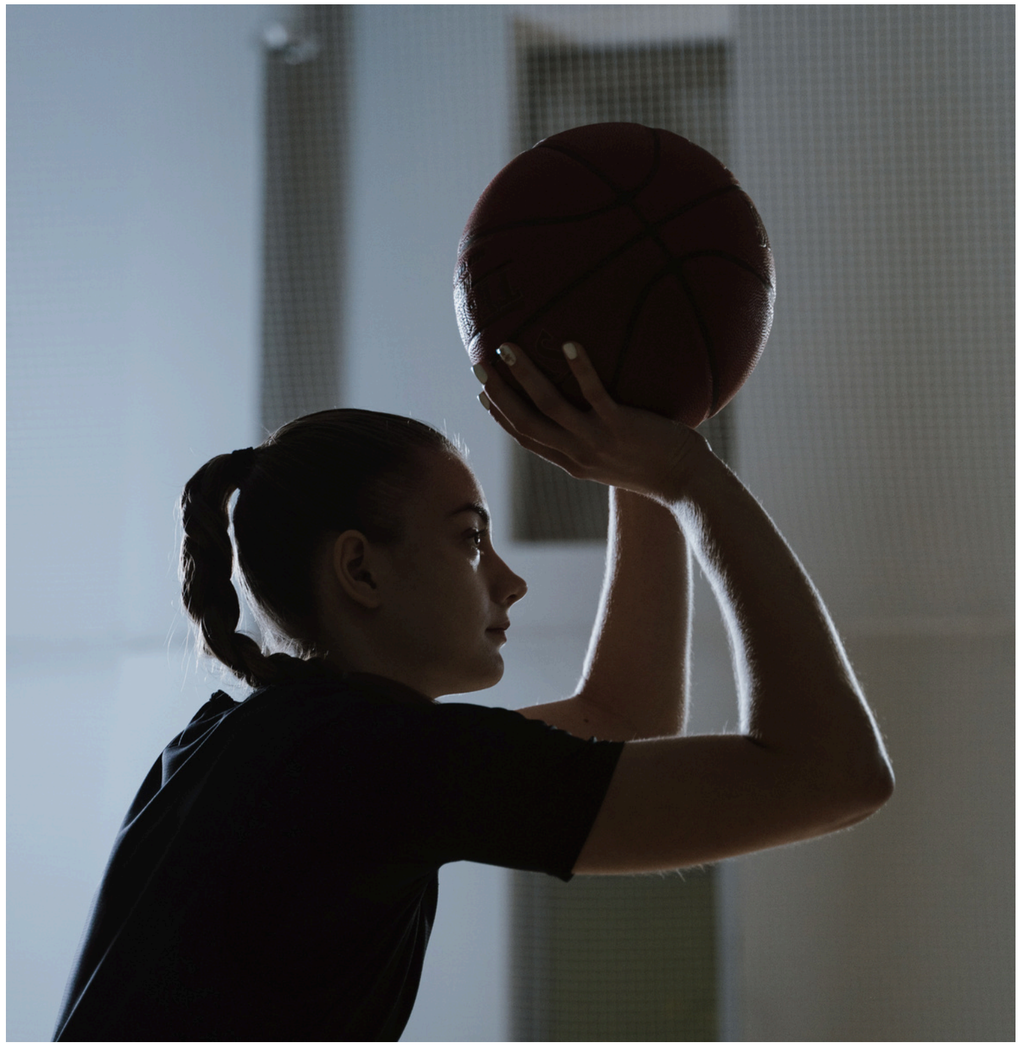
NCAA Championship Prediction and School Affinity Analysis

Crossroads Classic Analytics Challenge 25

PRESENTED BY NIMBUS 2025

DA FANG LIN, HUNG-CHEN HSU, YIRAN LIU



AGENDA

- Background & Problem statement
- Data Exploration
- Data Preprocess
- Model Building
- Model Result
- School Affinity Analysis

WAHT IS THE NCAA?

Deliver a world-class athletics and academic experience for student-athletes

Structure

Governs college athletics across three divisions

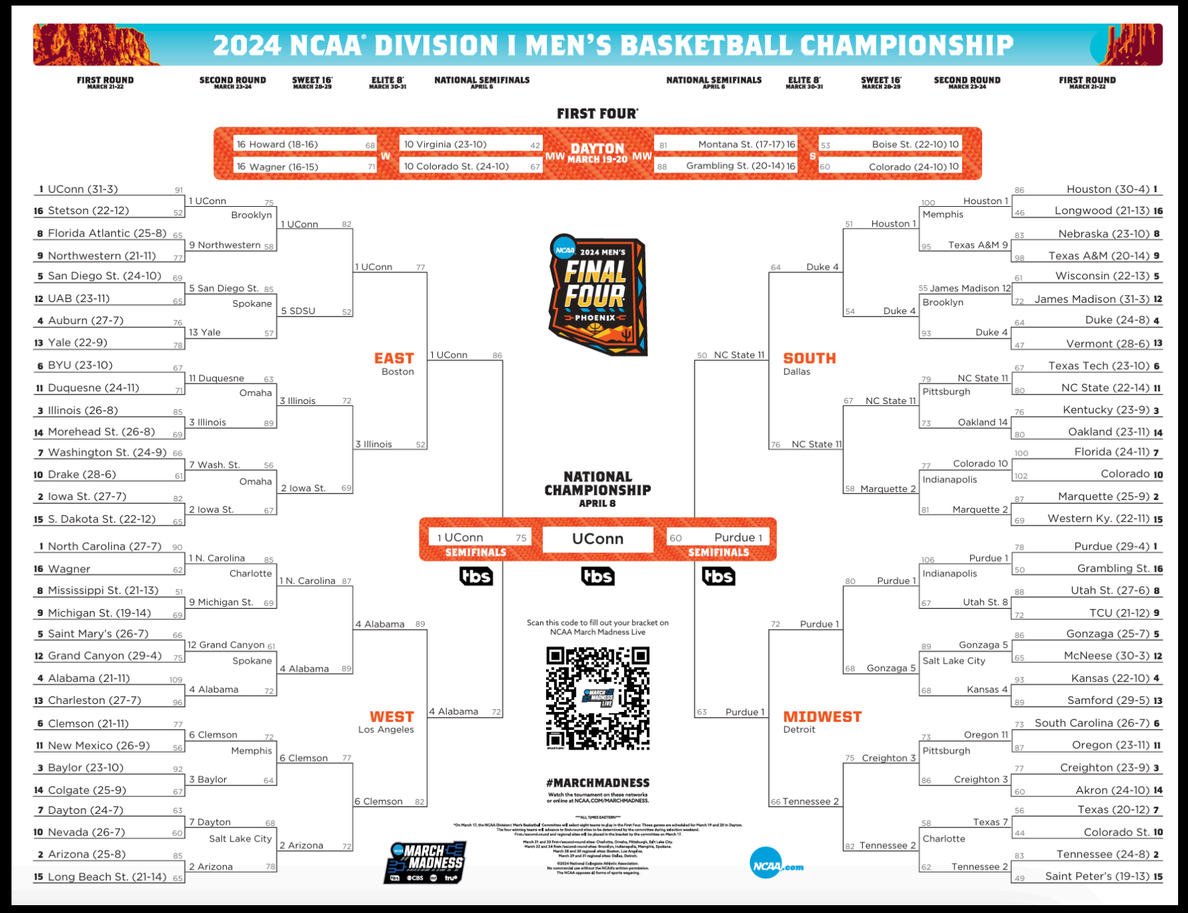
Scale

- 1,100+ member institutions
- 500K+ student-athletes

Competition

Hosts 91 championships in 25 sports

PROBLEM STATEMENT



Predict the Final Rounds

Develop models to forecast the semifinal and championship winners based on submitted brackets.

Understand School Affinity

Analyze whether a participant's school affiliation influences their bracket predictions.

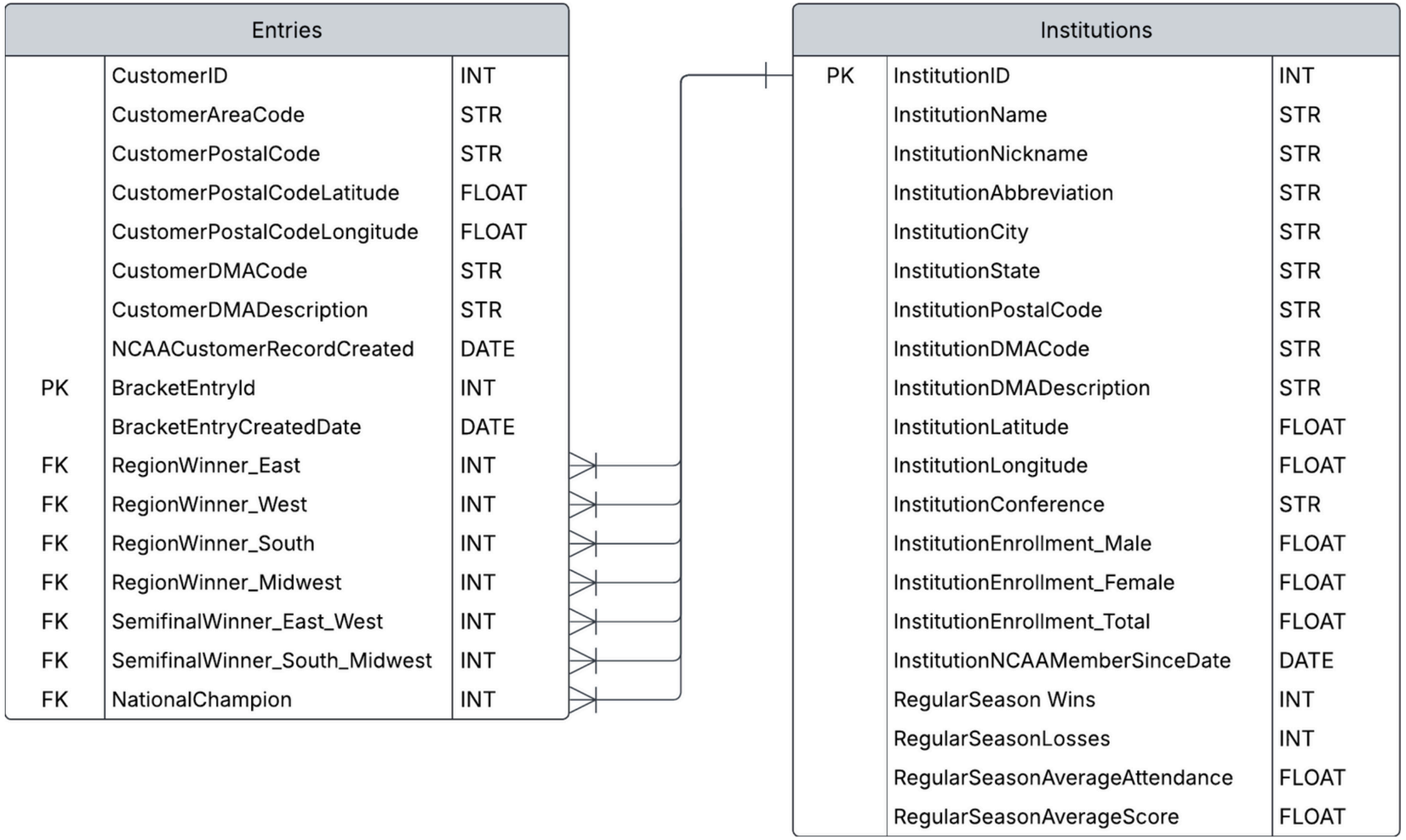
DATASET INTRODUCTION

Institutions

- 68 NCAA Division I teams competing in the 2024 Basketball Championship.

Entries

- 144,447 user-submitted entries for the Bracket Challenge.
- Each user can submit multiple entries.

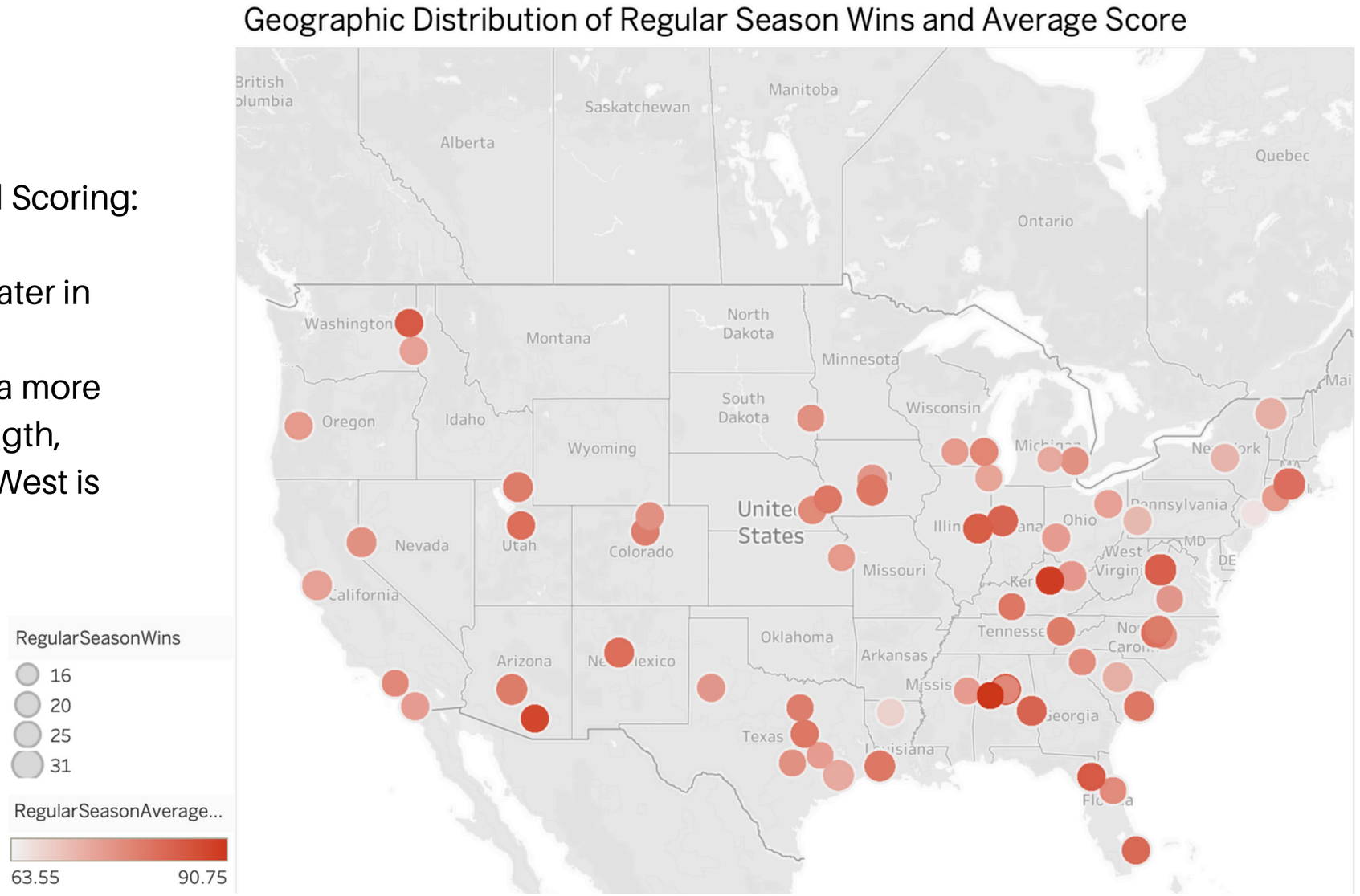


DATA EXPLORATION

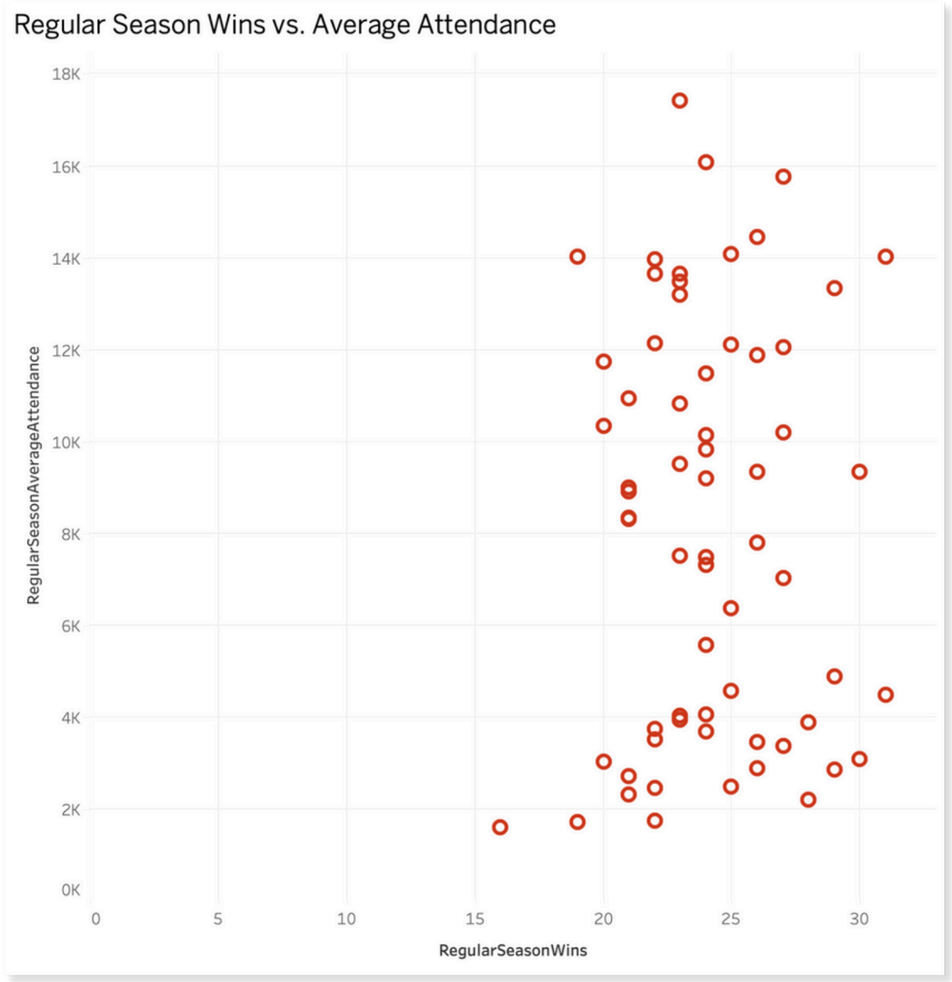
EDA

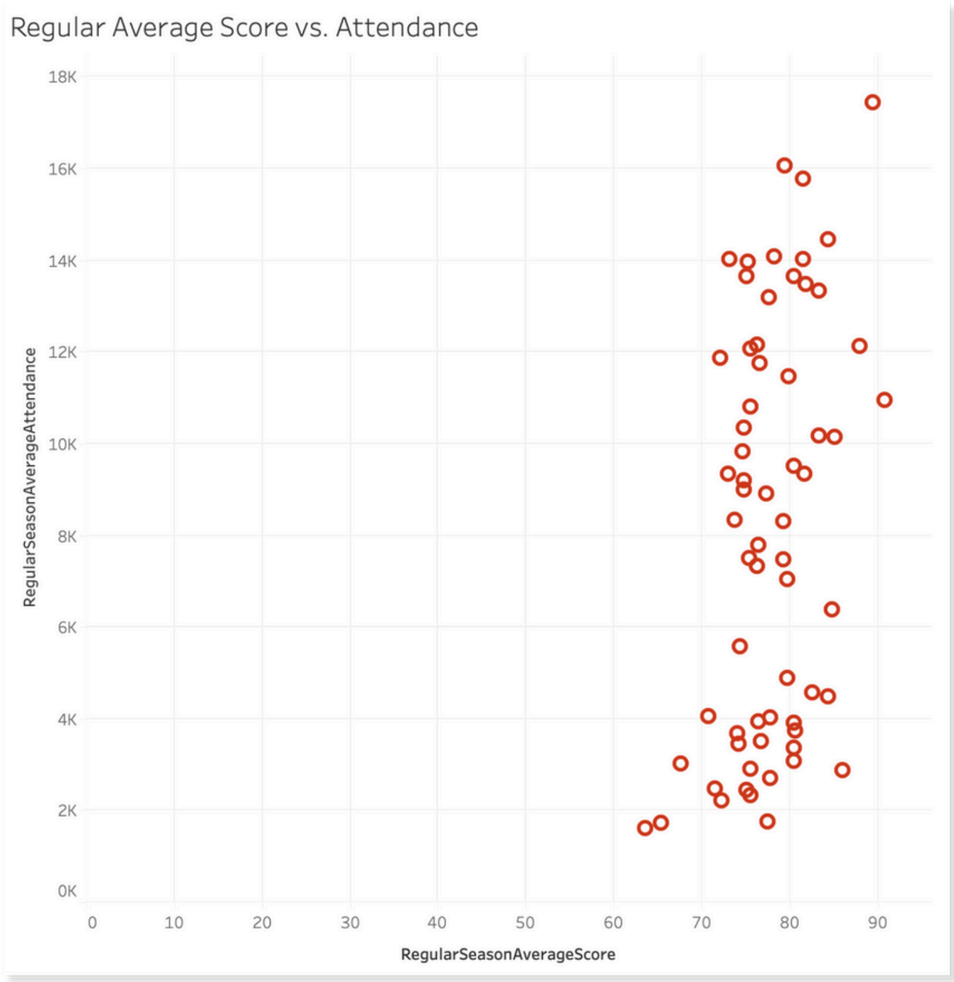
A Geographic View of Wins and Scoring:

- The number of teams is greater in the East than in the West.
- The teams in the East have a more noticeable disparity in strength, whereas the scoring in the West is relatively more balanced.



EDA



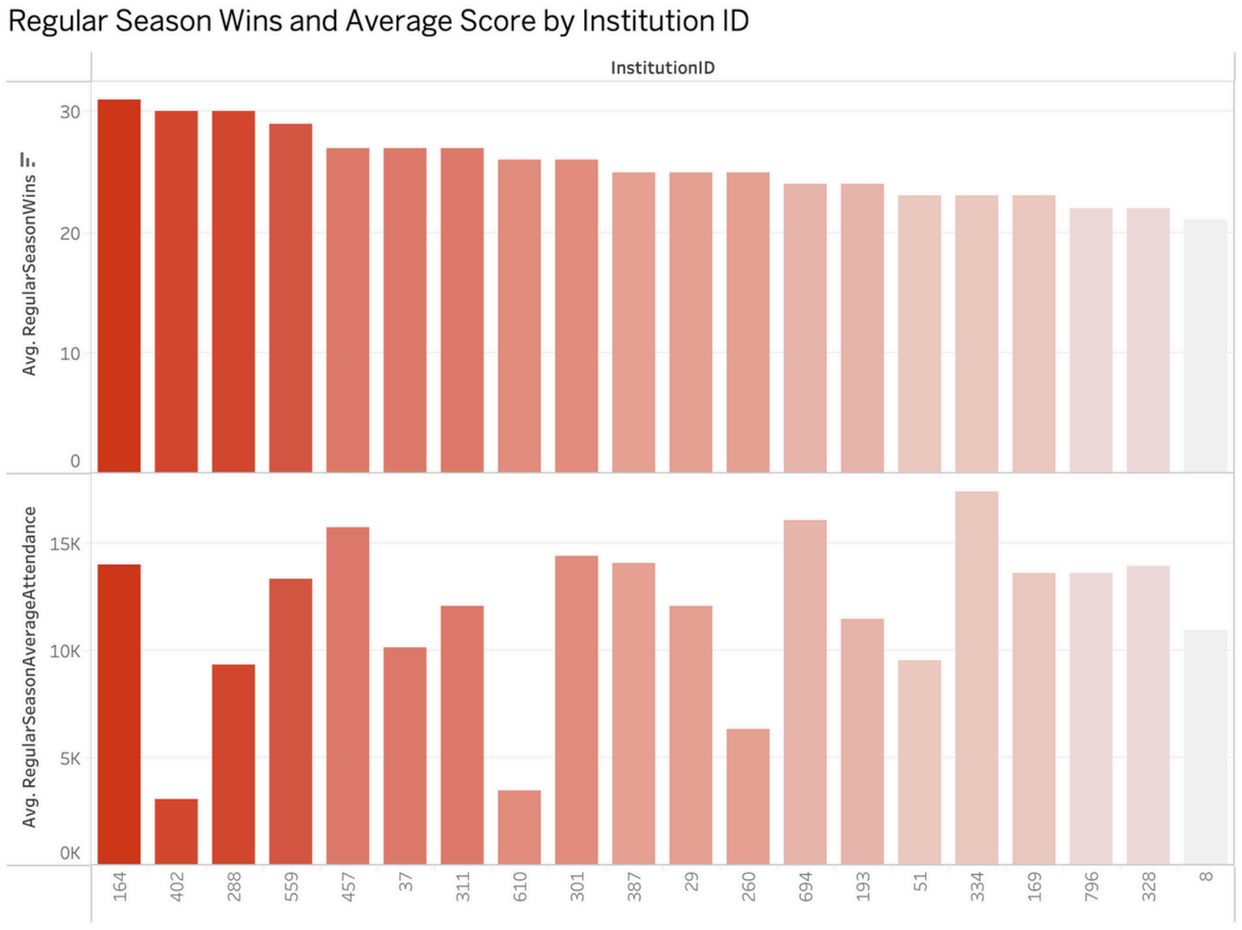


Examining the relationship between win count & average score and attendance.
We found that team performance does not directly impact audience size.

EDA

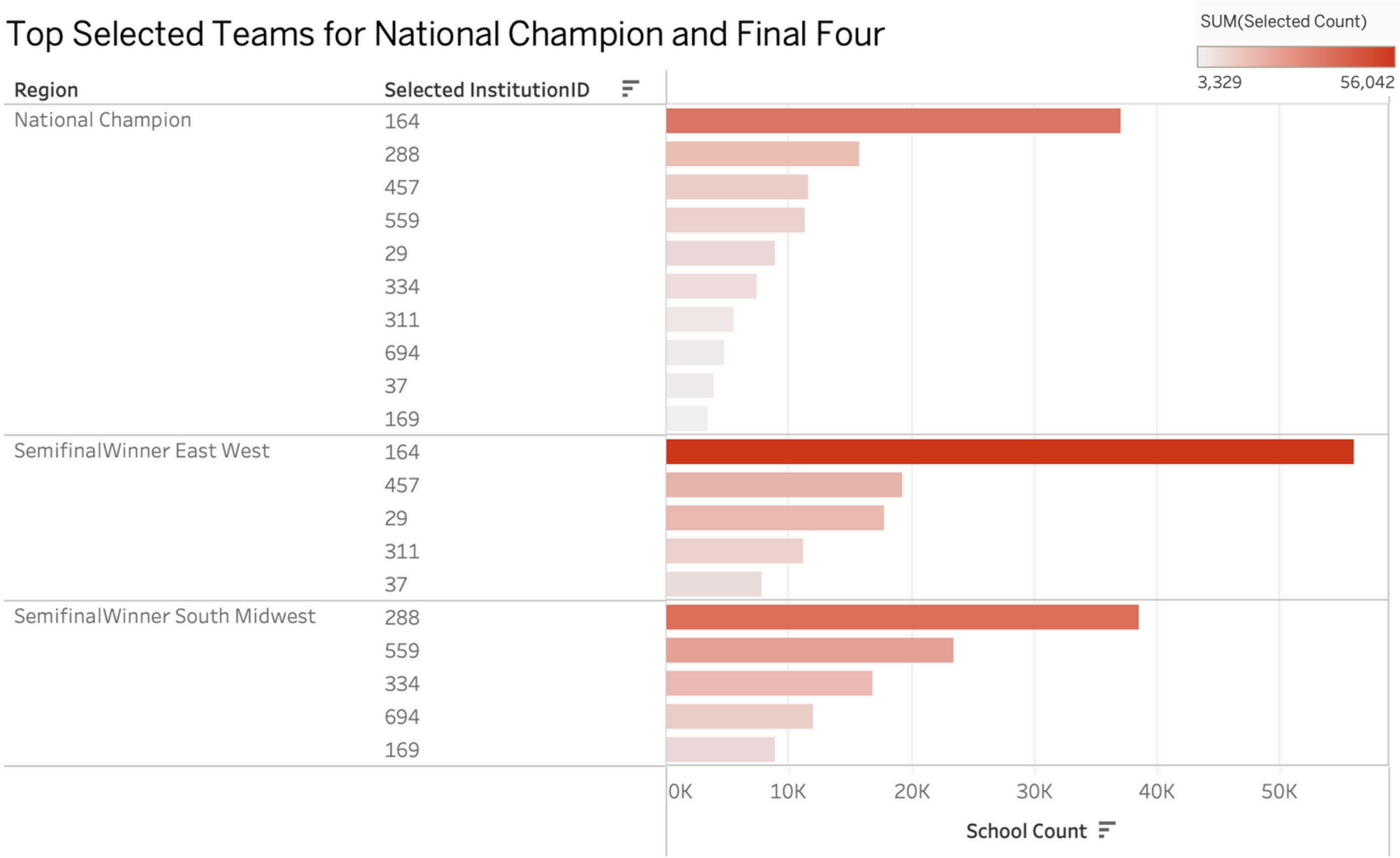
A school's performance (wins) does not directly impact local audience engagement. This may be due to lower regional interest in the tournament or a weaker college fan culture.

In our Prediction:
Audience selection behavior depends on more than just a school's performance; other factors must be considered.

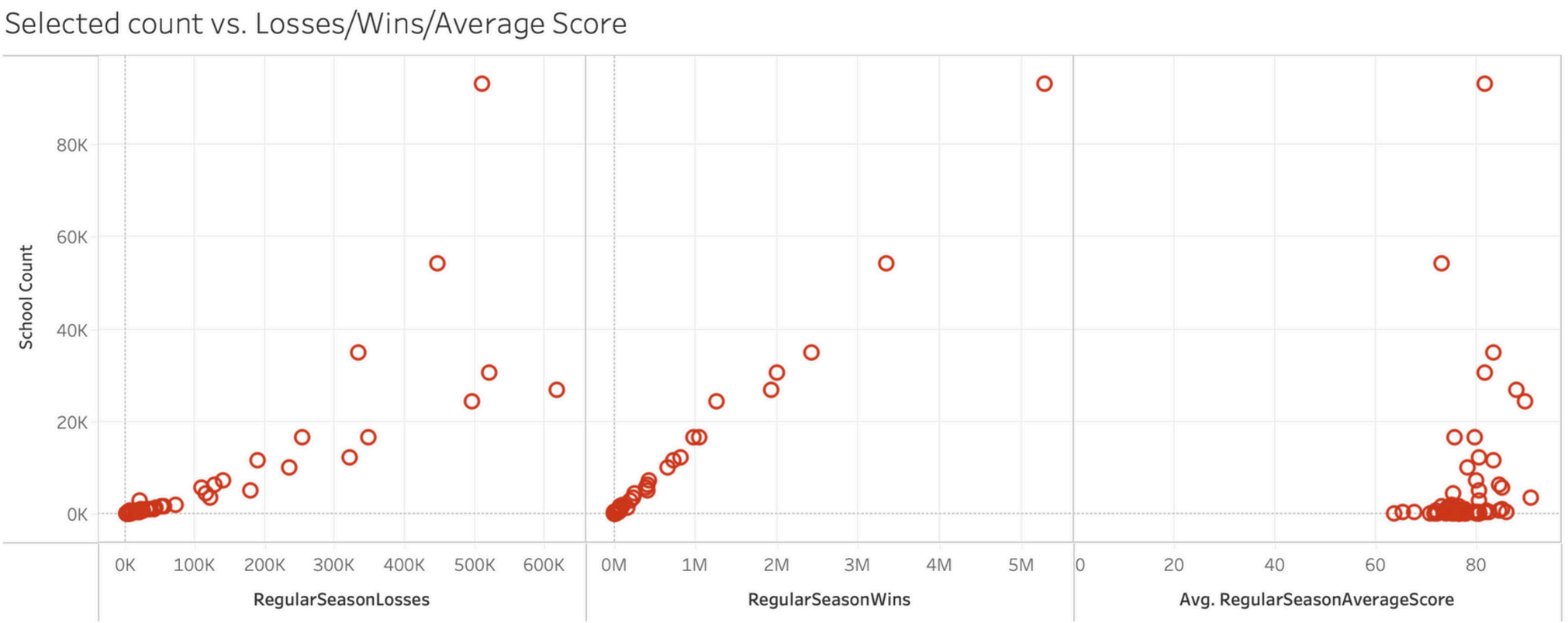


EDA

164 & 288 are the most picked teams for the final four and championship, continuing from the previous bar chart, they are indeed the teams that win more.



EDA



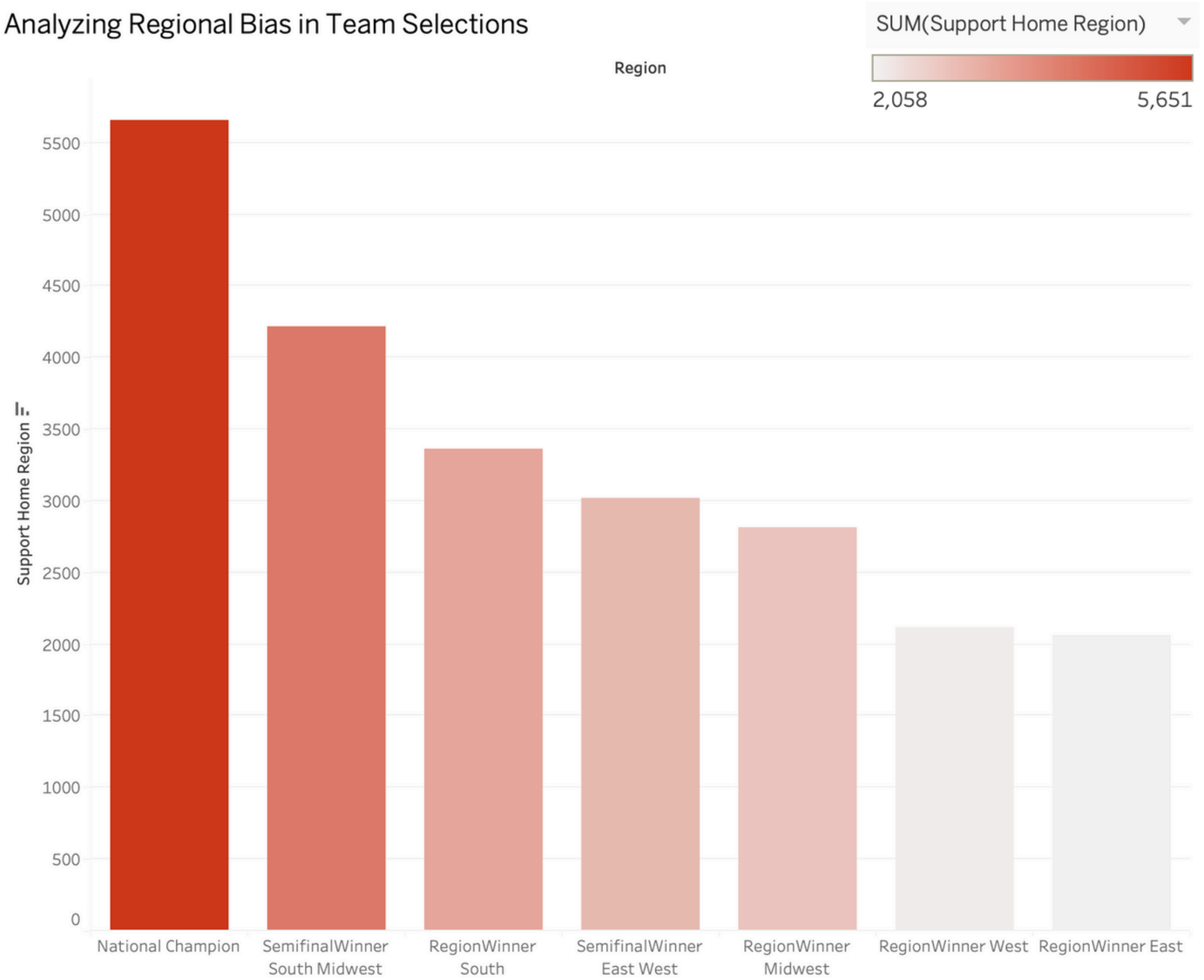
More selections -> More losses: Some well-known schools have a large fan base even if their performance is not perfect.

More selections -> More wins: This aligns with intuition, as strong teams usually attract more attention.

More selections -> No clear correlation with average score: Scoring may not be the primary factor influencing audience choices.

EDA

- Regional support is highest in the **championship game**, indicating that audiences are more likely to choose teams from their own region in the final stage.
- **Semifinal-South Midwest** follows, showing strong regional loyalty.
- **Region South** surpasses Semifinal-East West in support, suggesting higher local team loyalty.



DATA PREPROCESS

Feature Interaction
Missing Value Imputation
One-Hot Encoding
Feature Selection

FEATURE INTERACTION

East_West_Diff
South_Midwest_Diff

Calculate the differences between the winners in different regions

e.g. 113

DistanceToEast
DistanceToWest
DistanceToSouth
DistanceToMidwest

Calculate the geographic distance between Customer Postal Code Latitude/Longitude & Institution Latitude/Longitude

e.g. -923.981672

East_Wins
West_Wins
South_Wins
Midwest_Wins

East_Conference
West_Conference
South_Conference
Midwest_Conference

Merge school information to create more complete competition data.
Understand the season wins (_Wins) and the conference (_Conference) of the winning teams in the region.

e.g. 30

e.g. Big East Conference

DATA PREPROCESS

Missing Value

Categorical feature

Fill missing values with the ‘missing’.

e.g. ****_Conference

Numeric features

Fill missing values with the median.

e.g. CustomerPostalCodeLatitude, CustomerPostalCodeLongitude

Feature Engineering

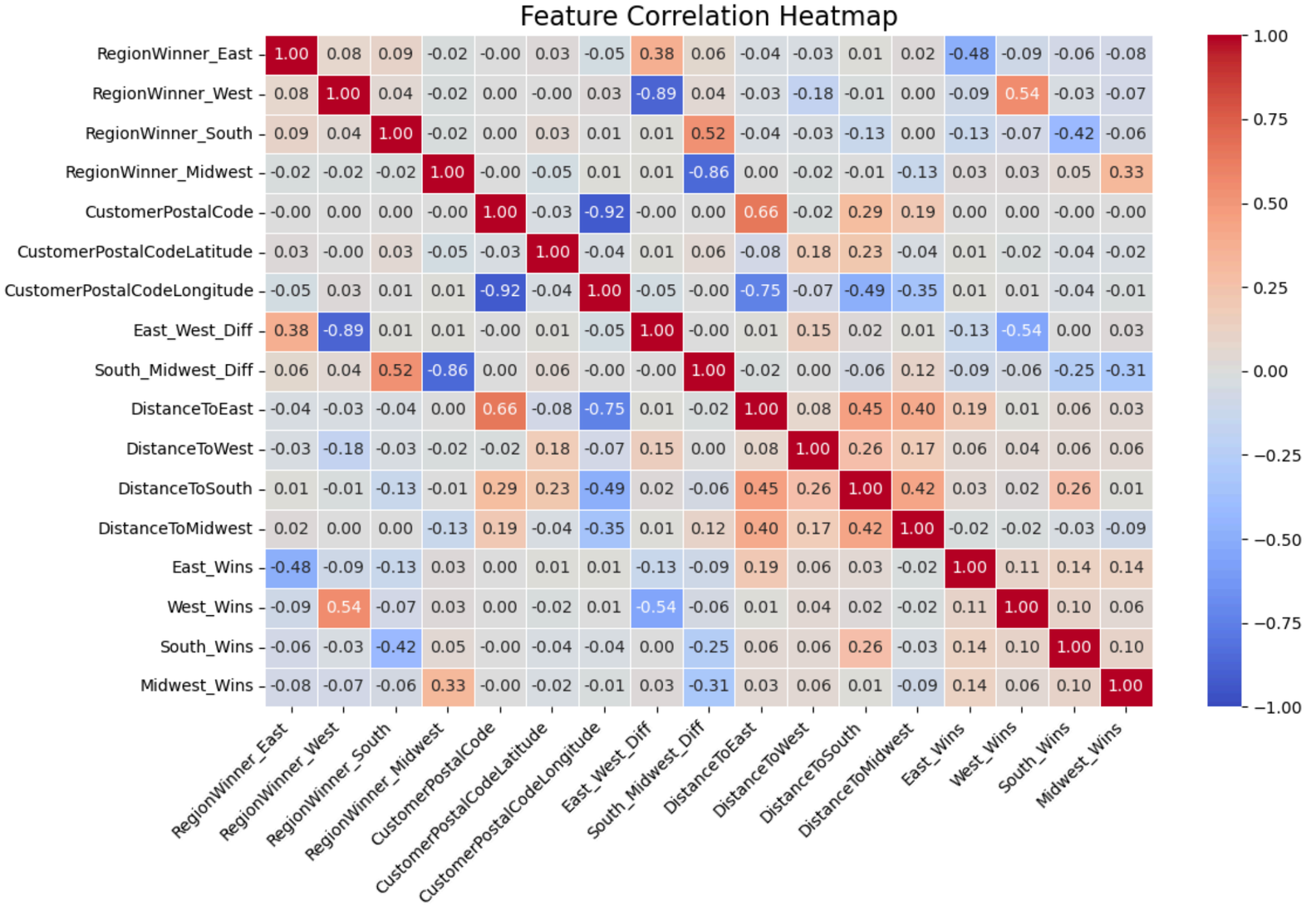
One hot enconding

Convert each categorical variable into binary format so that the model can process it.

Midwest_Conference_Big Ten Conference	Midwest_Conference_Metro Atlantic Athletic Conference	Midwest_Conference_Mid-American Conference
1	0	0
0	1	0
0	0	1

FEATURE SELECTION

Identifying highly correlated variables by using the **Heatmap**.



FEATURE SELECTION

Target

SemifinalWinner_East_West

SemifinalWinner_South_Midwest

NationalChampion

Existing Features

RegionWinner_East

RegionWinner_West

RegionWinner_South

RegionWinner_Midwest

CustomerPostalCode

CustomerPostalCodeLatitude

CustomerPostalCodeLongitude

New Features

DistanceToEast

DistanceToWest

DistanceToSouth

DistanceToMidwest

East_Wins

West_Wins

South_Wins

Midwest_Wins

East_Conference

West_Conference

South_Conference

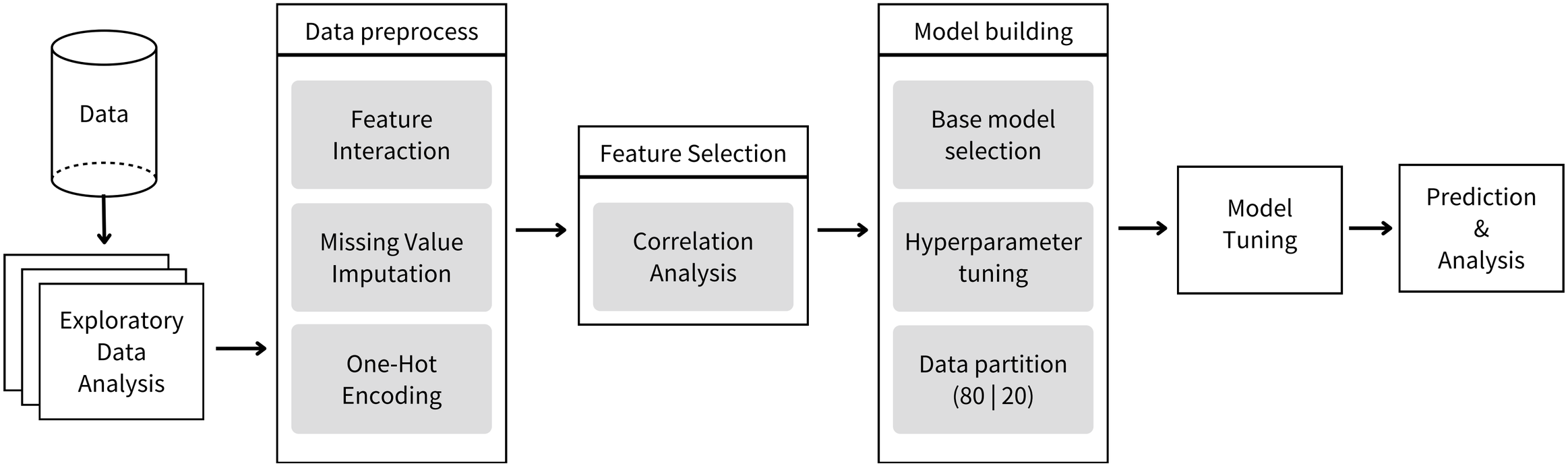
Midwest_Conference

East_West_Diff

South_Midwest_Diff

MODEL BUILDING & RESULT

METHODOLOGICAL WORKFLOW



BASE MODEL SELECTION

*validation accuracy score

Results for	Feature selection	Logistic Regression	Random Forest	XGBoost	LightGBM
SemifinalWinner_East_West	Existing Features	0.5303	0.6775	✅ 0.7015	0.3777
	with New Features	0.4311	0.6689	✅ 0.6982	0.1720
SemifinalWinner_South_Midwes	Existing Features	0.3919	0.6279	✅ 0.6554	0.2422
	with New Features	0.2963	0.6190	✅ 0.6527	0.1356
NationalChampion	Existing Features	0.3017	0.4508	✅ 0.4886	0.0863
	with New Features	0.2833	0.4428	✅ 0.4806	0.2788

MODEL BUILDING

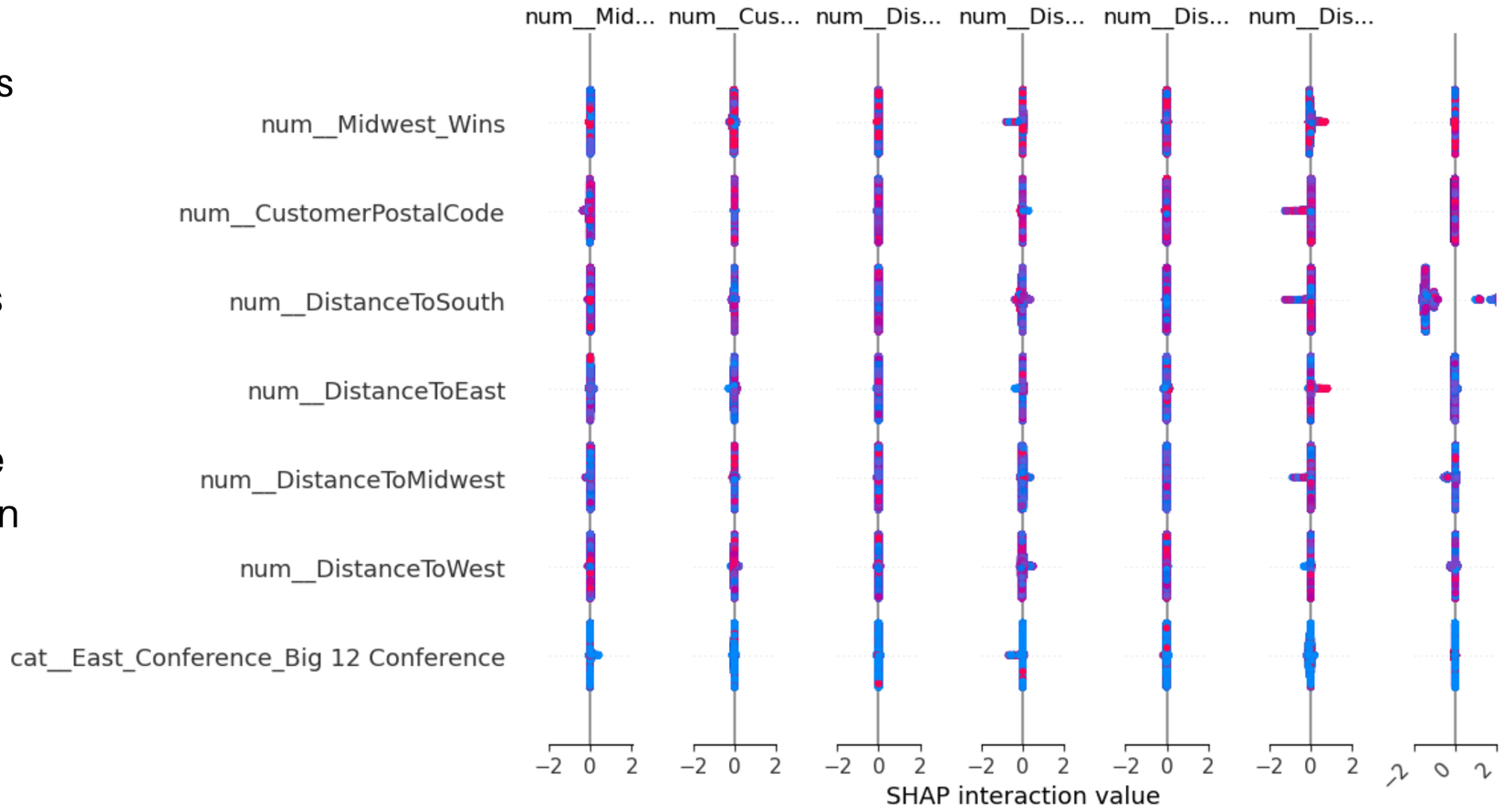
Hyperparameter Tuning

Max Depth of Tree: 9
Column Subsample Ratio of Tree: 0.5541
Min Sum of Instance Weight (hessian) per Child: 9
Gamma: 3.7034
Number of Trees (Estimators): 50
Evaluation Metric: mlogloss
Number of Jobs (Parallelization): -1

MODEL BUILDING

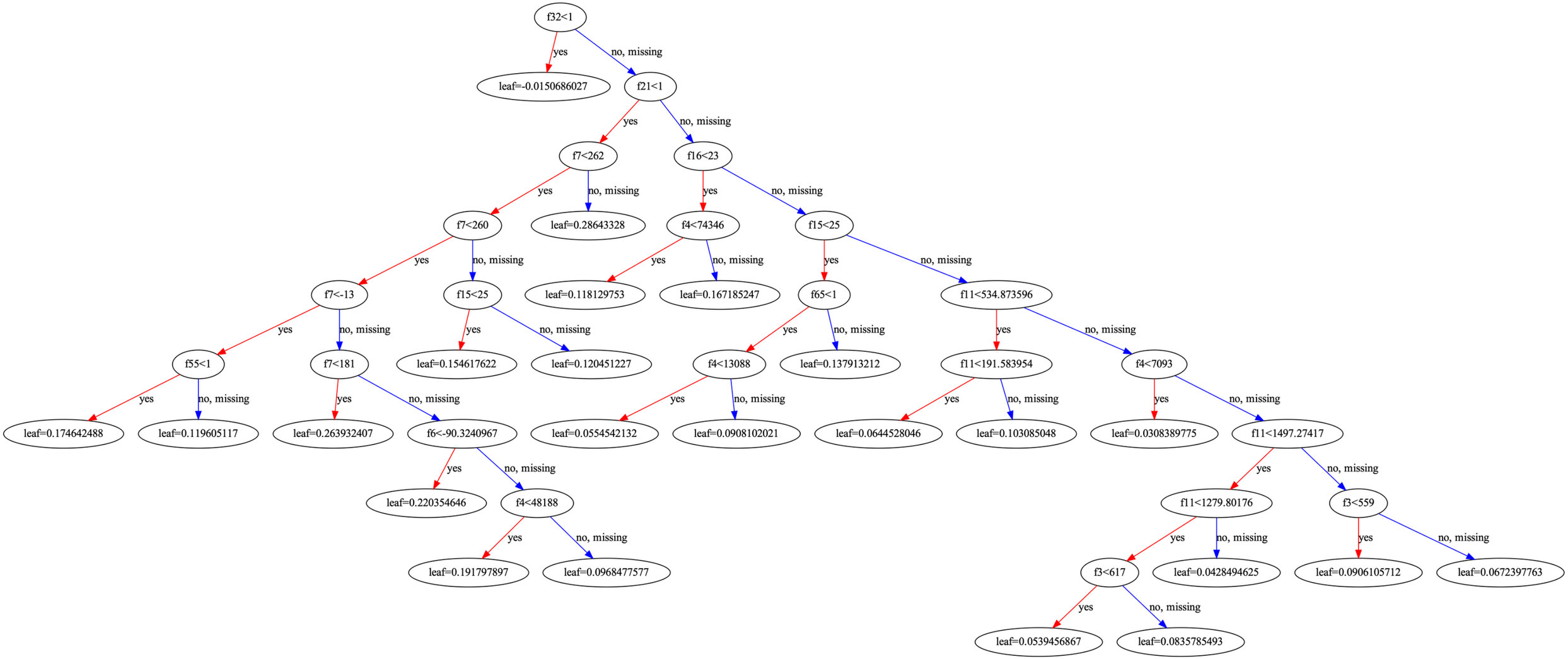
- We used SHAP to identify the most impactful features in our XGBoost model.
- These top 7 features contribute the most values in XGBoost.
- The Midwest_Wins feature actually takes a pivot role in contributing to our final model.

SHAP Feature Importance Bar Plot for Semifinal Winner East West

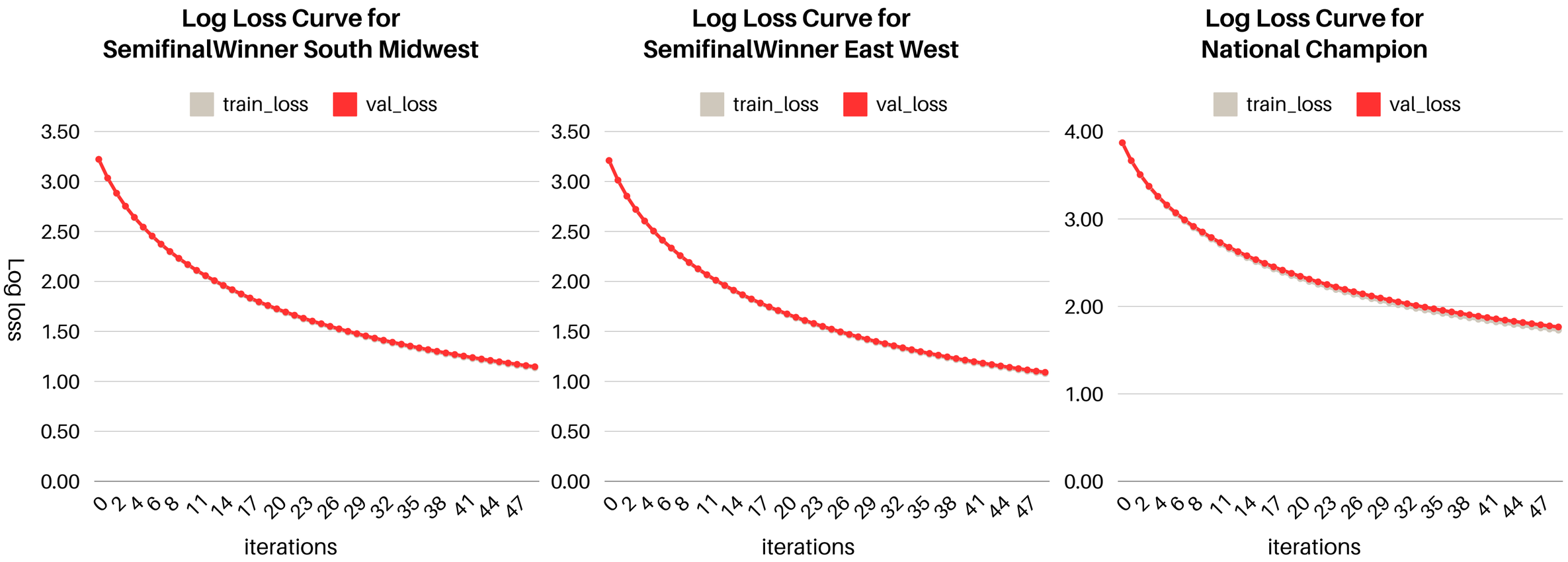


XGBOOST STRUCTURE

(20th tree as ref)

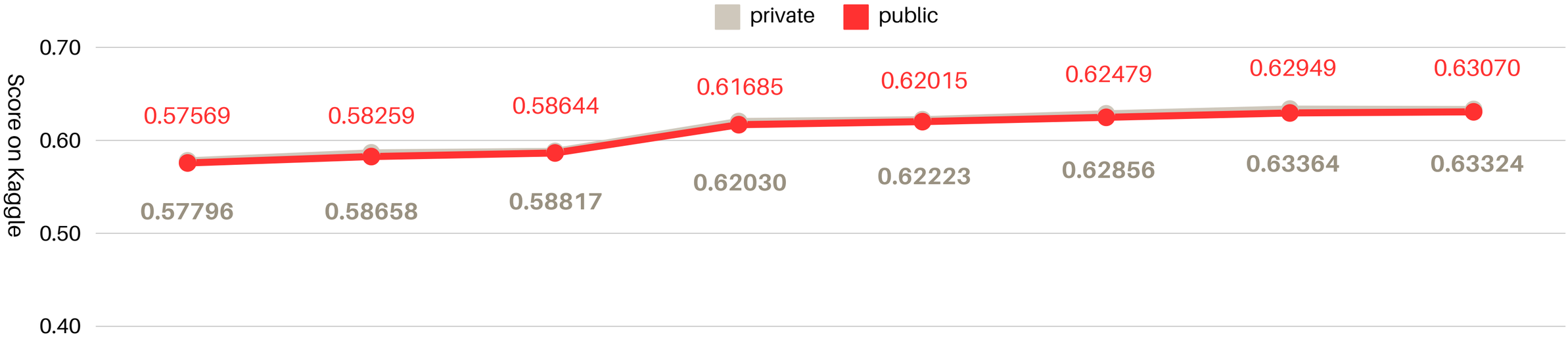


MODEL RESULT



The predictions for the three matches show that both the train log loss and validation log loss are gradually decreasing, indicating no overfitting issues.

MODEL TUNING



Model structure

Neural Network	Ensemble Random Forest + Catboost + XGBoost	Catboost	Ensemble Random Forest + Neural Network	Random Forest	Ensemble Random Forest + Catboost	Optimized ensemble	Optimized XGboost

SCHOOL AFFINITY ANALYSIS

Does consumer loyalty to the school influence our predictions?

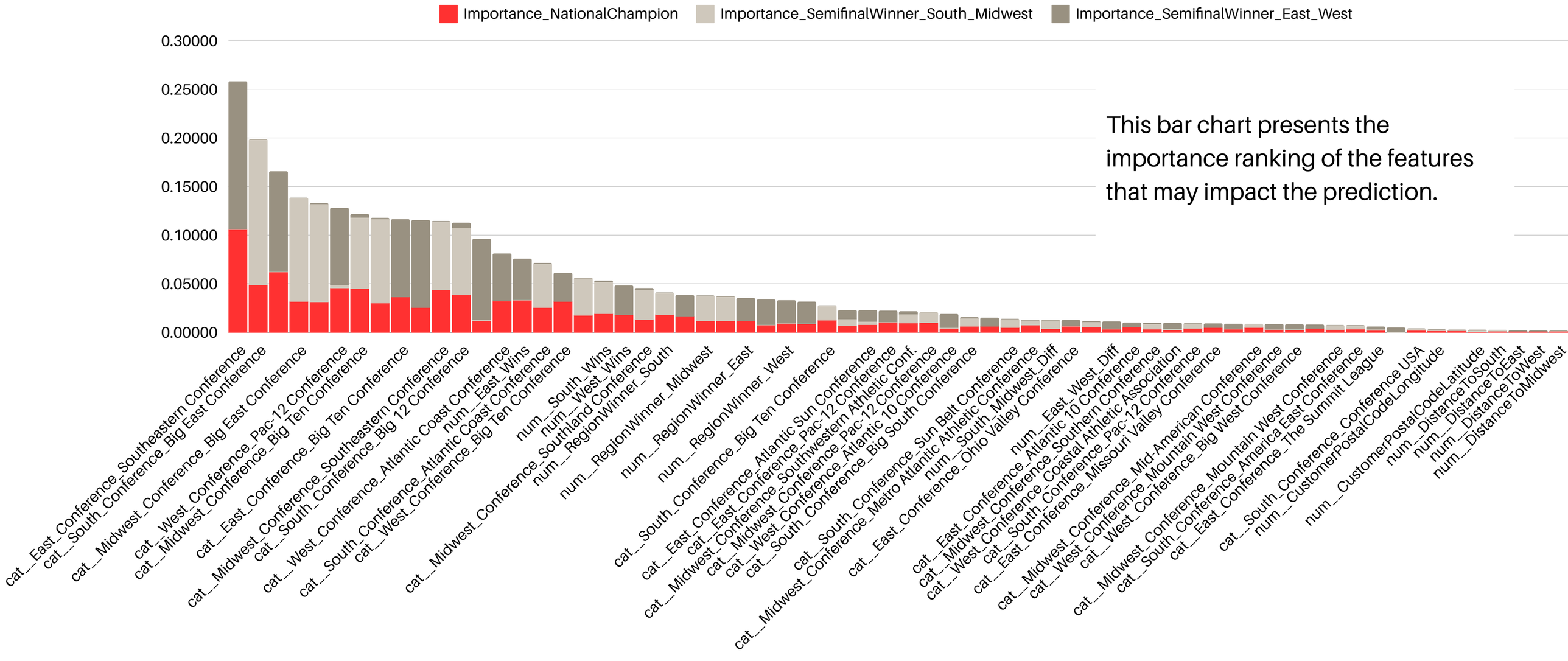
SCHOOL AFFINITY ANALYSIS

DistanceToEast
DistanceToWest
DistanceToSouth
DistanceToMidwest

Definition of School Affinity

Define the school affinity by the distance between the customer and institution, captured by the four new features. Then use feature importance ranking to determine if the features play a role in the predictions.

SCHOOL AFFINITY ANALYSIS



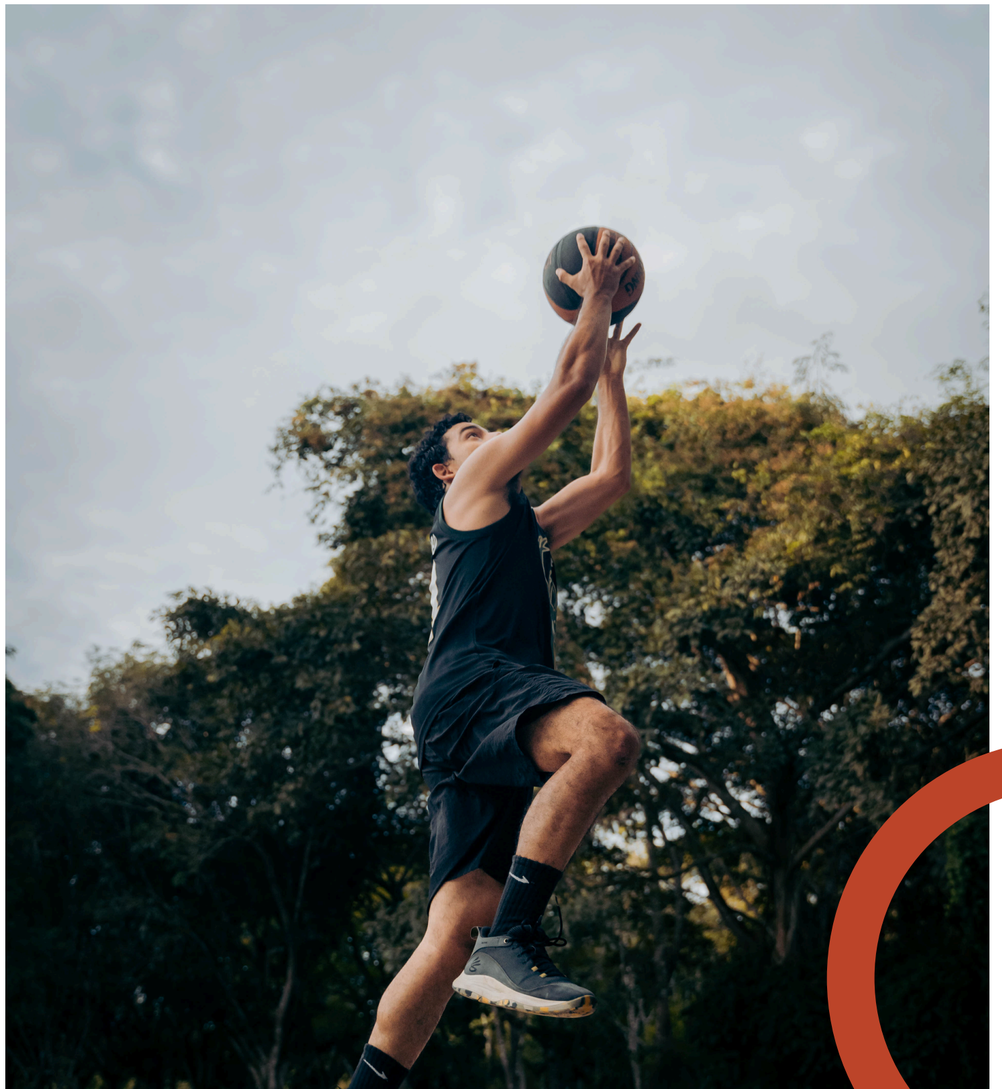
This bar chart presents the importance ranking of the features that may impact the prediction.

SCHOOL AFFINITY ANALYSIS

Conclusion

- As displayed in bar chart of feature importance, the four new distance features are ranked last. In conclusion, school affinity is unlikely to play a role in the predictions.

- The features in the higher position are conference features, indicating their strong influence in the predictions.



CCAC 25

THANK YOU

PRESENTED BY NIMBUS 2025

DA FANG LIN, HUNG-CHEN HSU, YIRAN LIU