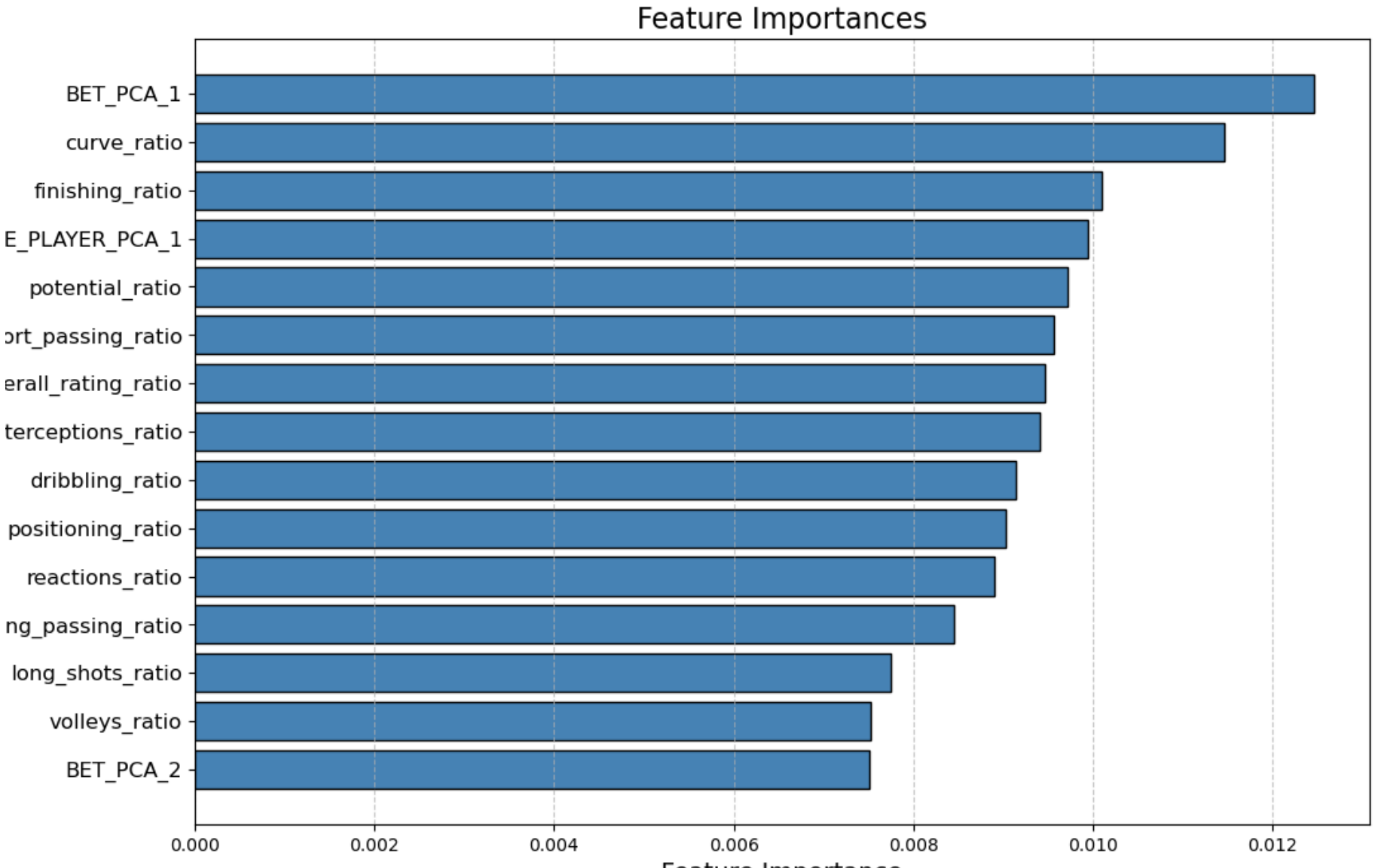
Thank you!

Fion Yang Ouyang, Mohammed Houssein Khosravi, Nanqiao Du & Nyah Bay

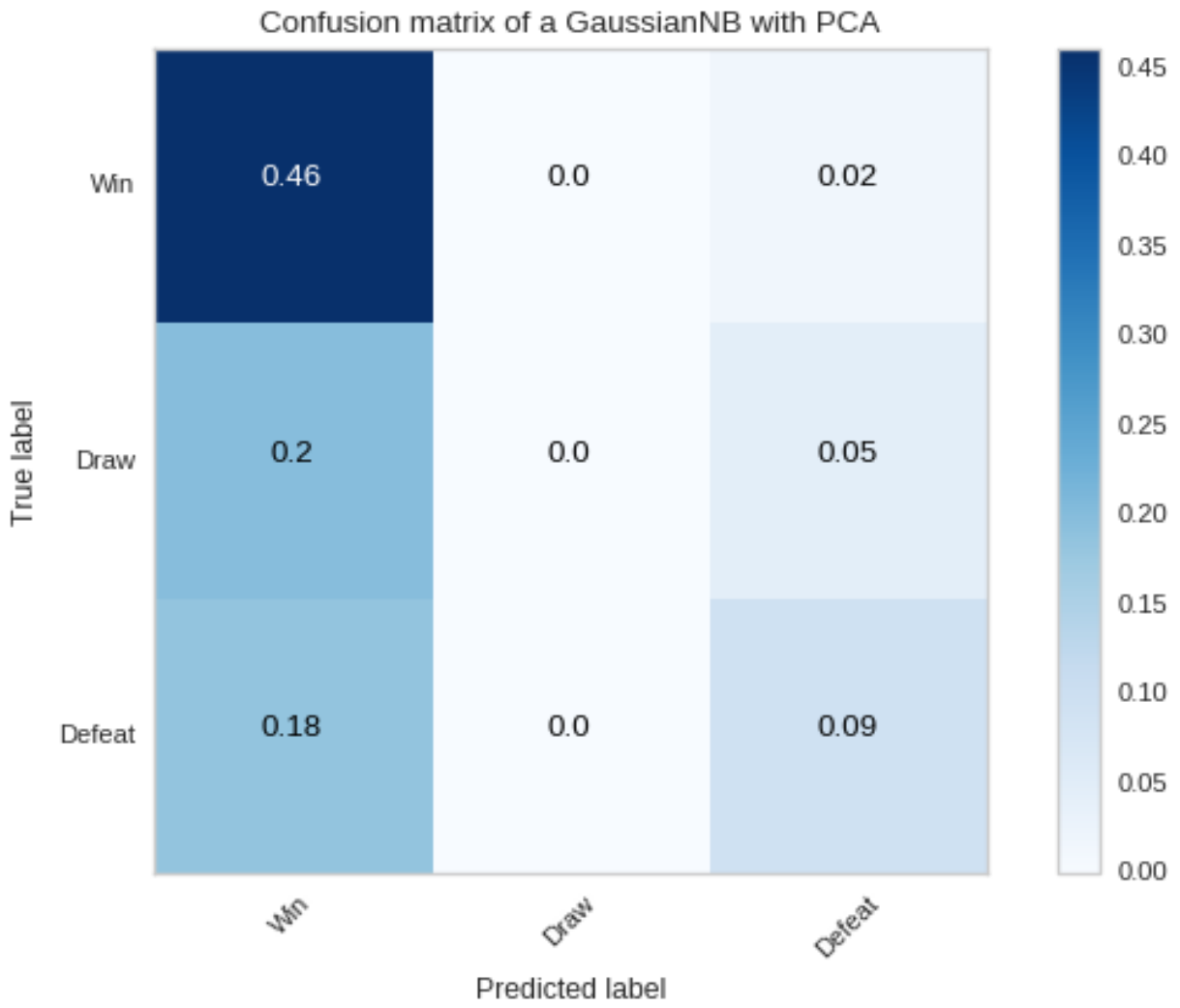
Accuracy of the teams based on their final ranking in the league



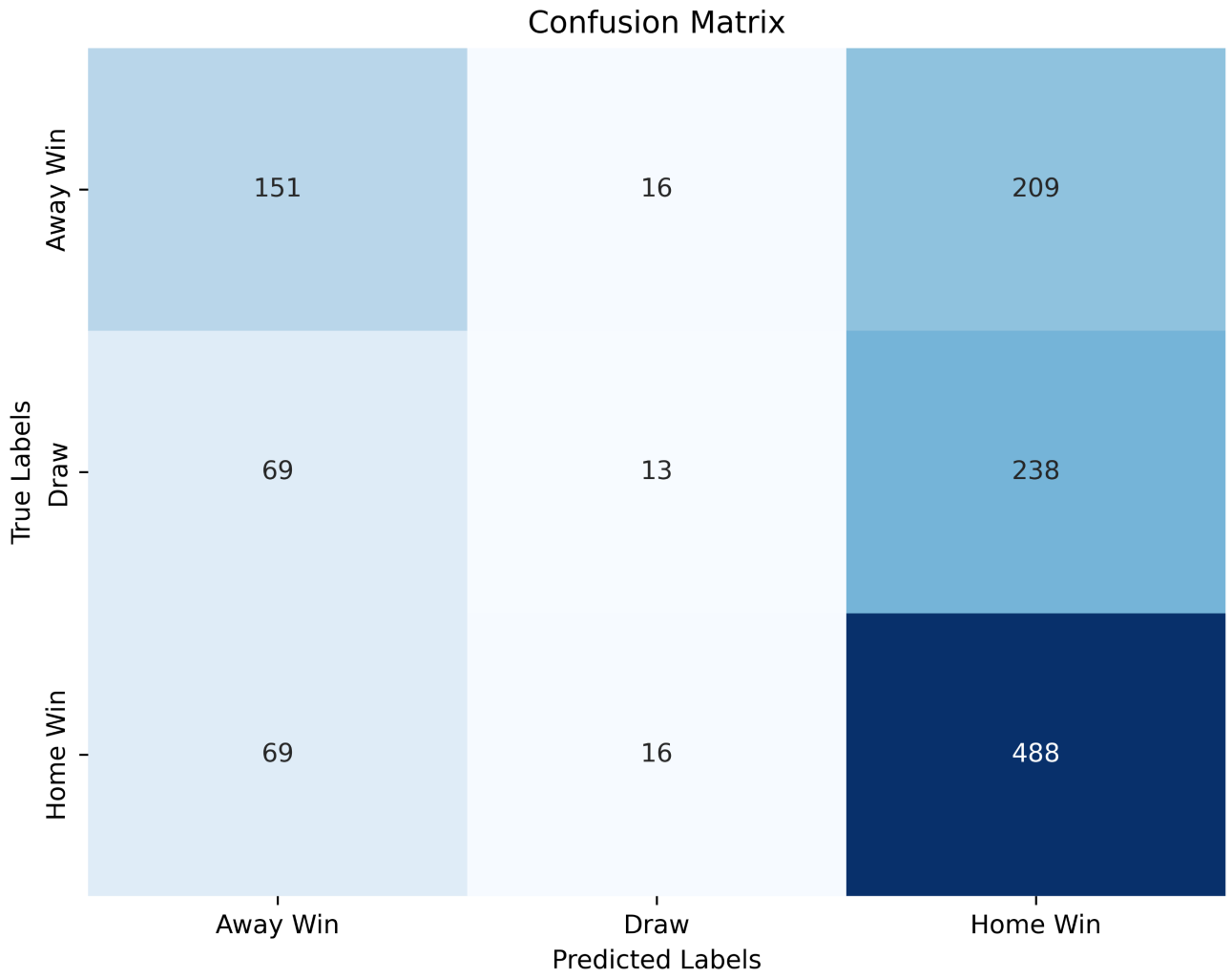
Feature importances



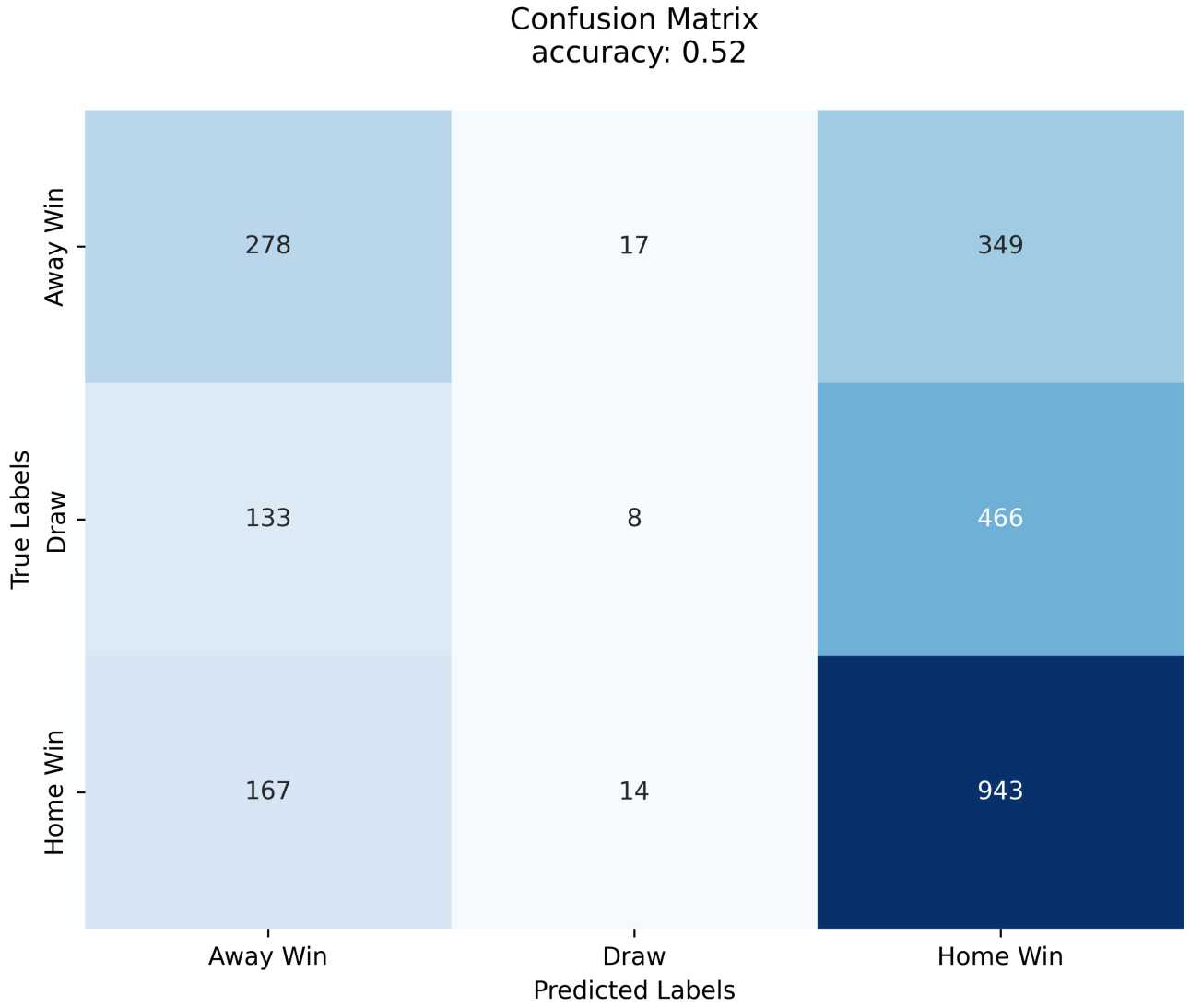
Confusion matrix of another project worked on this data



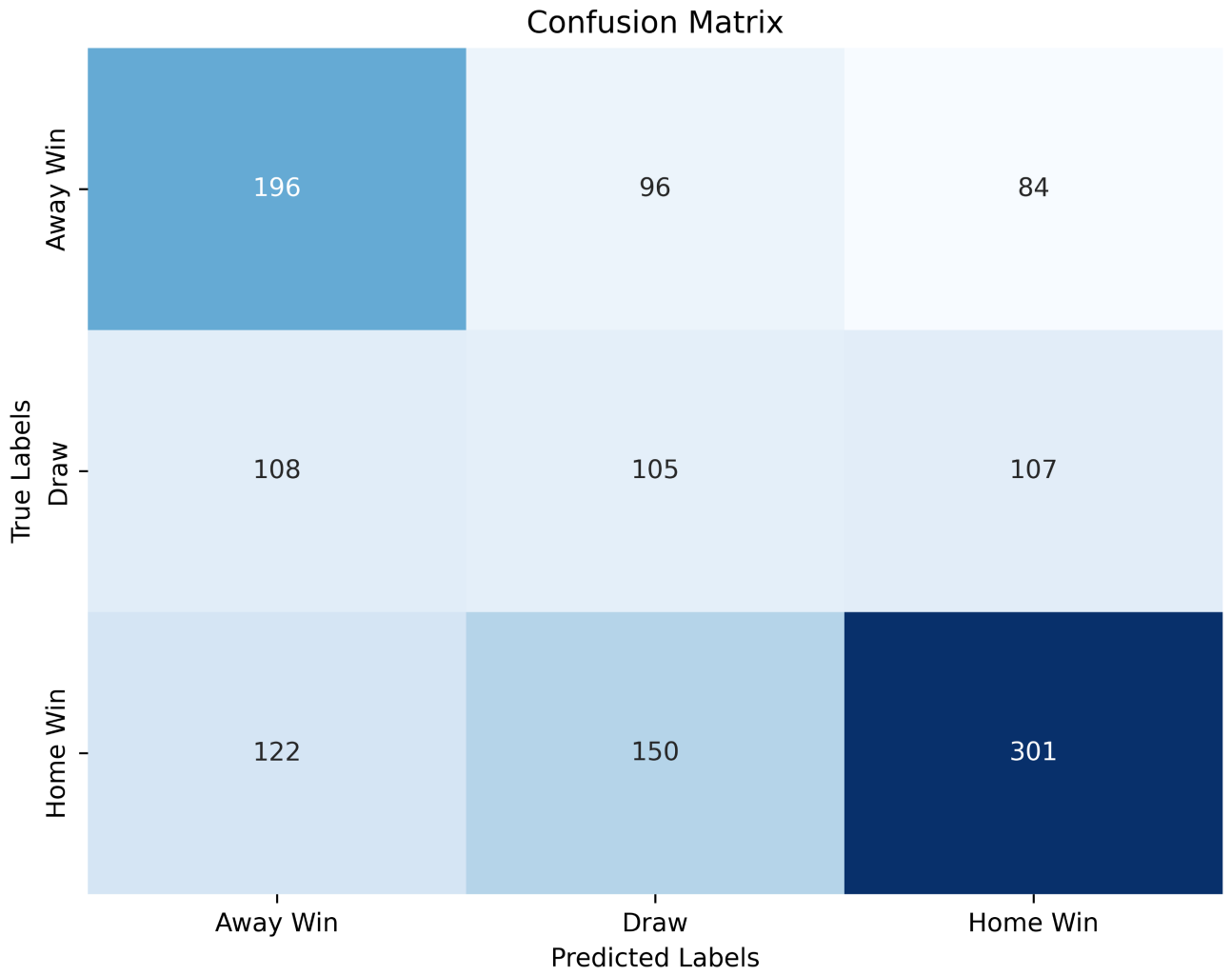
Model training on middle-class data

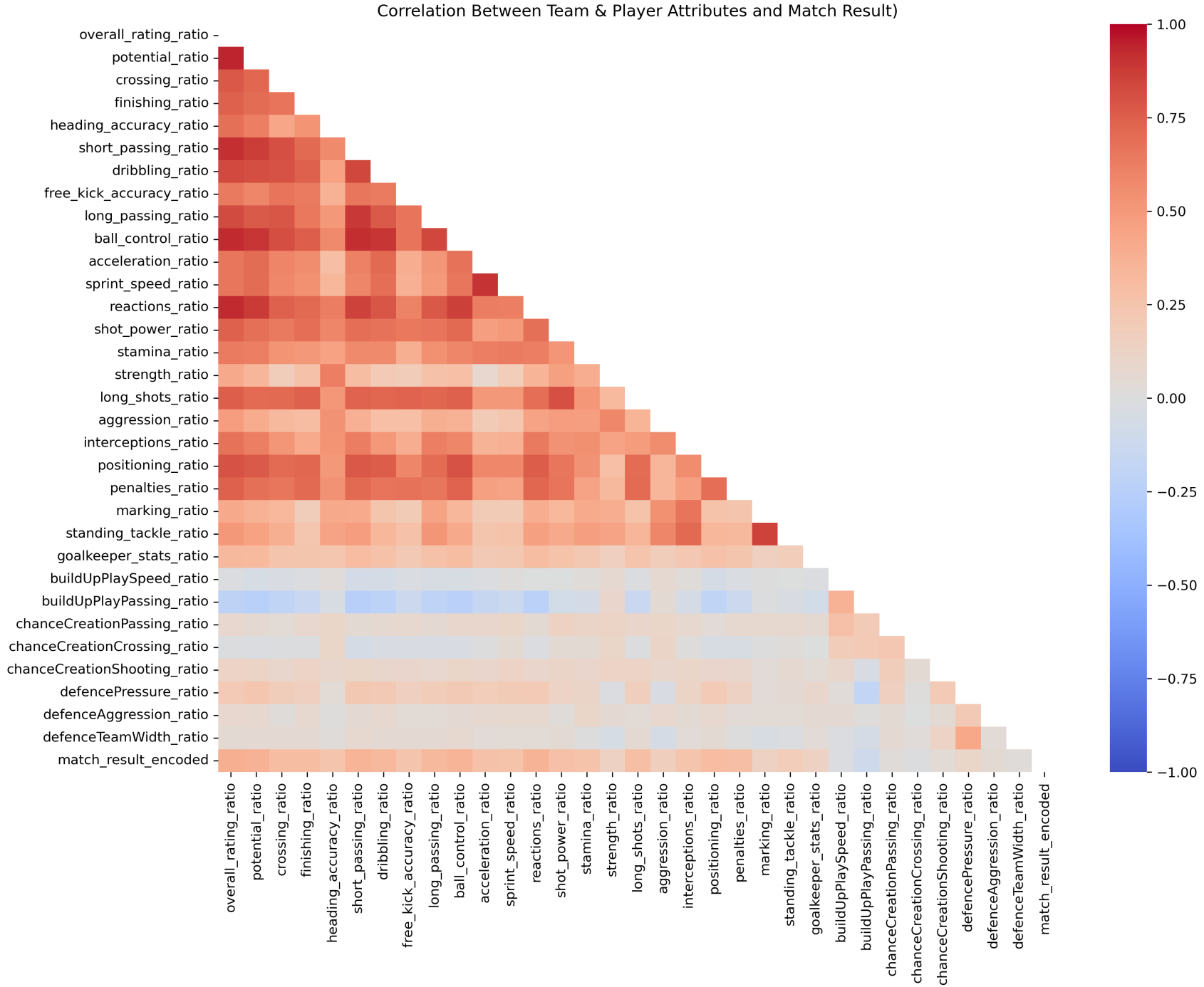


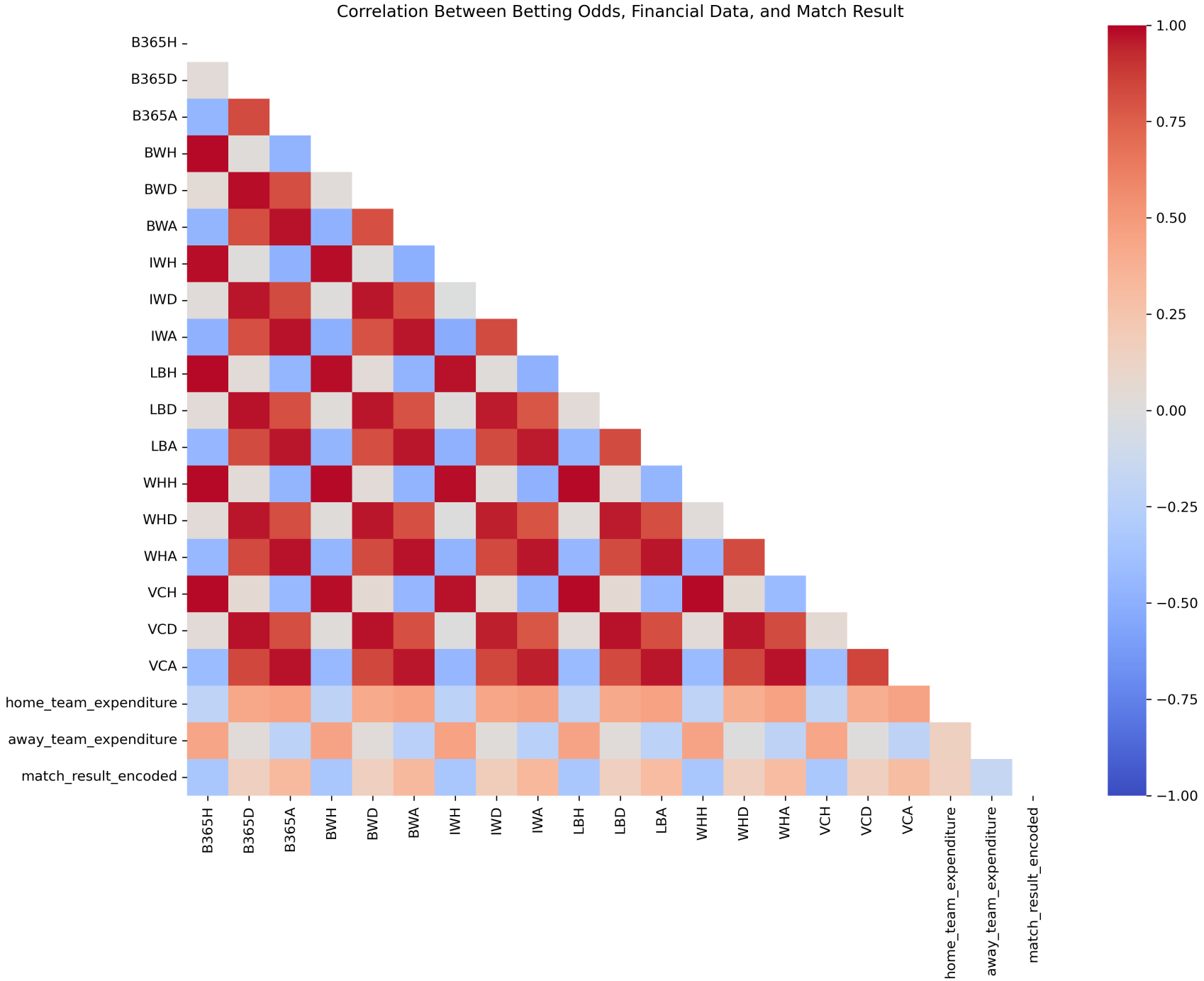
Model training on top-10%-excluded data



Model training on under sampled data




Correlation Between Team & Player Attributes and Match Result)
overall_rating_ratio
potential_ratio
crossing_ratio
finishing_ratio
heading_accuracy_ratio
short_passing_ratio
dribbling_ratio
free_kick_accuracy_ratio
long_passing_ratio
ball_control_ratio
acceleration_ratio
sprint_speed_ratio
reactions_ratio
shot_power_ratio
stamina_ratio
strength_ratio
long_shots_ratio
aggression_ratio
interceptions_ratio
positioning_ratio
penalties_ratio
marking_ratio
standing_tackle_ratio
goalkeeper_stats_ratio
buildUpPlaySpeed_ratio
buildUpPlayPassing_ratio
chanceCreationPassing_ratio
chanceCreationCrossing_ratio
chanceCreationShooting_ratio
defencePressure_ratio
defenceAggression_ratio
defenceTeamWidth_ratio
match_result_encoded
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Correlation Between Betting Odds, Financial Data, and Match Result
B365H
B365D
B365A
BWH
BWD
BWA
IWH
IWD
IWA
LBH
LBD
LBA
WHH
WHD
WHA
VCH
VCD
VCA
home_team_expenditure
away_team_expenditure
match_result_encoded
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00