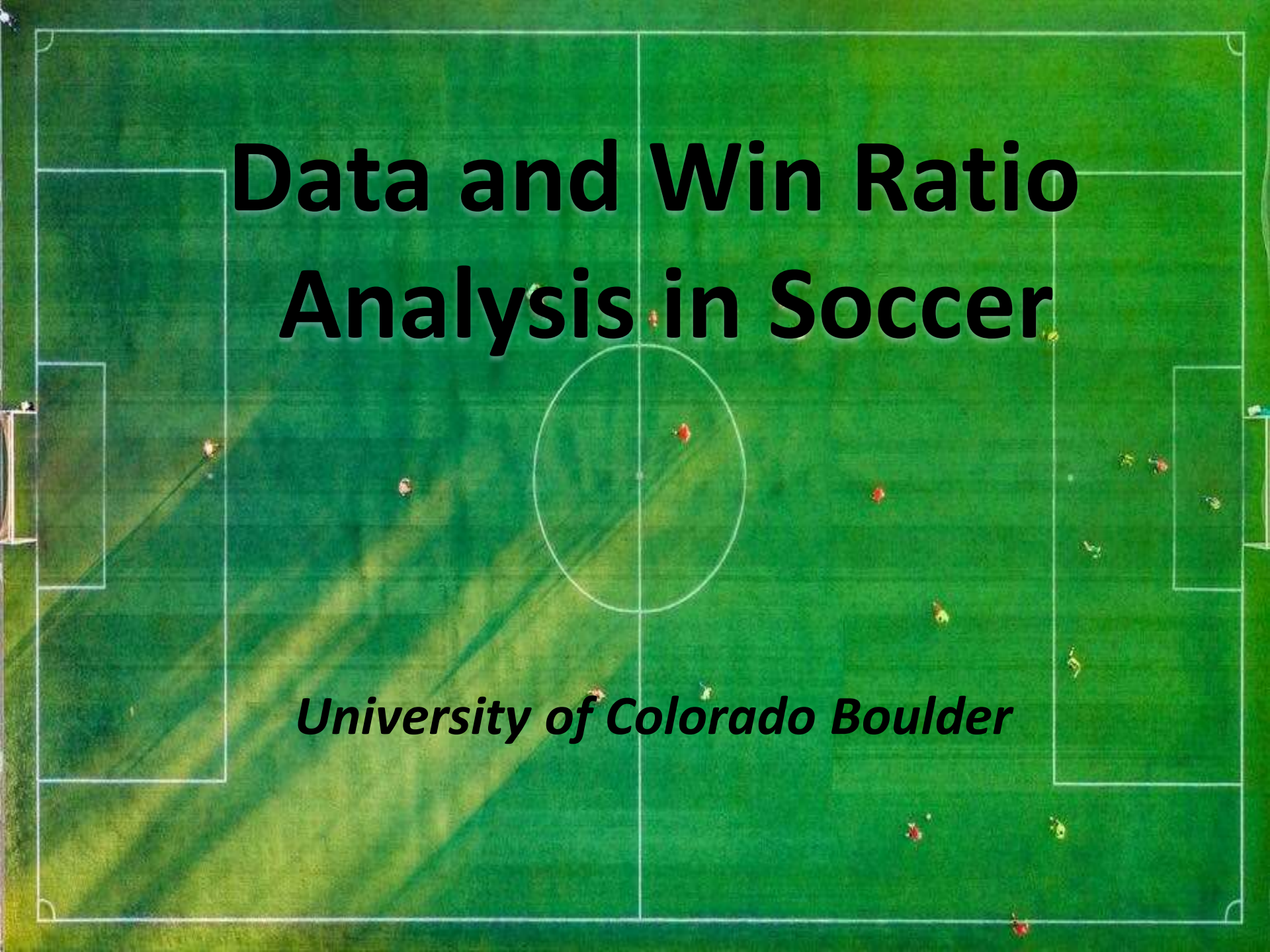


Data and Win Ratio Analysis in Soccer

University of Colorado Boulder

Problem Statement

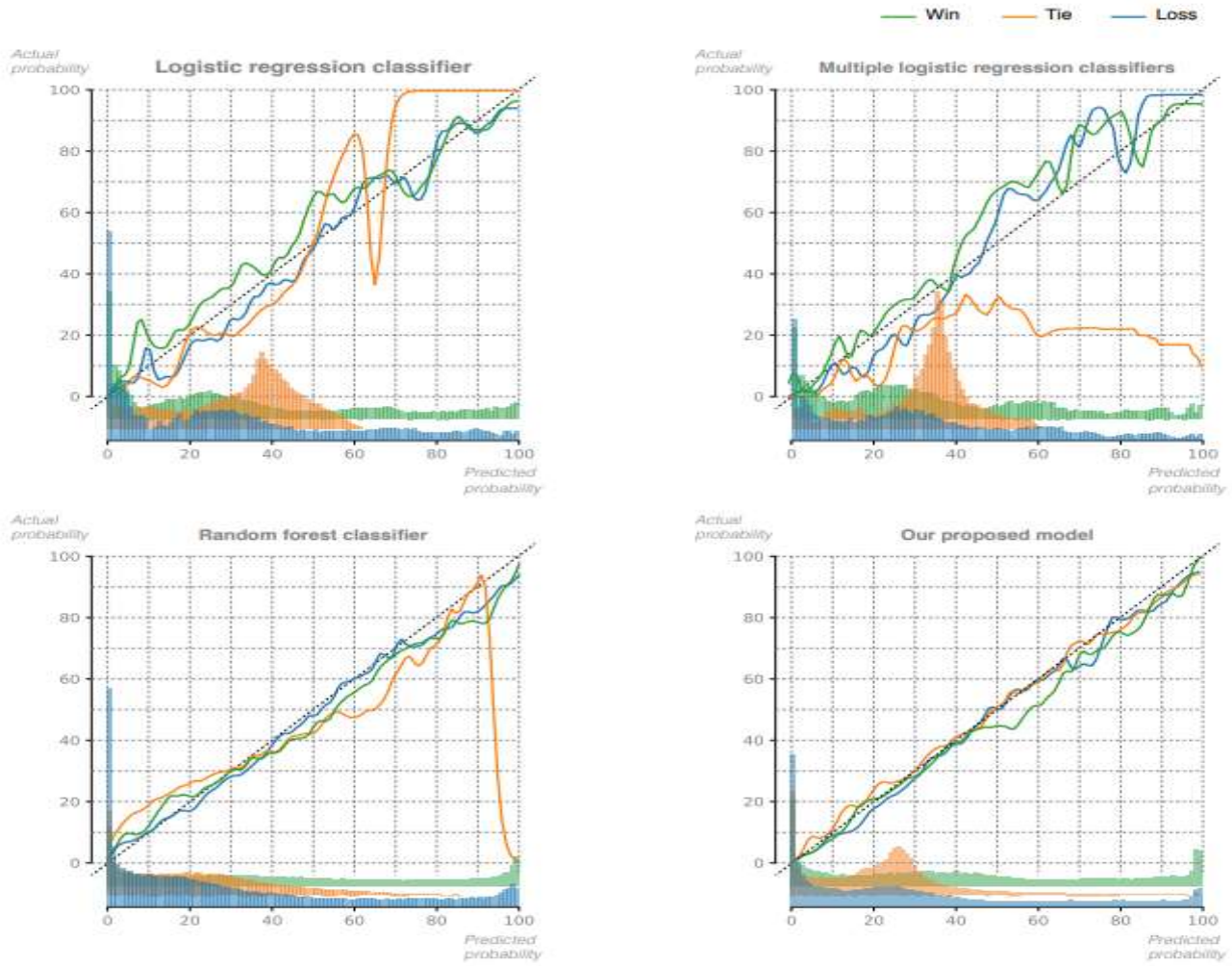
- *A vast amount of data generated from sports*
  - player skills
  - game results
  - seasonal performance
  - league management

- *The challenge for sports science*
  - analyzing this data to gain a competitive edge
  - various techniques and statistical methods to extrapolate valuable insights
  - result predictions

Related Work

- *A Bayesian Approach to In-Game Win Probability in Soccer*
  - Bayesian Model
  - predicts the result of future soccer matches
    - rank position
    - historical data

Bayesian model clearly outperforms the LR, mLR and RF models



Related Work

- *Application of the Apriori Algorithm in Soccer Games*
  - Apriori algorithm
  - around the concept of "scoring opportunities“:
    - scoring opportunities
    - individual players
    - frequent player combinations.

Proposed Work

- Tools
  - Python
  - SQL
- Datasets
  - European Soccer Database
- Tasks
  - Visualization
  - analyzing data
  - build result predictions
- Techniques
  - Cluster analysis
  - Bayesian Model
  - Apriori algorithm

Evaluation

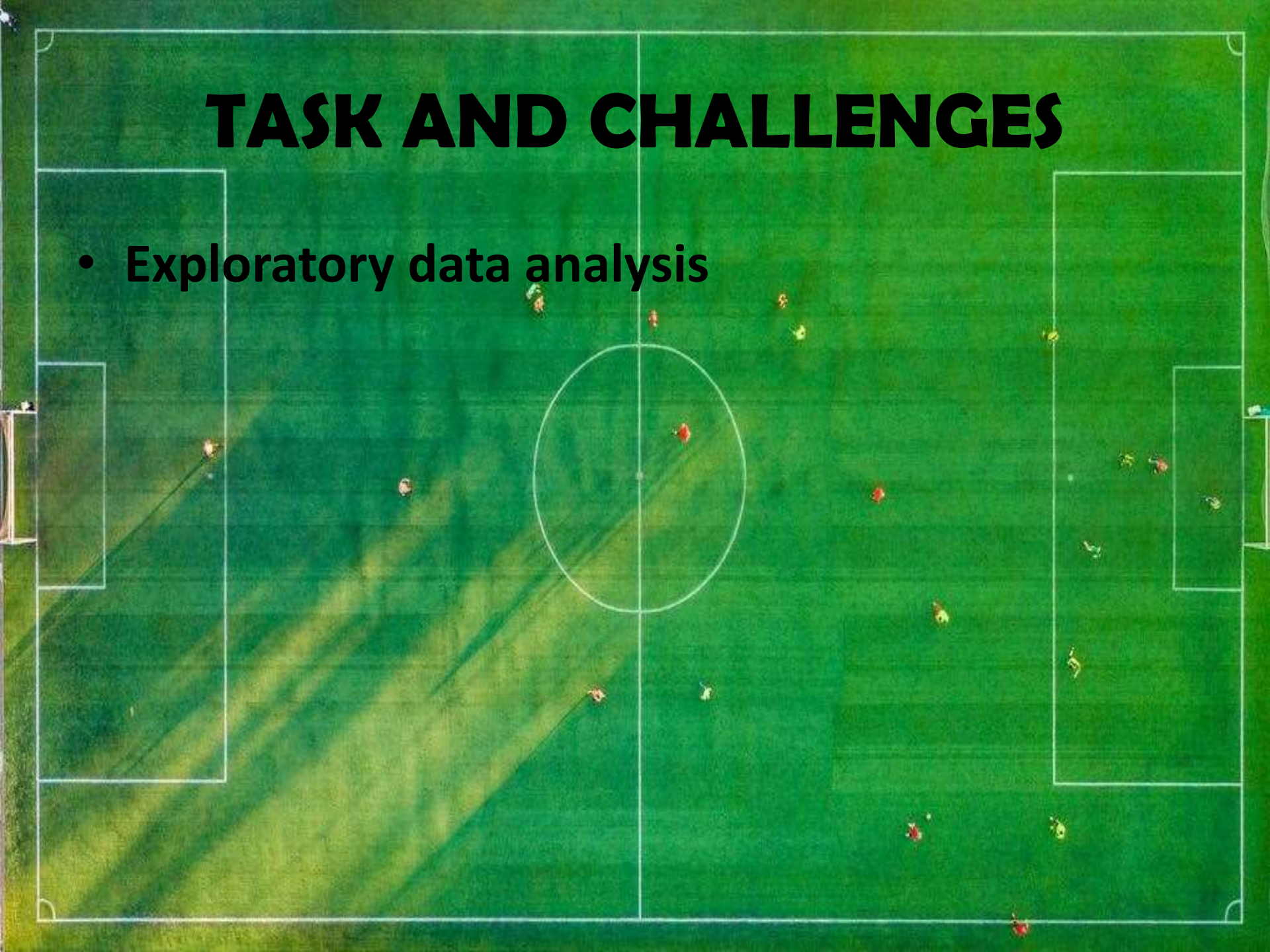
- **Model Validation**
  - test it against data
  - the effectiveness
  - the accuracy
- **Prediction Accuracy**
  - comparing with actual outcomes
  - win ratio of the outcome between the actual outcome
- **Improvements Over Time**
  - progressive improvement in prediction over time
- **Comparison With Other Models:**
  - assess effectiveness
  - compared with other similar models
  - standard statistical analysis

Timeline

- *Week 2 Project Proposal*
  - *Finished*
- *Week 3 Checking Project*
  - *Finished*
- *Week 4*
  - *Finished*

TASK AND CHALLENGES

- Exploratory data analysis

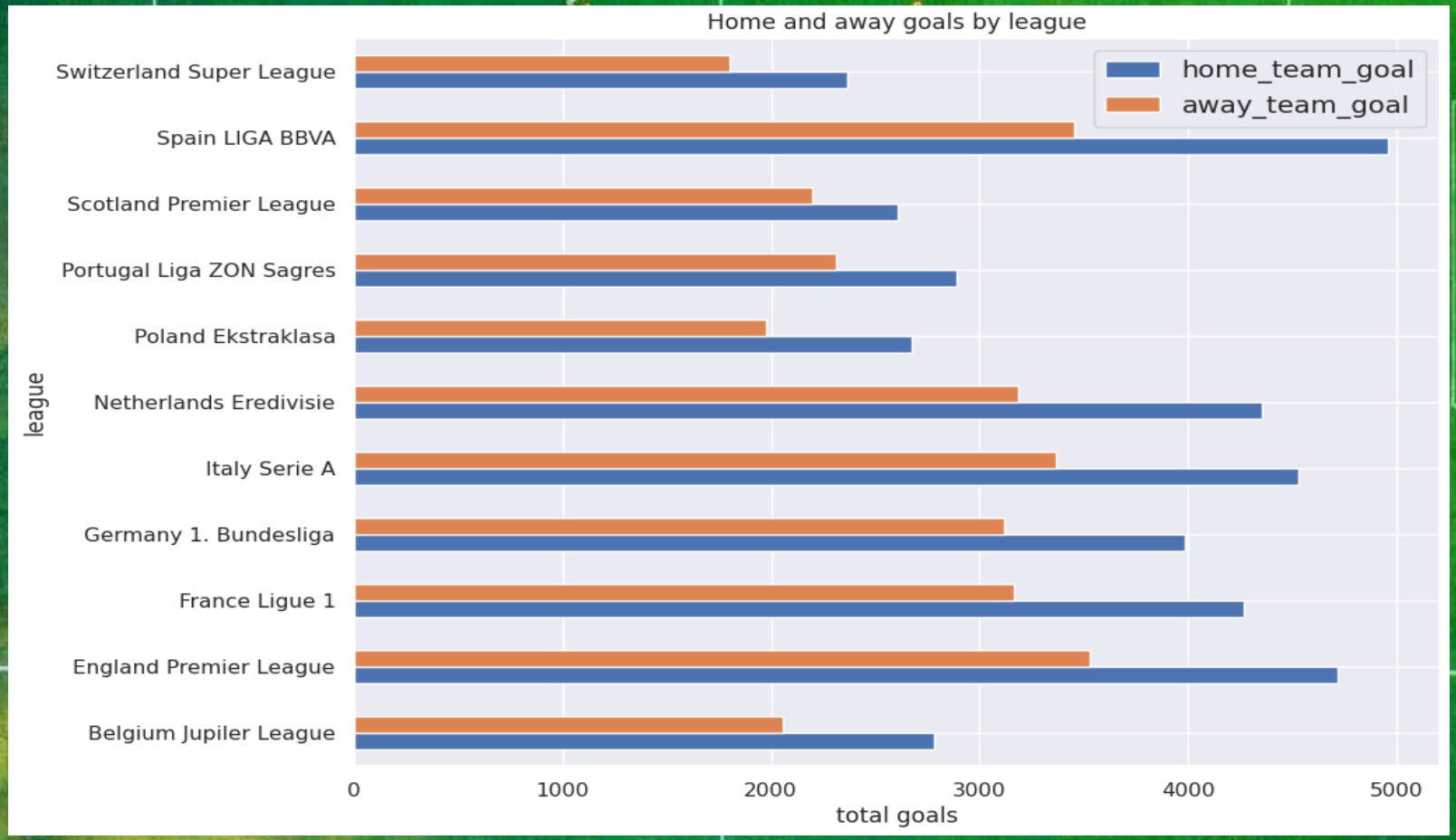


Exploratory data analysis

- **Matches by League**

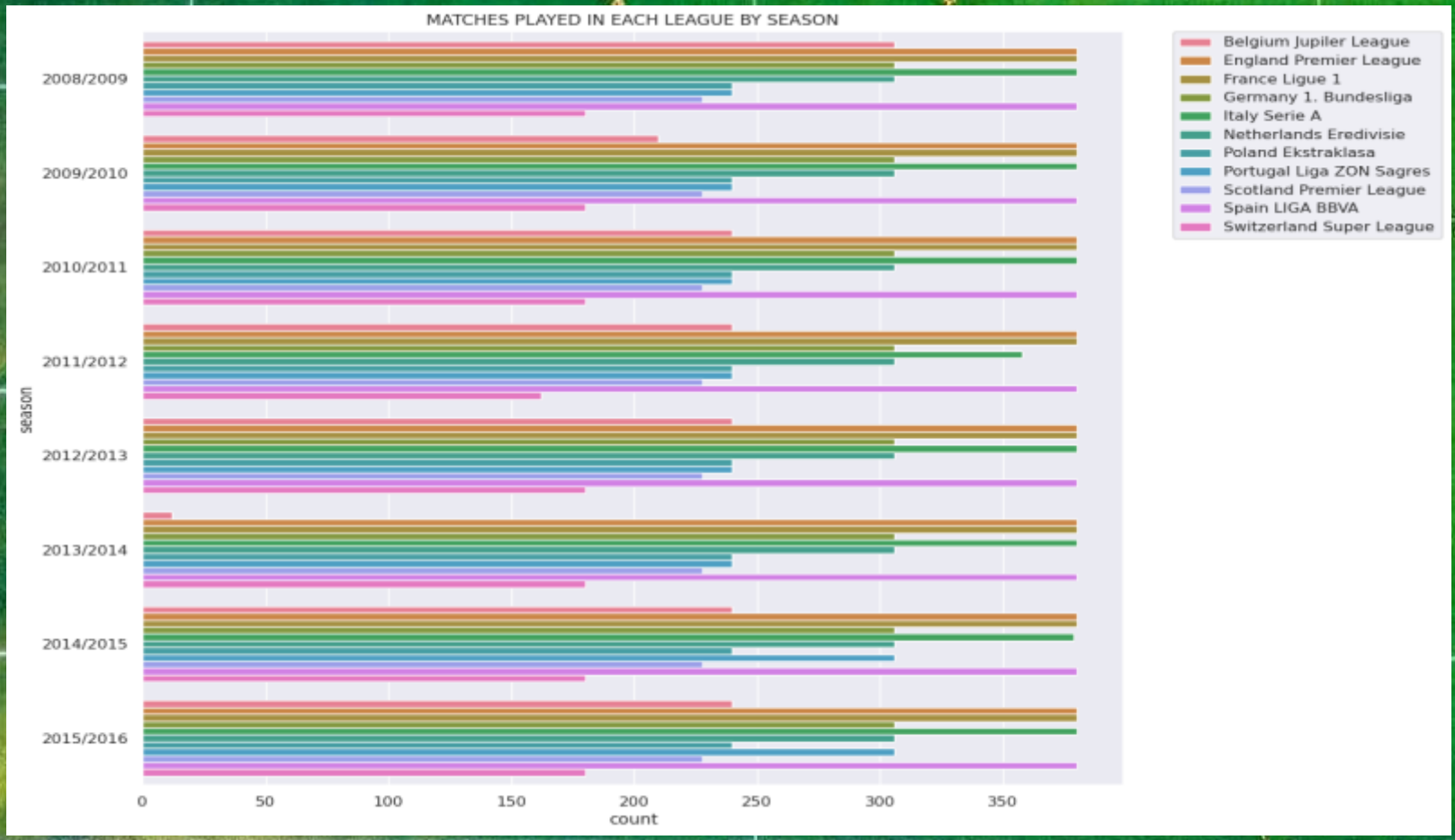
Exploratory data analysis

- **Matches by League**



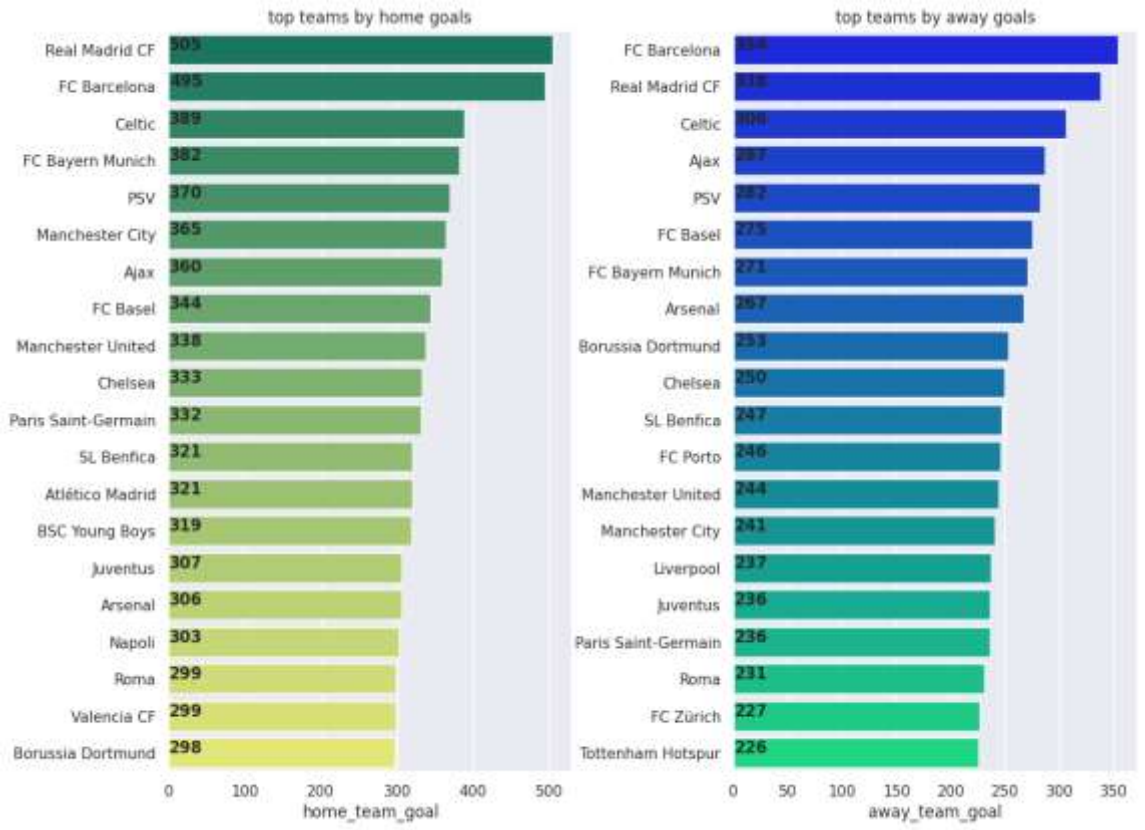
Exploratory data analysis

- **Matches by League**



Exploratory data analysis

- **Matches by League**



Exploratory data analysis

- **Correlation analysis**

Bayesian data analysis

- distribution of future goals:

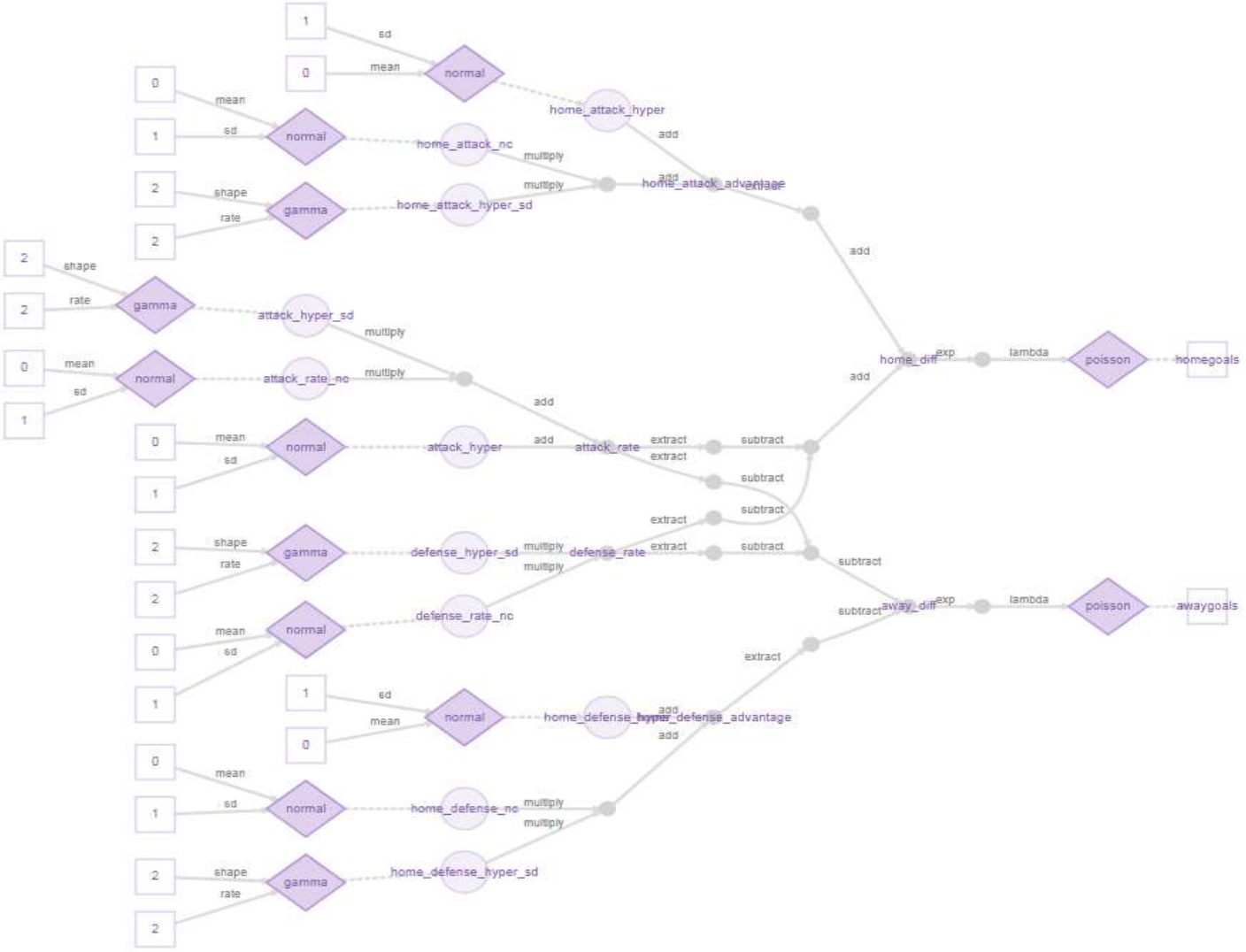
$$homegoals \sim Poisson(\exp(homediff))$$
$$awaygoals \sim Poisson(\exp(awaydiff))$$

Bayesian data analysis

- **Likelihood:**

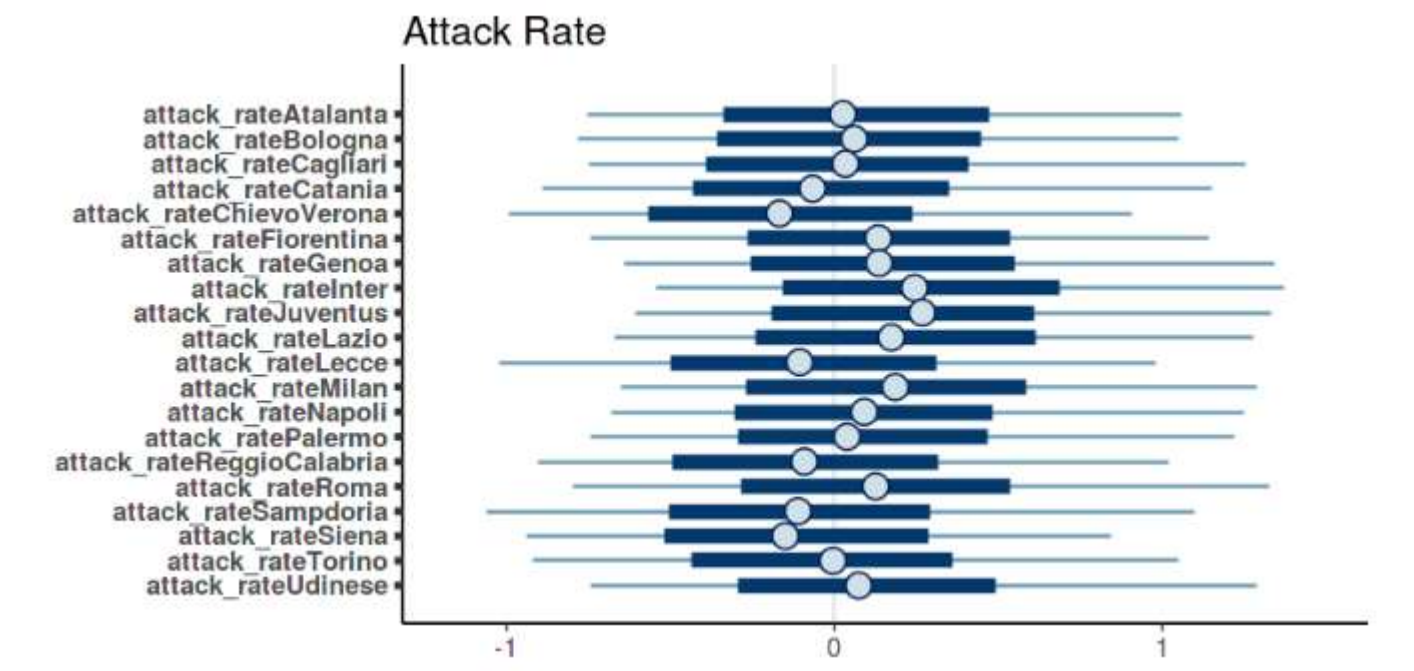
$$
\begin{aligned}
homediff &= AttRate_{home} - DefRate_{away} + HomeAttAdvantage_{home} \\
awaydiff &= AttRate_{away} - DefRate_{home} - HomeDefAdvantage_{home} \\
AttRate_i &= BaseAtt + AttRate_{i,non-centered} * \tau_i^{att} \\
DefRate_i &= DefRate_{i,non-centered} * \tau_i^{def} \\
HomeDefAdvantage_{home} &= BaseHomeAtt + HomeAttRate_{i,non-centered} * \tau_i^{HomeAtt} \\
HomeDefAdvantage_{home} &= BaseHomeDef + HomeDefRate_{i,non-centered} * \tau_i^{HomeDef} \\
BaseAtt &\sim Normal(0,1) \\
AttRate_i &\sim Normal(0,1) \\
DefRate_i &\sim Normal(0,1) \\
BaseHomeAtt &\sim Normal(0,1) \\
HomeAttRate_i &\sim Normal(0,1) \\
BaseHomeDef &\sim Normal(0,1) \\
HomeDefRate_i &\sim Normal(0,1) \\
\tau_i^{att} &\sim Gamma(2,2) \\
\tau_i^{def} &\sim Gamma(2,2) \\
\tau_i^{HomeAtt} &\sim Gamma(2,2) \\
\tau_i^{HomeDef} &\sim Gamma(2,2)
\end{aligned}
$$

Bayesian model



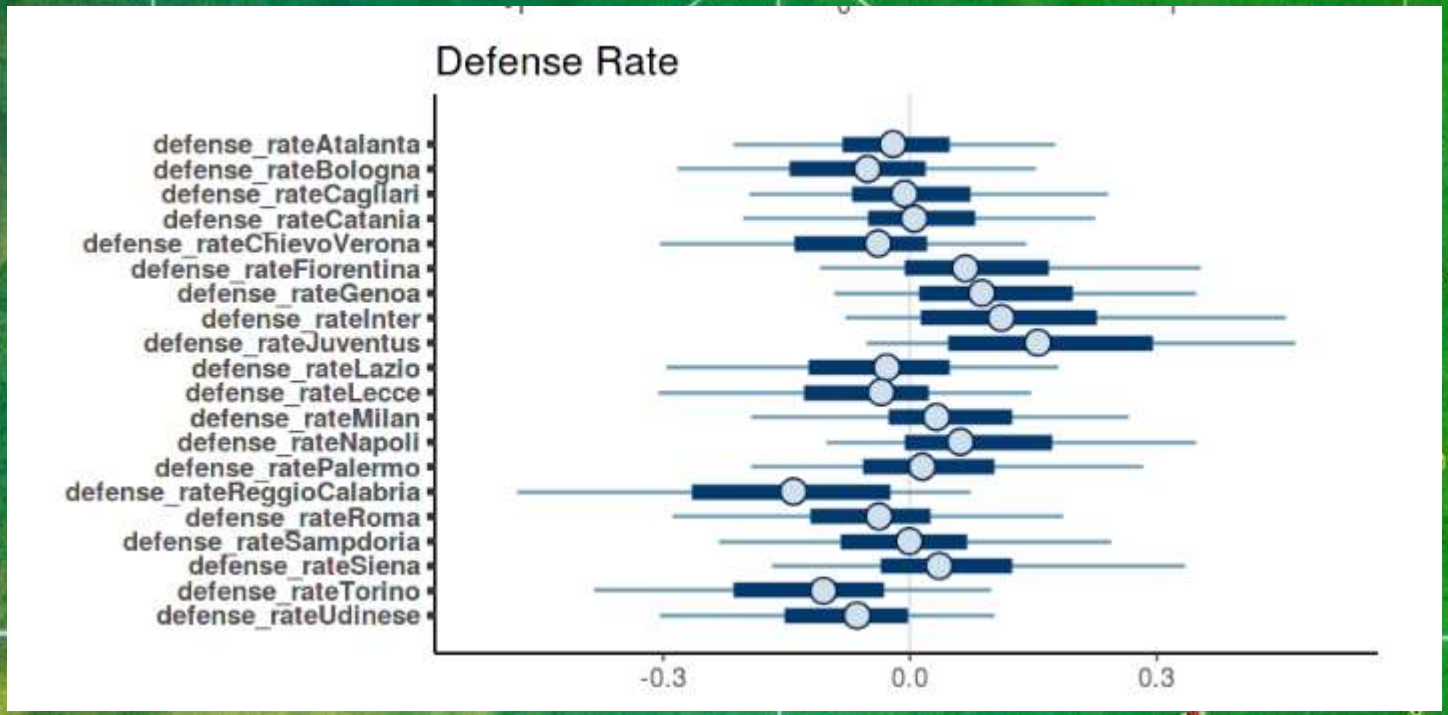
Evaluation and feature importance

- base features



Evaluation and feature importance

- base features



CONCLUSIONS

- Bayesian model utilizes team-specific features and predicts the future number of goals scored
- requires only a relatively small amount of data to learn effectively.
- In the era of Big Data, where not all datasets are large, this technique offers a reliable way to create robust models
- Bayesian multilevel modeling a powerful and valuable tool for sports analytics