MARCH DATA CRUNCH MADNESS 2023

Predicting NCAA Tournament Champion

Team Jordan Year

Yujoon Jang
Jennifer McFadden
Masaya Sugimoto
Jingqing Zhou

Sponsored by: Deloitte.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Introduction

Problem Statement

Can we predict the 2023 March Madness NCAA Men's Basketball tournament bracket winners?

Objective

Utilize historical data, player performance, and additional features to build a model that accurately predicts 2023 March Madness NCAA Men's Basketball tournament result.

Methodology

Data Pre-Processing

- Data Validation
- Feature Engineering

Feature Selection

- PCA
- RFE
- Correlations

New Features

- Research
- Data Cleaning
- Data Merge

Model Selection

- Logistic Regression
- Random Forest
- Decision Tree
- XGBoost
- LightGBM

Model Evaluation

- Log Loss
- ROC AUC
- Precision
- Recall
- F-Measure

Data Pre-Processing & Feature Engineering

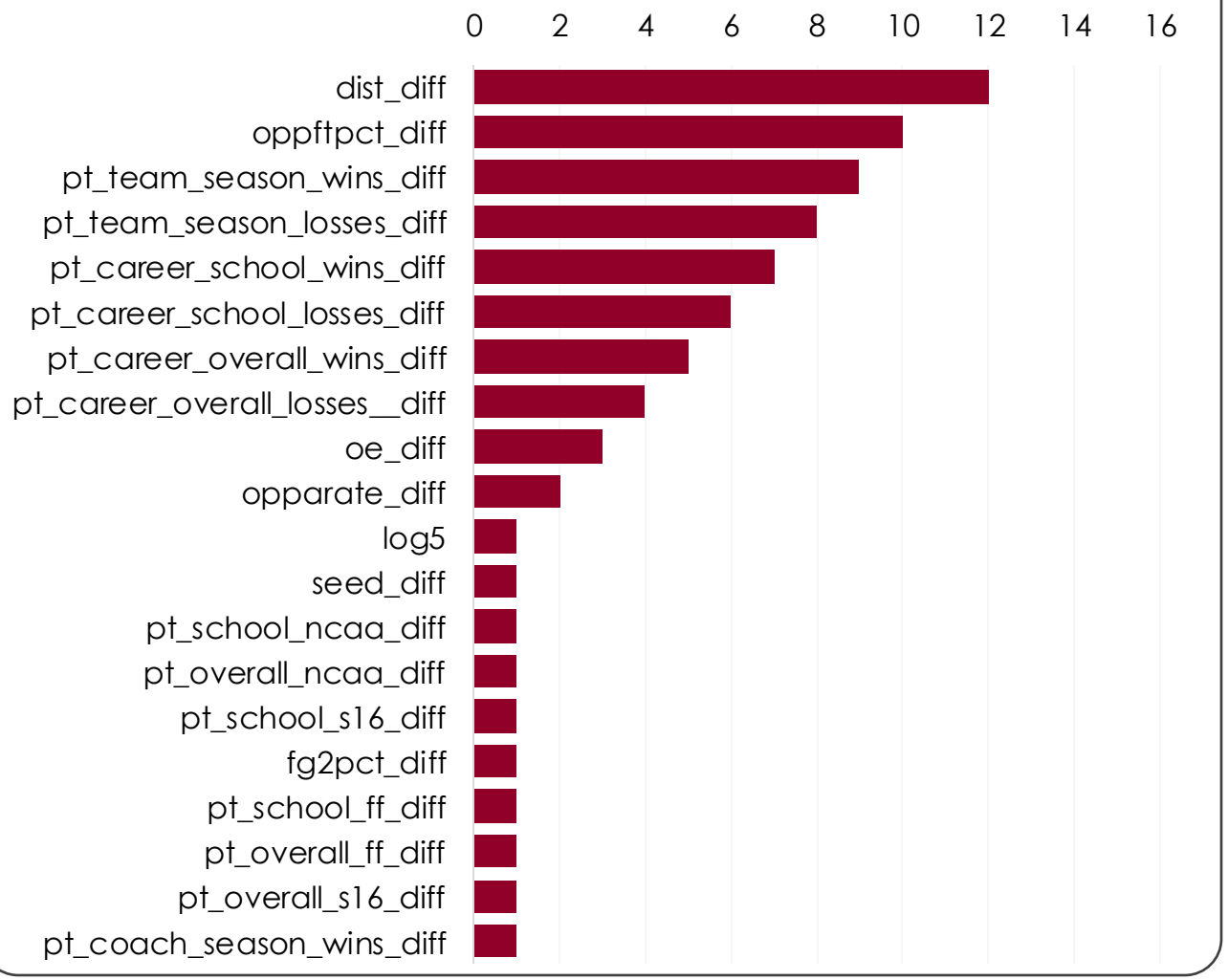
New features derived from existing data to enhance model accuracy

Pythagorean Win Percentage	Team's expected win percentage based on adjusted offensive & defense efficiency
Distance From Home	Distance between home court and game location court
Team 1 & Team 2 Difference	Difference between Team 1 and Team 2 for all performance metrics

Feature Selection

Used RFE and PCA techniques to identify features with strongest predicting power

RFE Feature Ranking



PCA

Principal Component 1

Represent 87% of the dataset

- dist_diff
- pt_career_overall_wins_diff
- pt_career_school_wins_diff
- pt_coach_season_wins_diff
- pt_team_season_losses_diff
- score_diff
- seed_diff
- adjoe_diff
- adjde_diff
- oe_diff

Principal Component 2

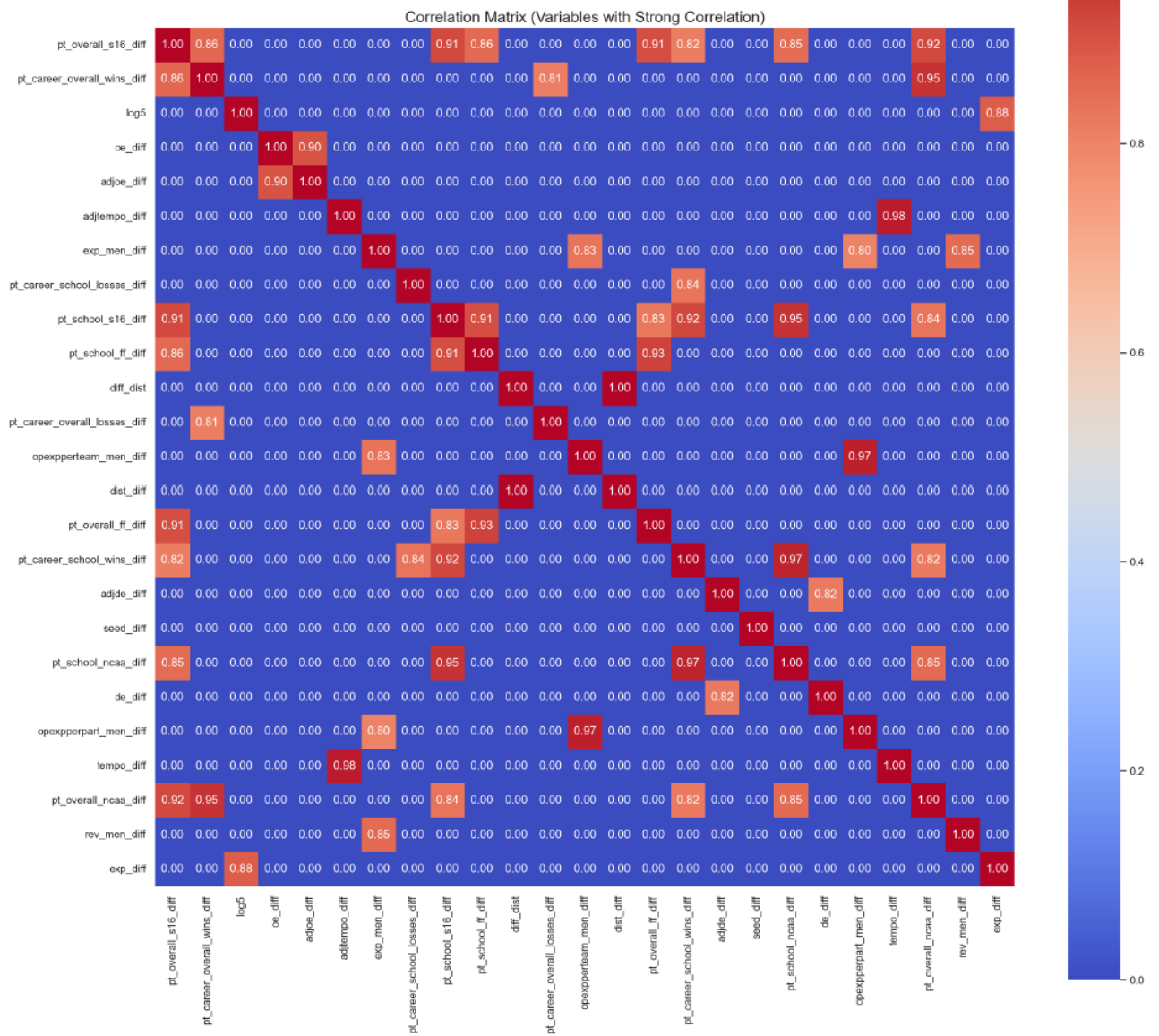
Combined with PC1 represent 95% of the dataset

- pt_career_overall_losses_diff
- pt_career_school_losses_diff
- pt_overall_ncaa_diff
- pt_school_ncaa_diff
- pt_overall_s16_diff
- pt_school_s16_diff

Feature Selection

Correlation analysis performed post feature selection to further refine feature list

- Heat map shows correlation between features remaining after selection process
- “Hot” features removed due to high correlation



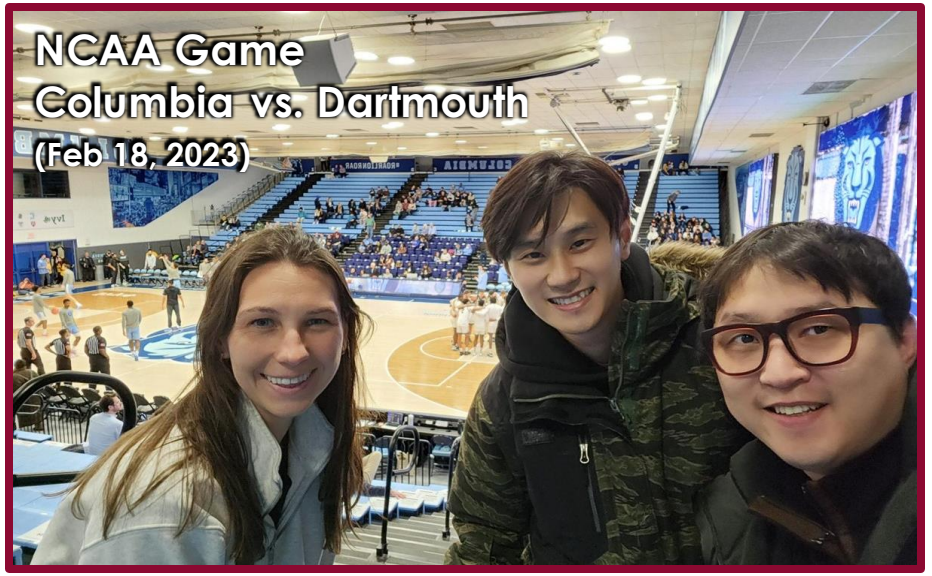
New Feature Research

Scan Me!
QR code to
Qualtrics survey

Conducted preliminary research to understand most impactful indicators for tournament success

Research

- 1st party Qualtrics survey conducted at NCAA regular season game (15 total responses)
- NCAA super-fans family members and friends
- Articles and insights from top analysts



Top-cited Indicators

- Money
- Team and coach historical performance
- Scoring ability

Areas of opportunity based on data source availability

1) Team budget
2) Top scorers per team
3) Injured players

New Features

Expense & Revenue

- Men's team expenses
- Revenue generated by Men's Team

Source: US Department of Education Equity in Athletics Data Analysis

Star Player Rating

- Top scoring player's avg. points-per-game (ppg) proportion of team avg. points-per-game

Source: NCAA Statistics

Player Injury Adjustment

- 10% reduction on adjusted offensive & defensive efficiency for season-ending key player injuries

Source: Rotowire & News Coverage

* Applied to 2023 data only

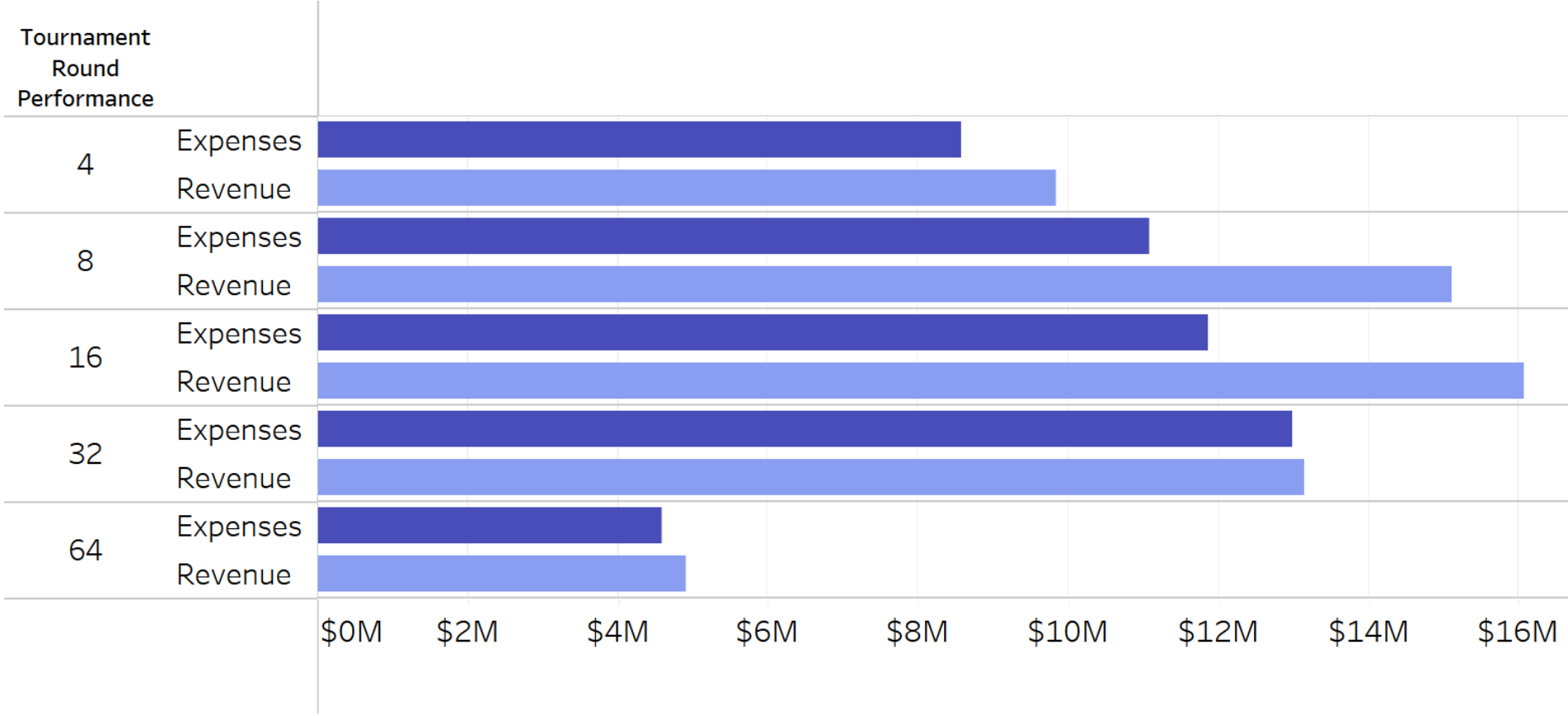
Enhanced Historical Training Data:

- Historical data for new features only available for 2011-2023, so 2002-2010 season data were removed from training set.
- However, predicting value from new features outweighs the loss of a smaller training data size, and removes potentially outdated seasons.

New Features

Higher expense budget and revenue advantageous in early rounds, importance waned in later rounds

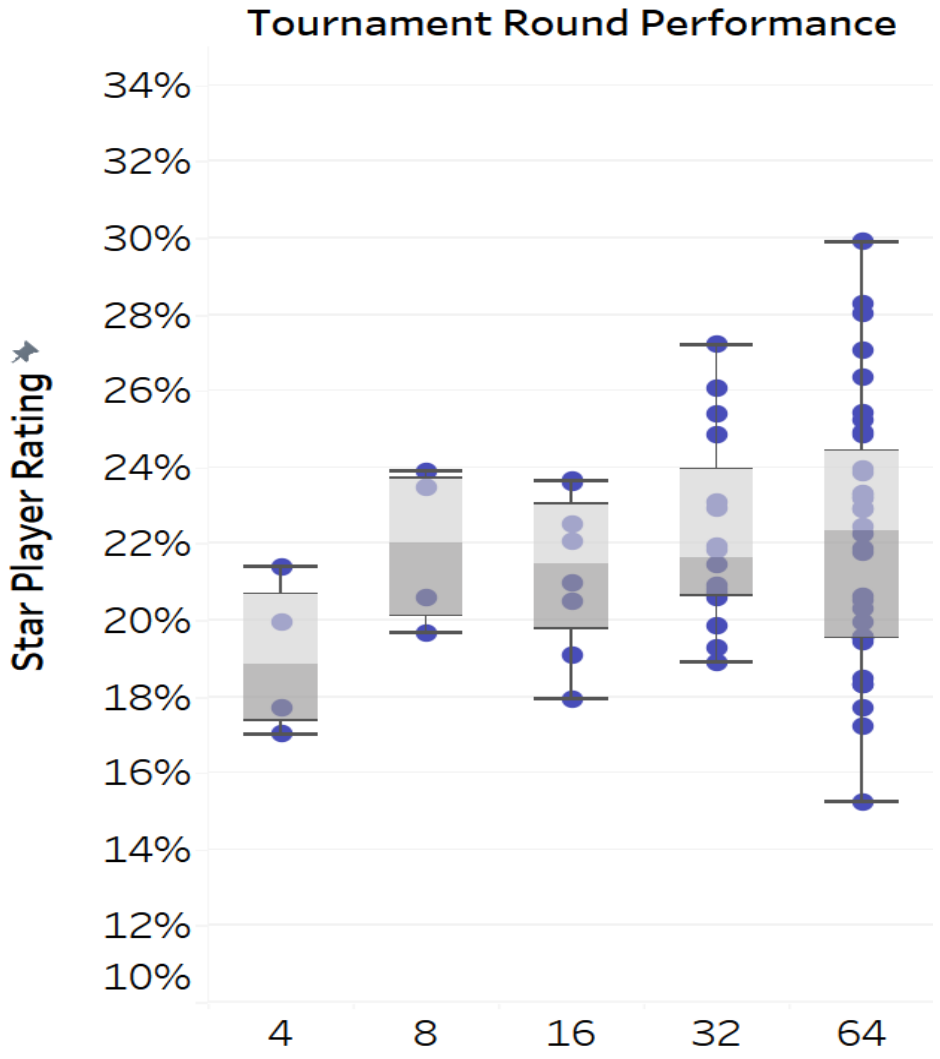
Average team expenses and revenue by tournament round performance (through final four)



New Features

- Bigger is not necessarily better for star player rating
- Top scorer should account for middle ground of ~20% total team points

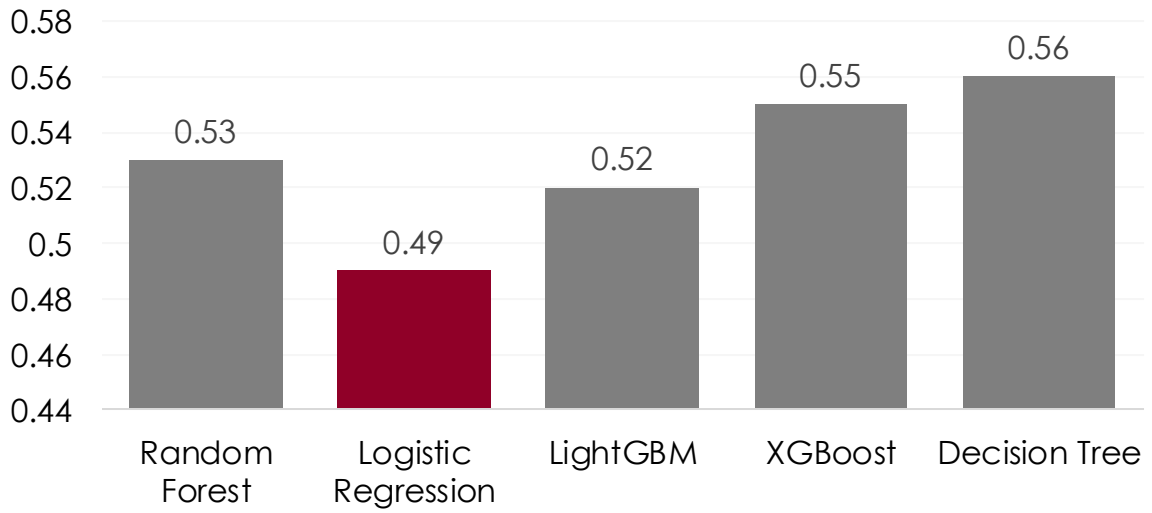
Team star player rating by tournament round performance (through final four)



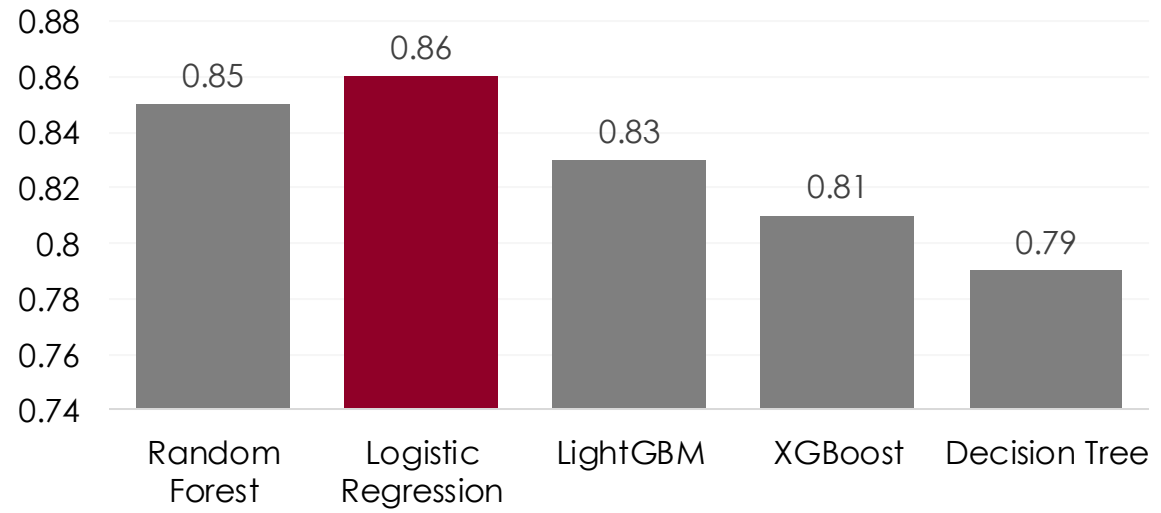
Model Performance Comparison

Out of five models tested, Logistic Regression strongest based on log loss

LOG LOSS



ROC AUC

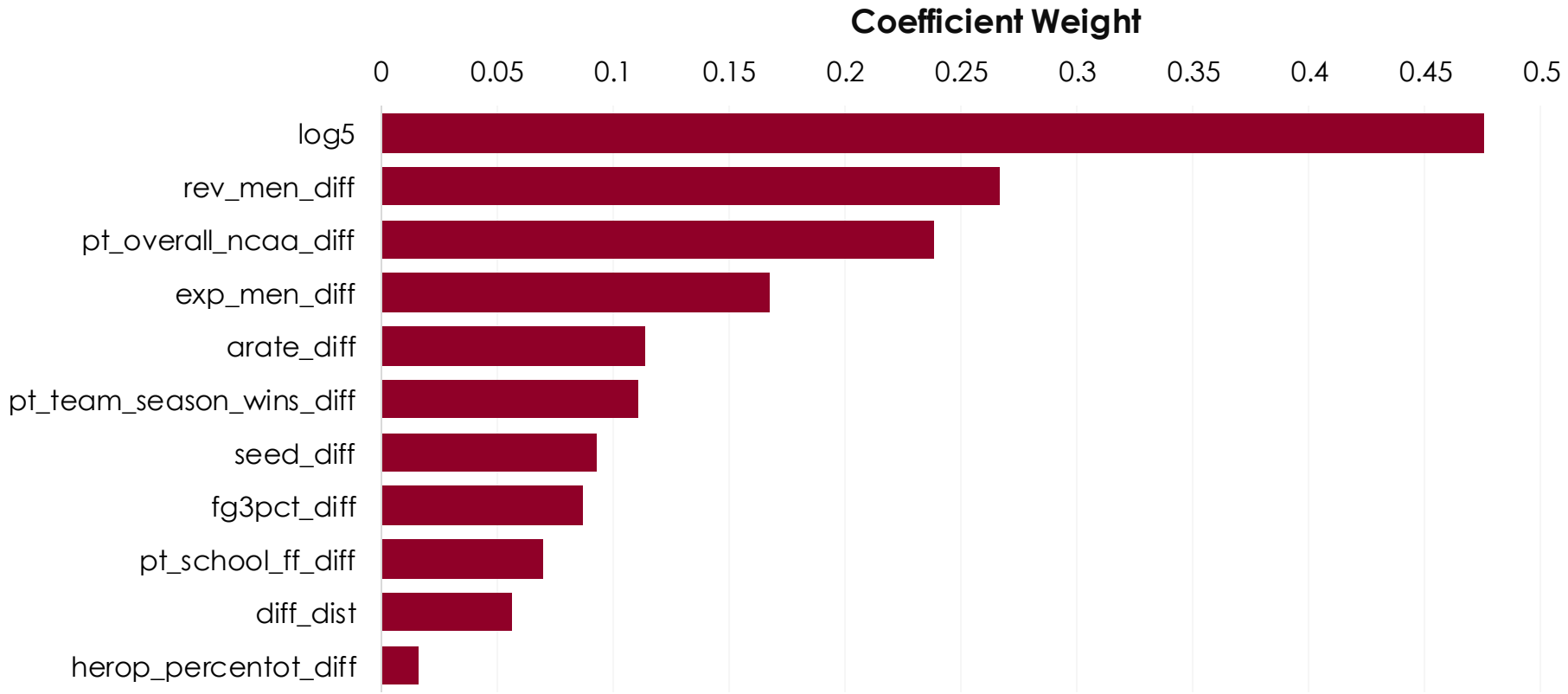


Name	Accuracy	Precision	Recall	F1 Score	Log Loss	ROC AUC
Random Forest	**0.82**	**0.75**	0.89	**0.81**	0.53	0.85
Logistic Regression 🏆	0.80	0.73	0.85	0.79	**0.49**	**0.86**
LightGBM	0.72	0.62	**0.97**	0.76	0.52	0.83
XGBoost	0.74	0.66	0.89	0.76	0.55	0.81
Decision Tree	0.75	0.71	0.75	0.73	0.56	0.79

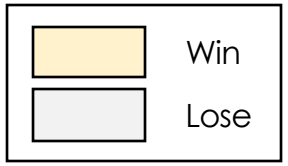
Model Evaluation

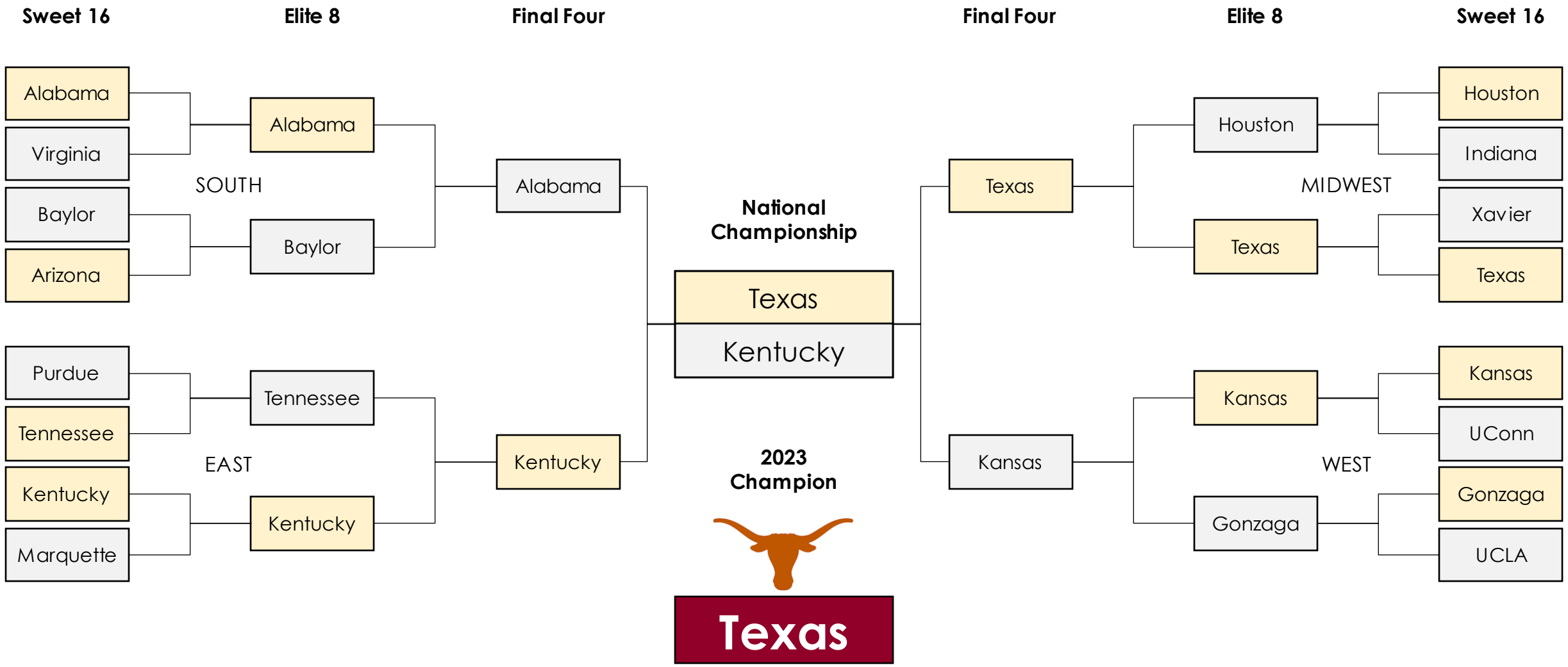
New features Revenue and Team Expenses increased model accuracy

Logistic Regression Feature Importance



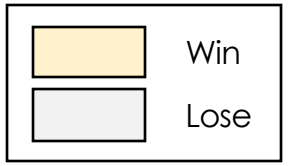
Tournament Result

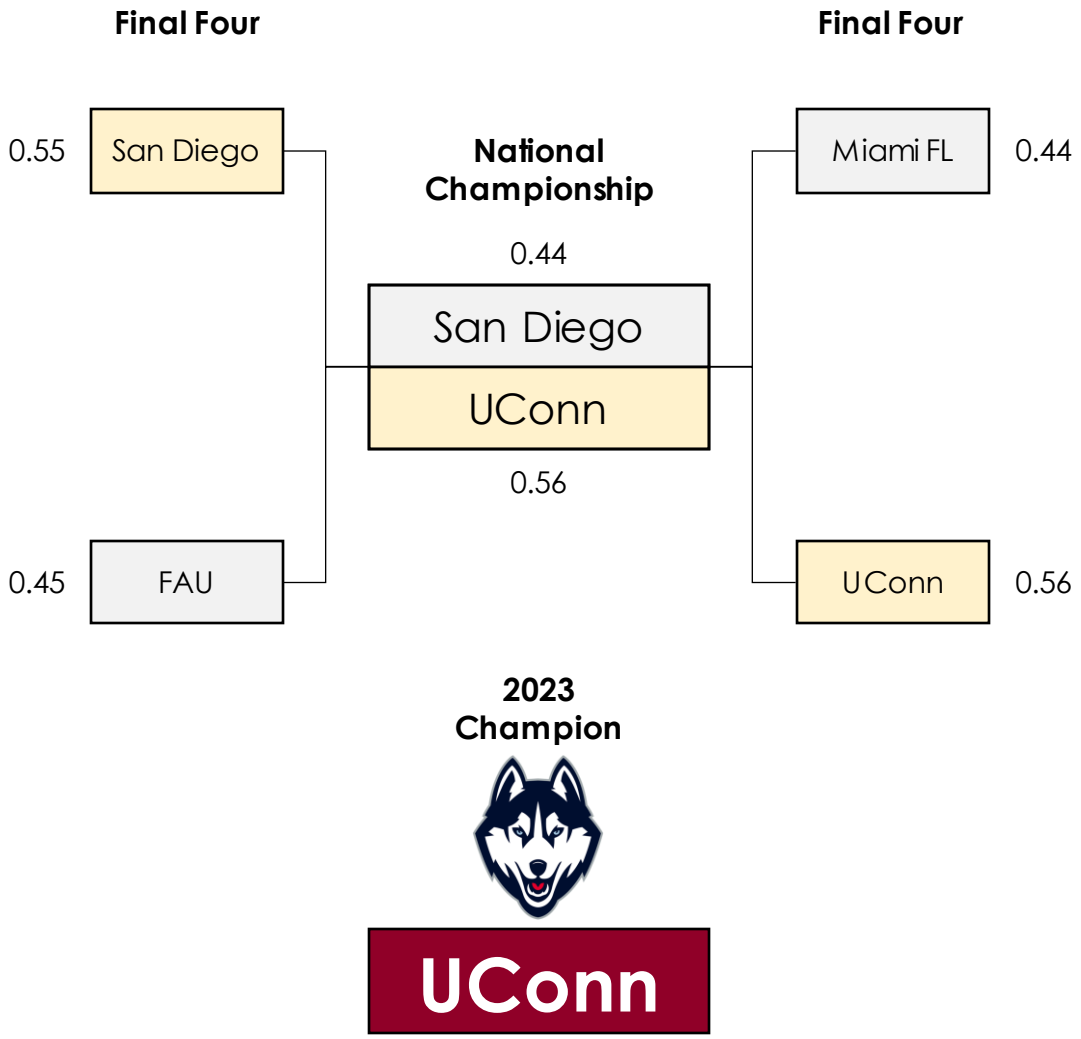




* Based on Logistic Regression Model

Tournament Result (Amended)





* Based on Logistic Regression Model

Thank You!

If you have any questions about the data,
please contact us at:

Yujoon Jang – yjang12@fordham.edu

Jennifer Cate McFadden – jmcfadden4@fordham.edu

Masaya Sugimoto – msugimoto@fordham.edu

Jingqing Zhou – jzhou109@fordham.edu

Sponsored by: Deloitte.

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK