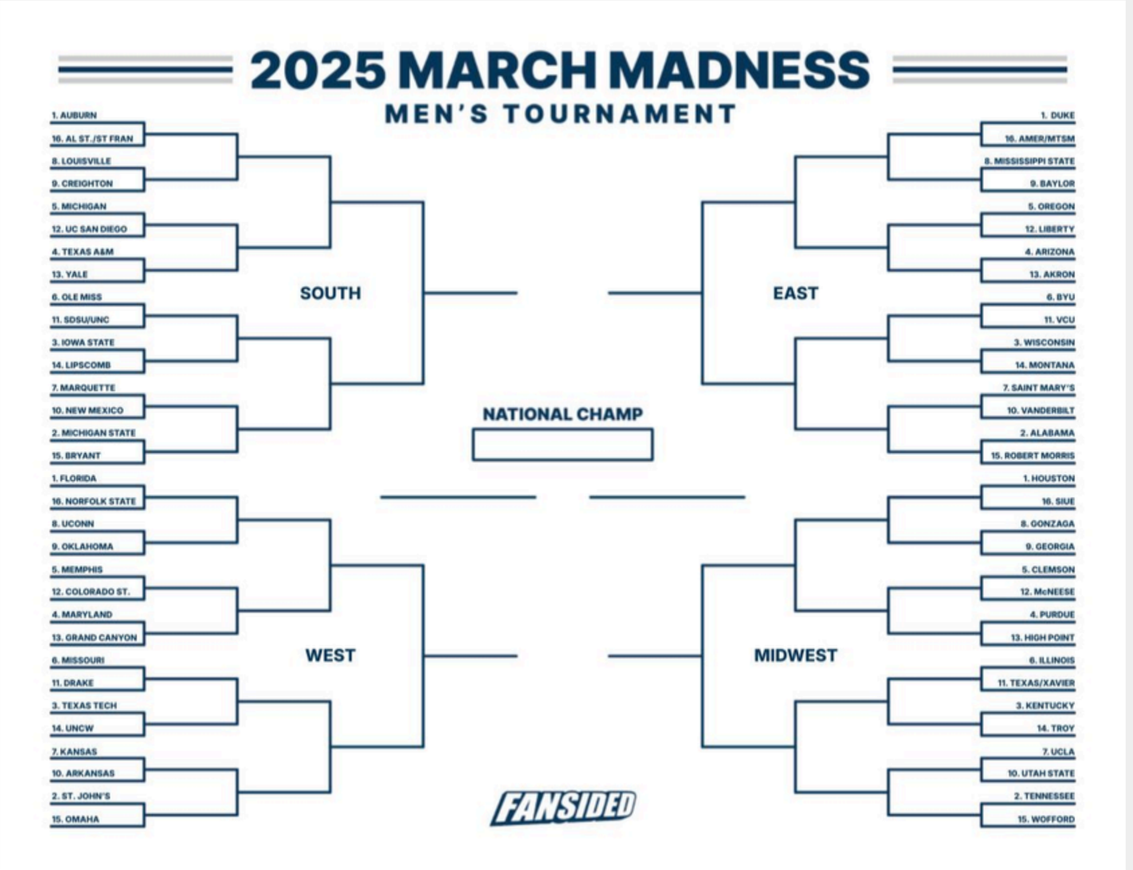
NCAA March Madness Analysis & Prediction

CS 670 Data Science Project

Topic Background

- NCAA Basketball = College Basketball
- March Madness is a series of single elimination games to determine the winner team out of 68 that are selected based on the preceding stage

- **March Machine Learning Mania 2025 - kaggle competition designed for predicting the win probability for every possible matchup during the tournament**



Data

- Team Seeds
- Public Rankings
- Game Locations (City, State)
- Coach Data
- Regular Season game outcomes and scores
- Team Conferences
- History of all NCAA Tournaments - games and scores:
  - Men since 1985
  - Women since 1998

Goal

- Predict win probablity for every possible matchup in the Season
  - metric: Brier score = (Actual result − Forecast Probability)
- Data questions?
  - Are there any notable differences between Men's and Women's data?
  - Does the Coach have a significant impact on the result?
  - **What will be the most impactful features?**

Brier Score

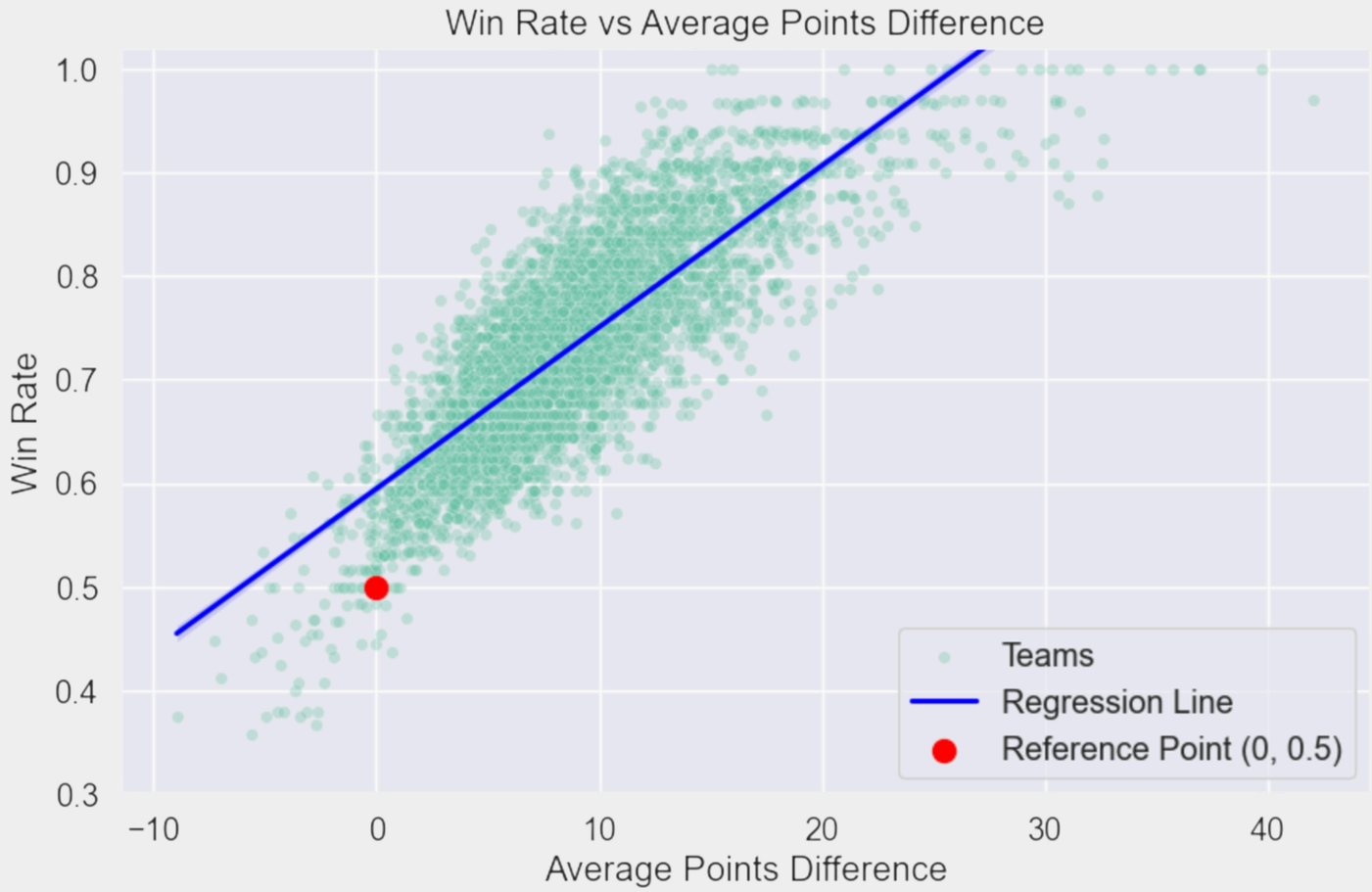
$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Game	Result	Predicted Team 1 Win Proba	Brier Score
1	Team 1 wins	0.5	0.25
2	Team 2 wins	0.5	
1	Team 1 wins	0.6	0.16
2	Team 2 wins	0.4	
1	Team 1 wins	1	0
2	Team 2 wins	0	

Added Features

Team-specific

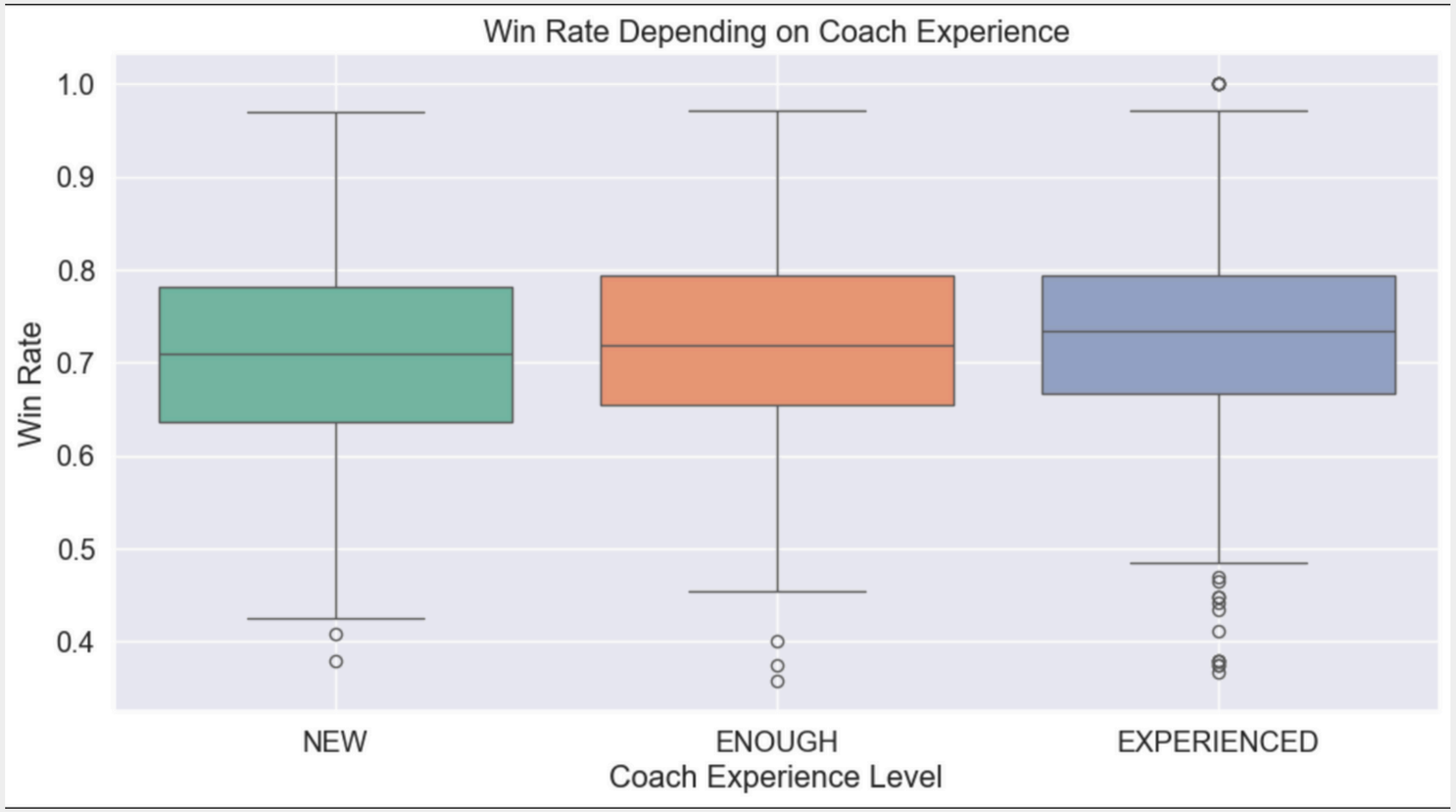
- Avg Points delta during the regular season: pts scored by team - pts scored by opponent
- Win Rate during regular season
- Coach fired during the season flg
- Coach consecutive years in the team
- Men's or Women's NCAA



Added Features

Team-specific

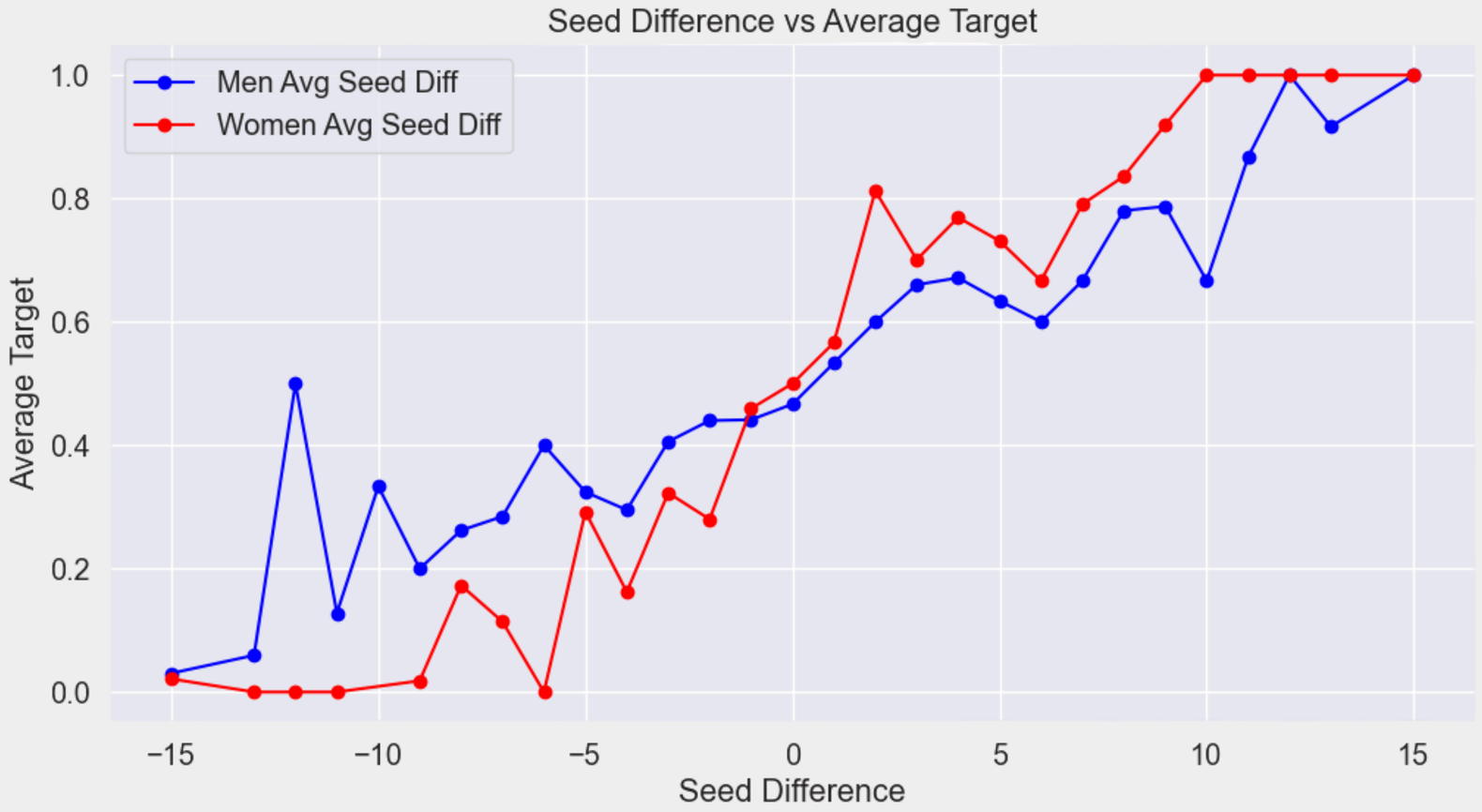
- Avg Points delta during the regular season: pts scored by team - pts scored by opponent
- Win Rate during regular season
- Coach fired during the season flg
- Coach consecutive years in the team
- Men's or Women's NCAA



Added Features

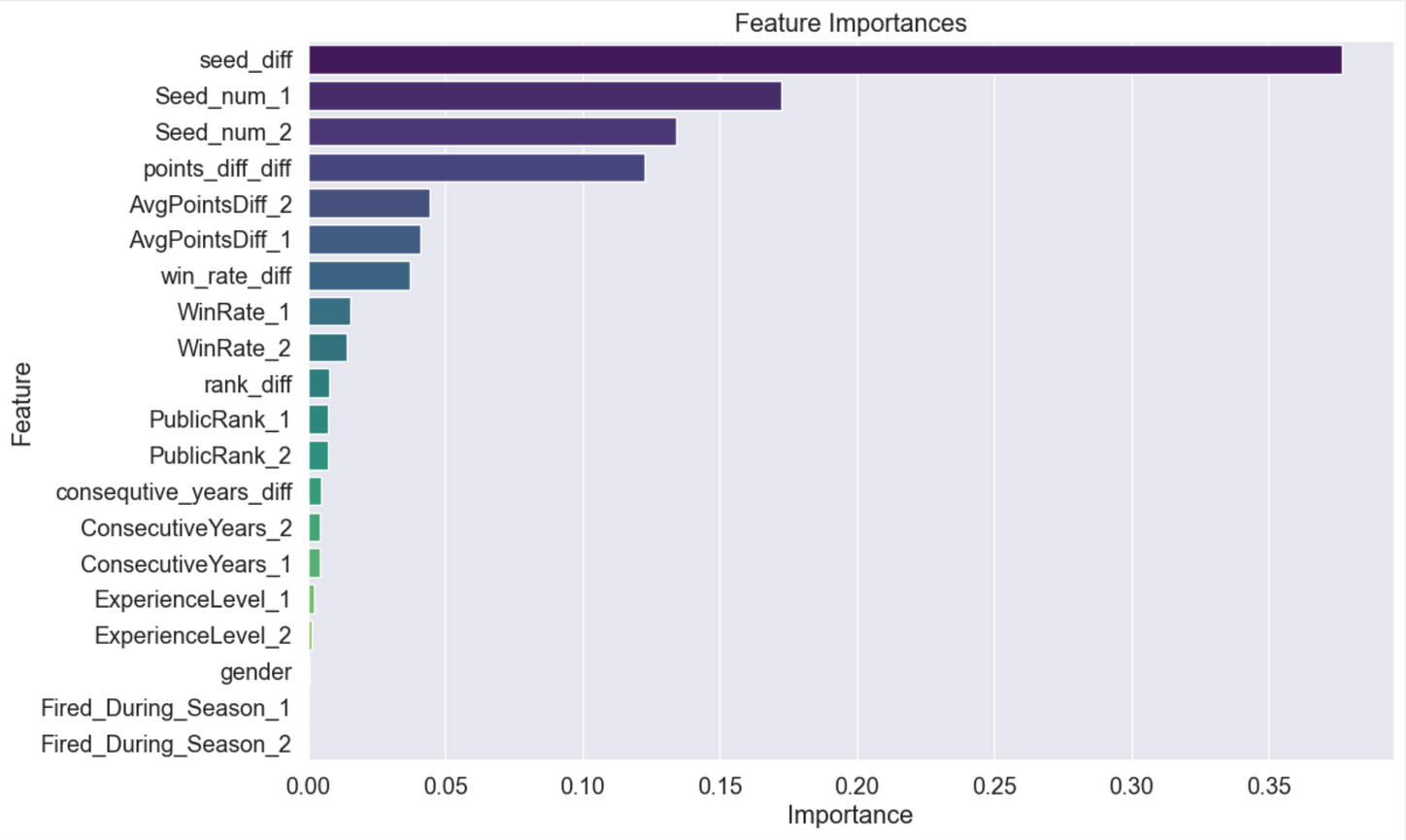
Pair-specific

- Difference between seeds
- Difference between public rankings
- Difference between point differences in regular season 🤯
- Difference between seeds

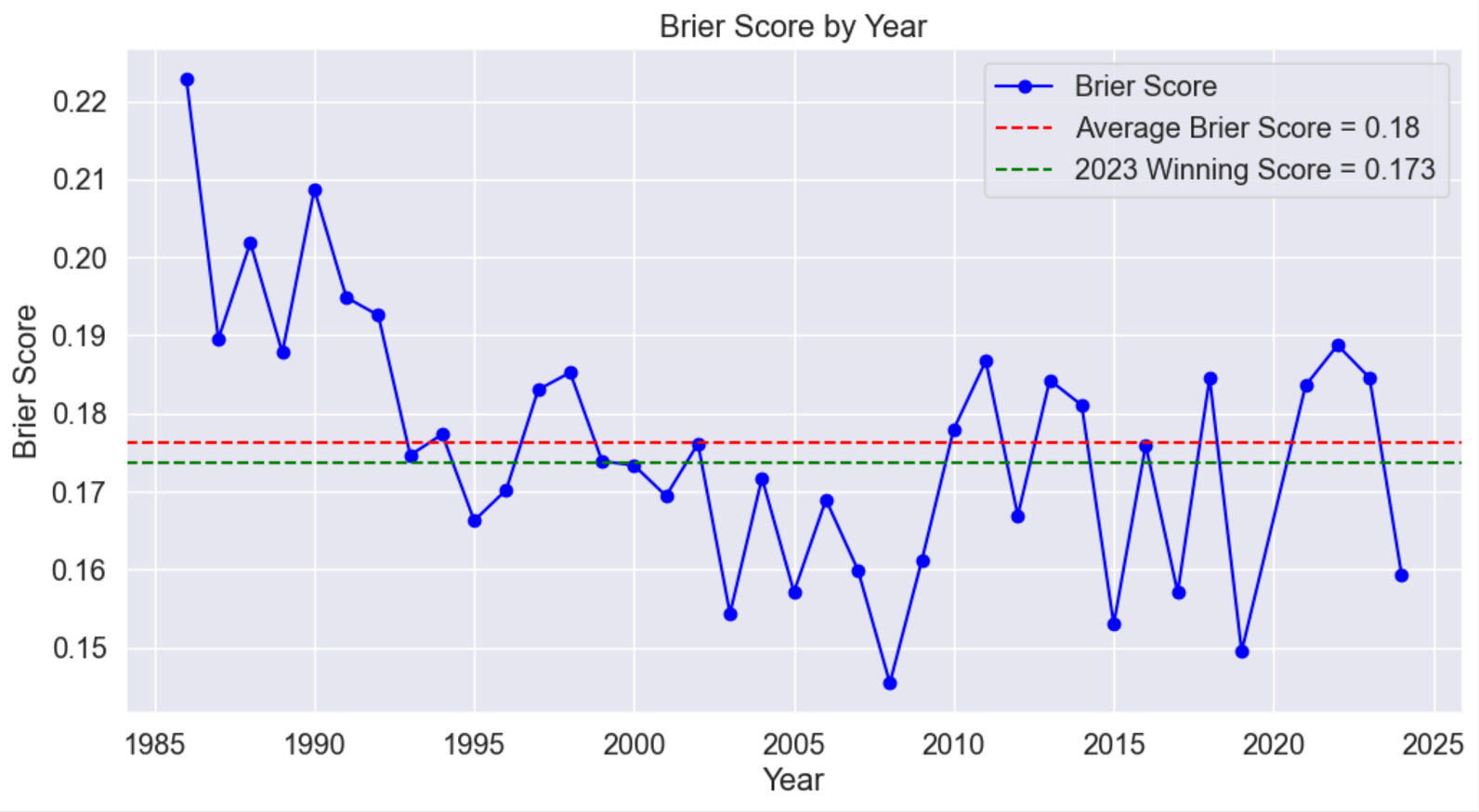


Model application

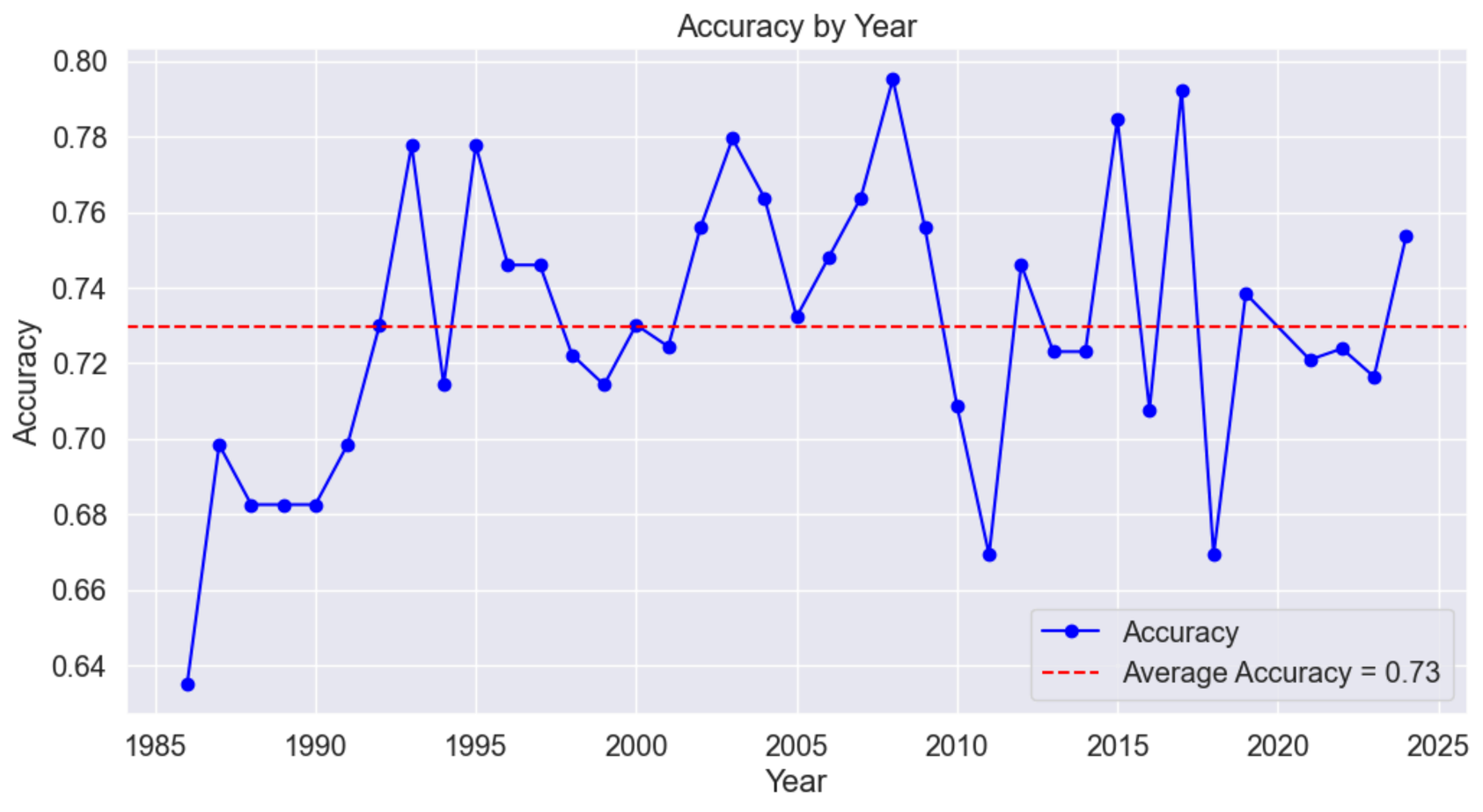
- Selected model: **Random Forest.** Why?
  - Easy to adjust hyperparameters to prevent overfitting since the dataset is small (~4500 samples)
  - Can capture complex dependencies
  - Can estimate feature importances
- Selected parameters after grid search with Cross-Validation:
  - n_estimators=1000
  - min_samples_split=5
  - min_samples_leaf=15
  - max_features=0.3
  - max_depth=5



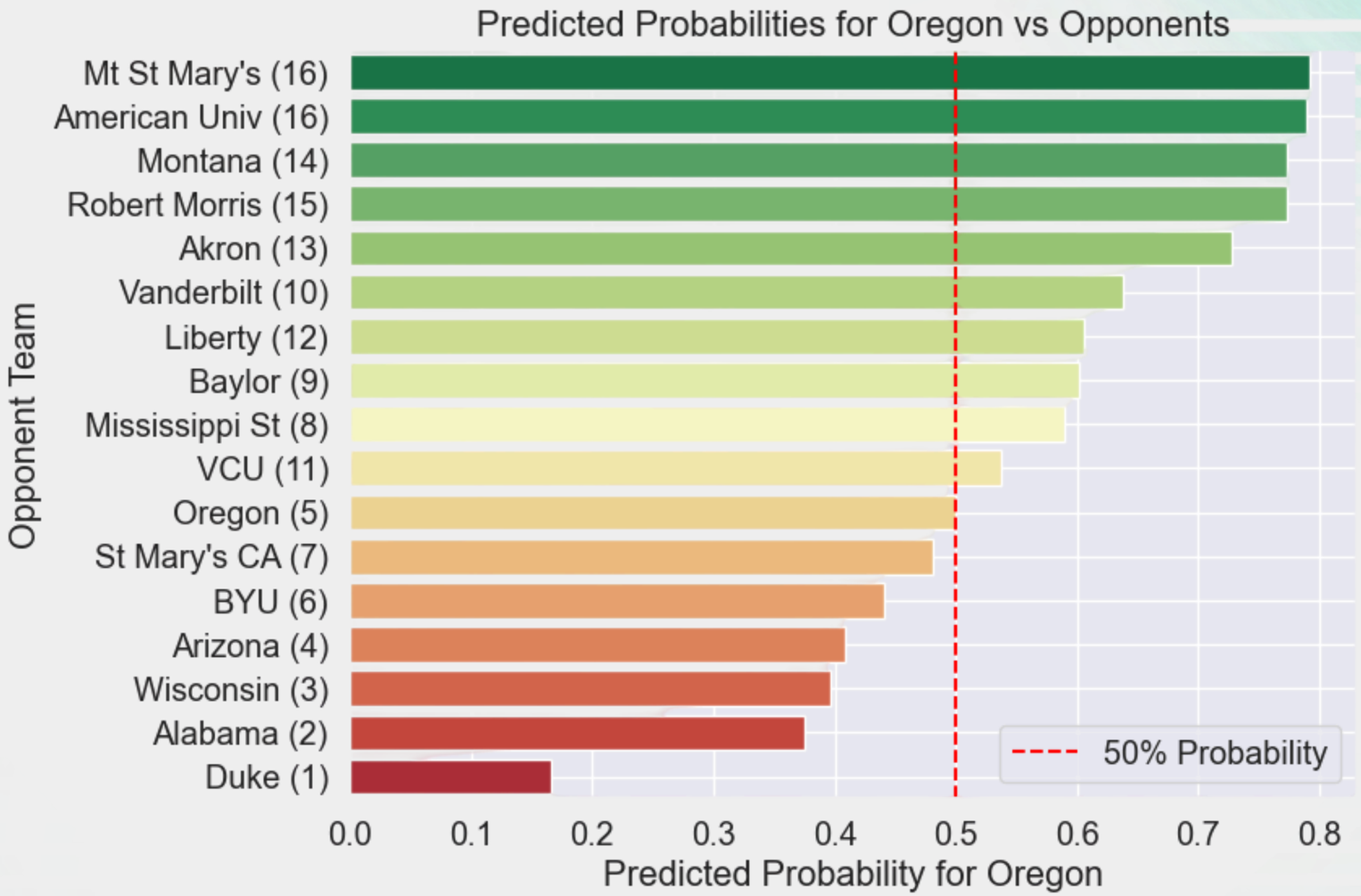
Results: Brier Score



Results: Accuracy



Predictions for UO Men's



Conslusion

- Identified the most important predictions
- Achived performance comparable to the top solutions of previous years
- Applications: sports analytics and betting industry

Thank you!