CROSSROADS CLASSIC ANALYTICS CHALLENGE ‘25

TEAM: DATA CURRY

MEET THE TEAM

Ajay Shankar

Rohan

Krithiga Rajan

Subbaiah

Keerthi Anand

Problem Statement

🏀 **What?**

Analyze school affinity and other institution based feautures’ impact on a customers March Madness predictions.

🏀 **Why?**

Understand patterns and biases in predictions to refine models and understand more about the customer preferences.

🏀 **How?**

Build predictive models and visuals using bracket entry data, external sources, and Tableau.

How does the data look?

🏀 Bracket Entry Dataset

130,002 rows

17 columns

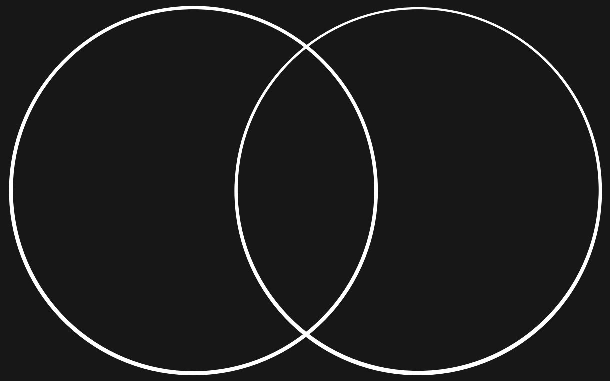
Has bracket entry predictions of customers and their postal codes, latitudes & longitudes

🏀 NCAA Institution dataset

68 rows

20 Columns

Has 68 institutes’ data, including wins, losses average score, attendance enrollment etc.



🏀 Consolidated data set

Has 130,002 rows

135 columns

Consolidated daset after joining the above 2 dataset using 4 region IDs

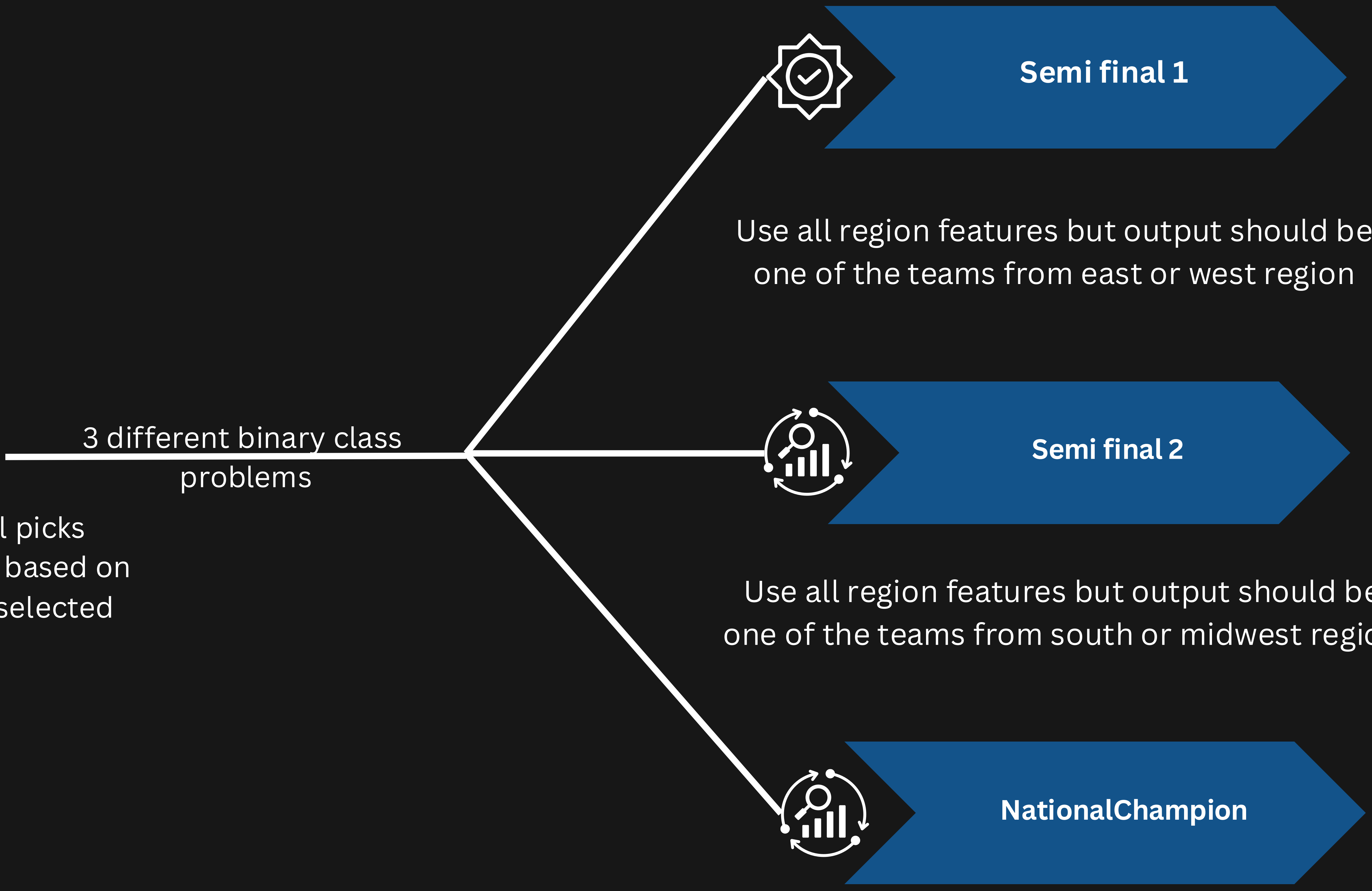
Exploratory Data Analysis in Tableau

Now Let’s talk about the Model

Model Strategy

Goal

The goal is to predict a customer's semifinal picks (Semifinals 1, 2,) and national champion model based on their bracket entries, their distance from the selected institute and team stats



3 different binary class problems

Semi final 1

Use all region features but output should be one of the teams from east or west region

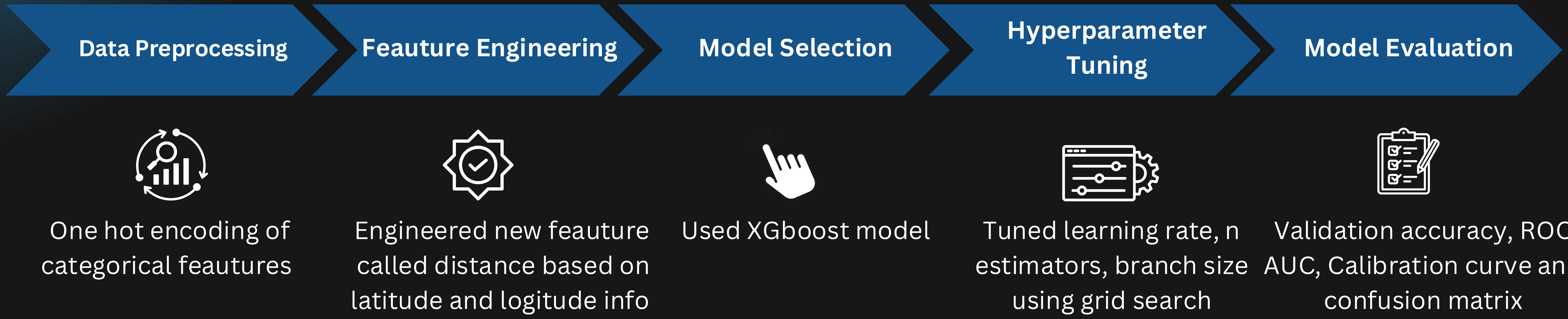
Semi final 2

Use all region features but output should be one of the teams from south or midwest region

NationalChampion

Use all region features but output should be one of the teams from semi 1 or 2

Model Building



Data Preprocessing

One hot encoding of categorical feautures

Feauture Engineering

Engineered new feauture called distance based on latitude and logitude info

Model Selection

Used XGboost model

Hyperparameter Tuning

Tuned learning rate, n estimators, branch size using grid search

Model Evaluation

Validation accuracy, ROC AUC, Calibration curve and confusion matrix

Our Winning Model

XGboost

We selected XG boost because of the feature interpretability and its ability to handle missing and skewed data

Binary class

Treated the problem as 3 different binary class problems-predicting semi 1, 2 and national champion

Accuracy

Semi 1 Accuracy - **69.4%**
Semi 2 Accuracy - **64.9%**
National Champion - **63.6%**

Other models considered

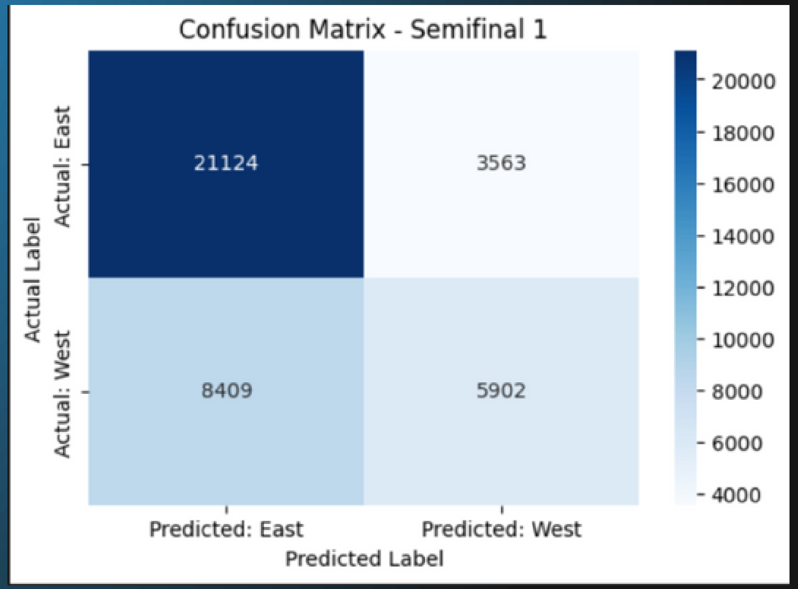
Considered logistic regression and random forest techniques.

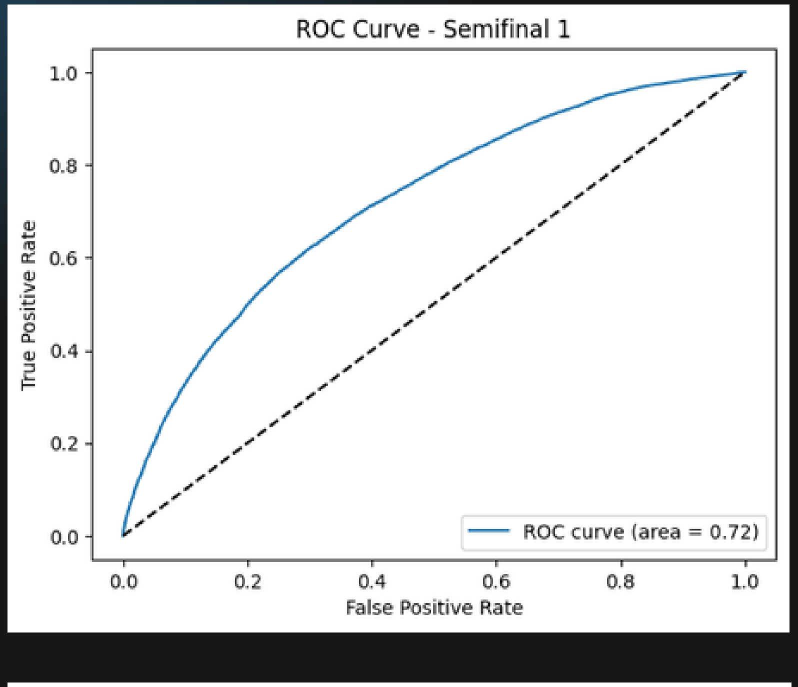
SHAP

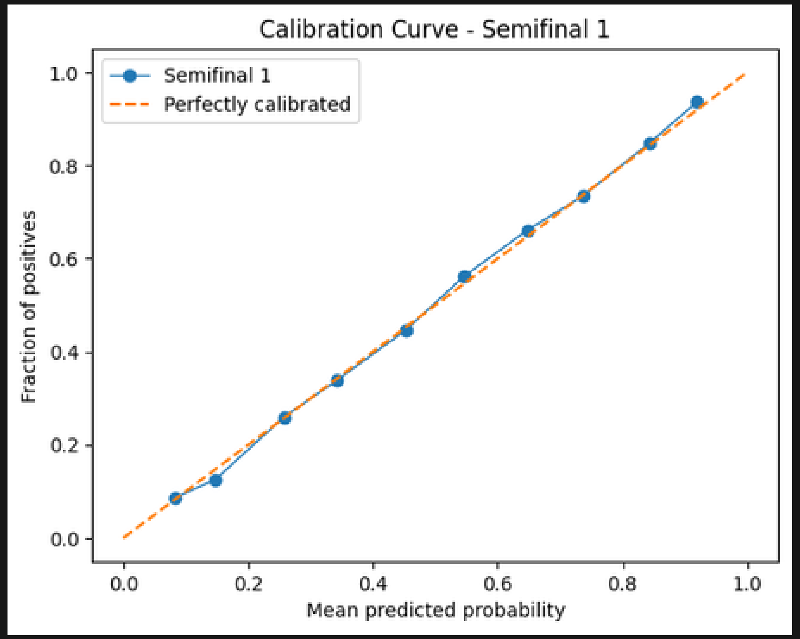
Used SHAP (SHapley Additive exPlanations) to get feature importance and infer results

Model Evaluation

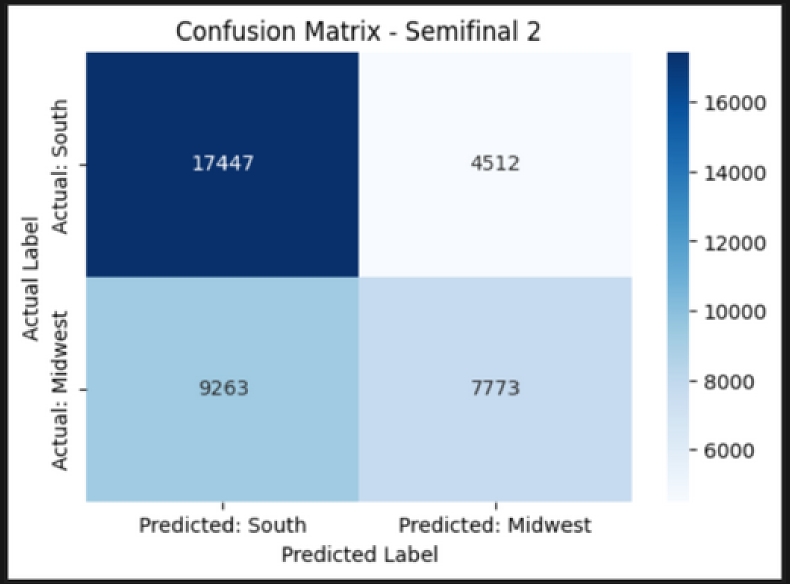
Semifinal 1

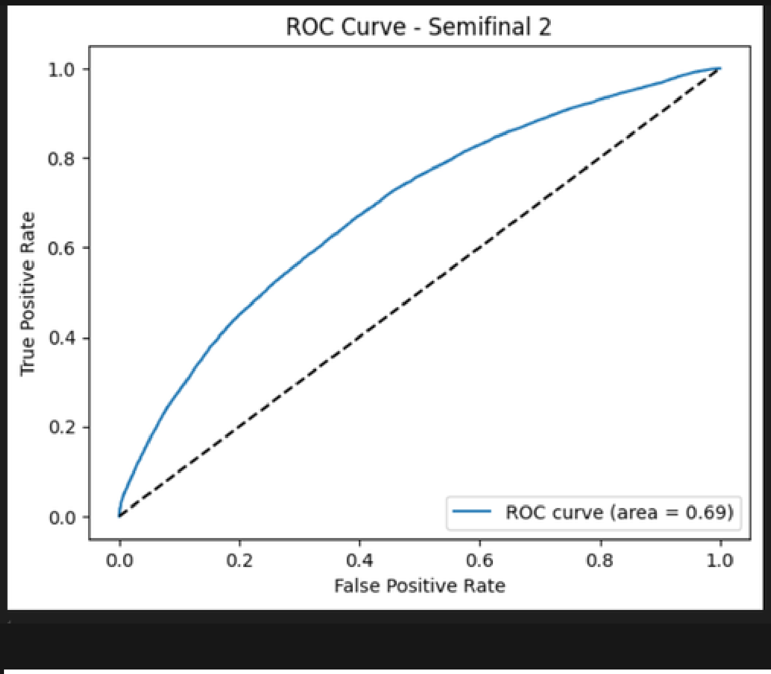


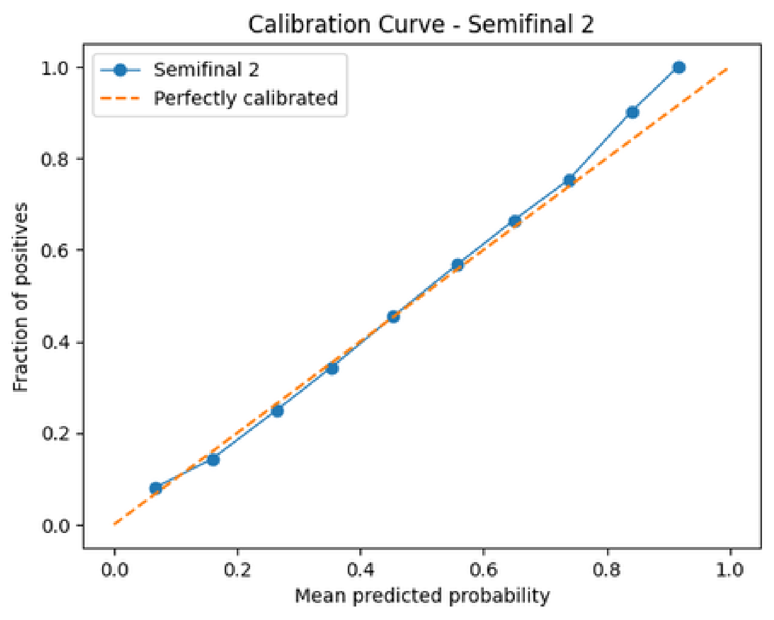




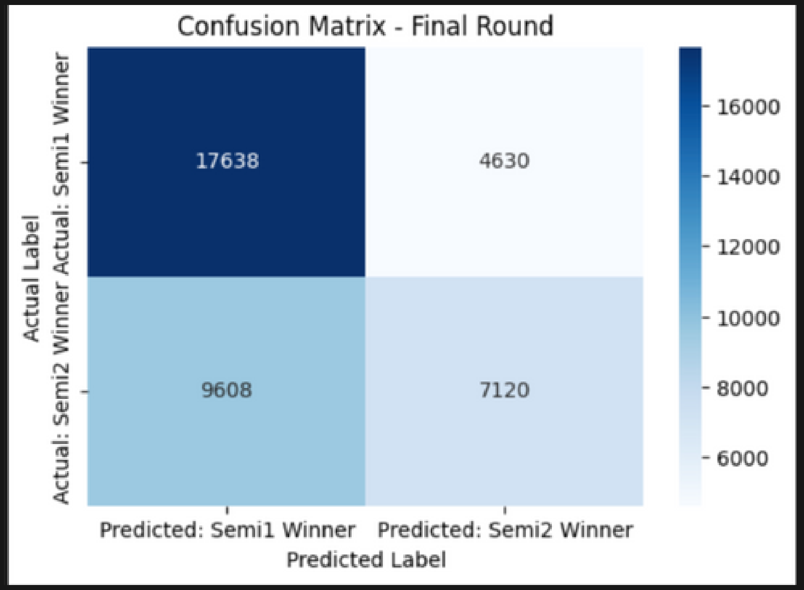
Semifinal 2

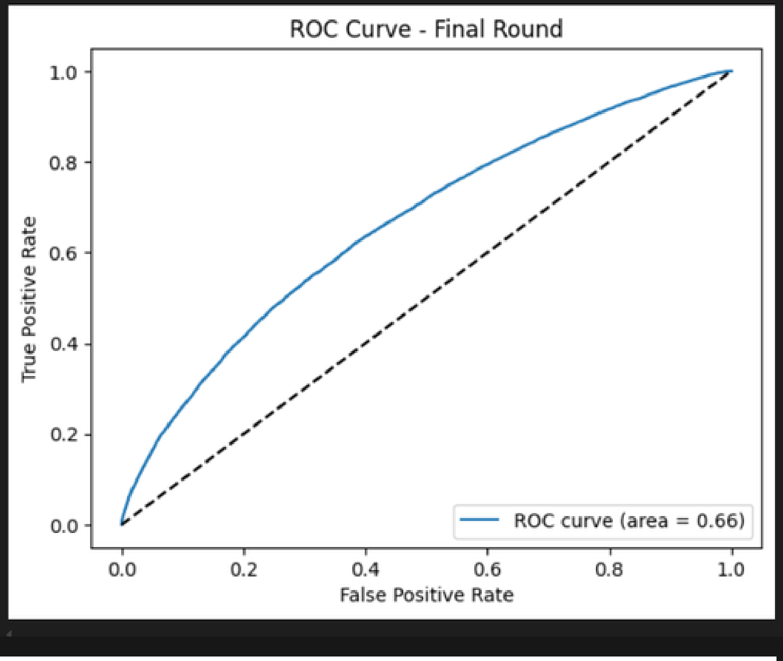


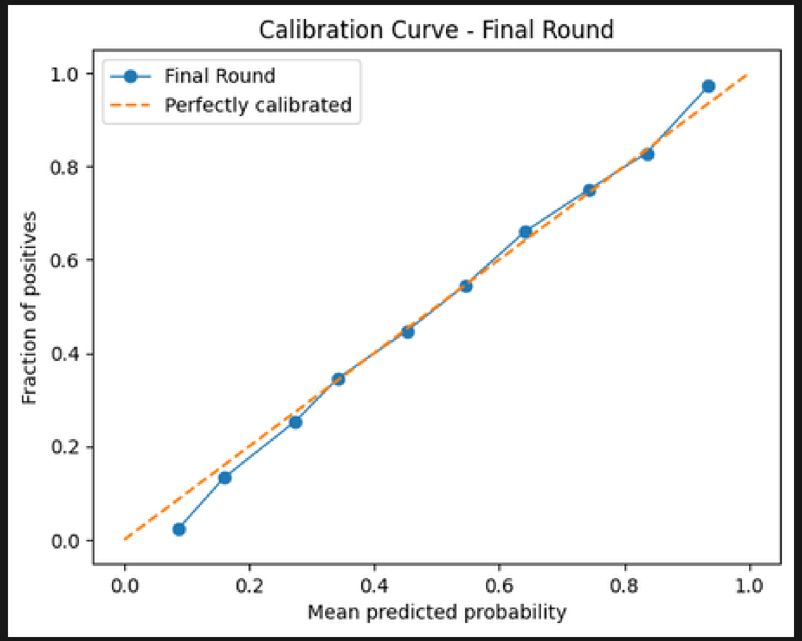




National Champion

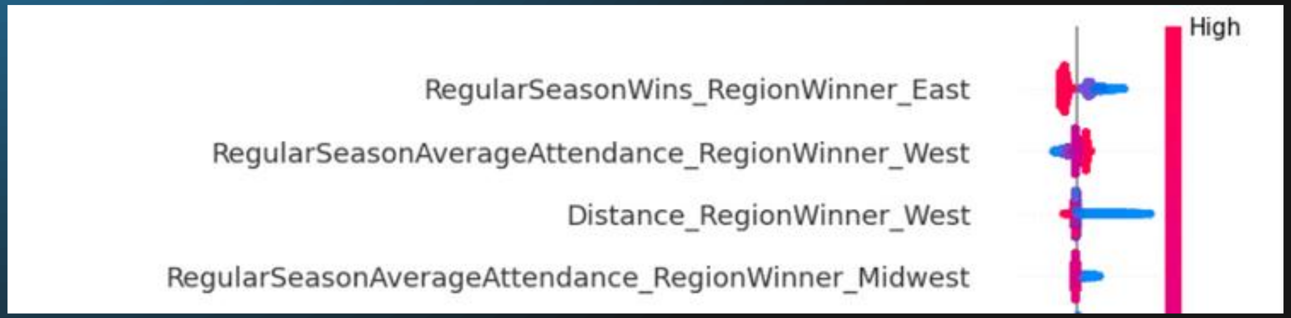






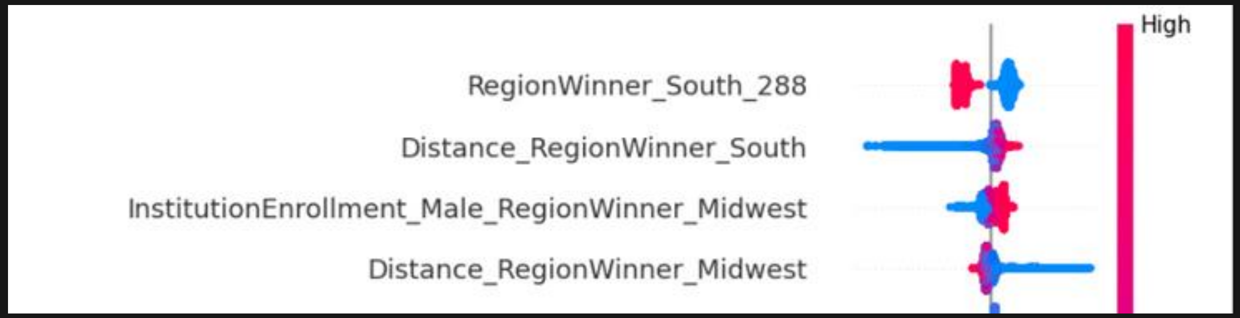
Model Inferences

Semifinal 1



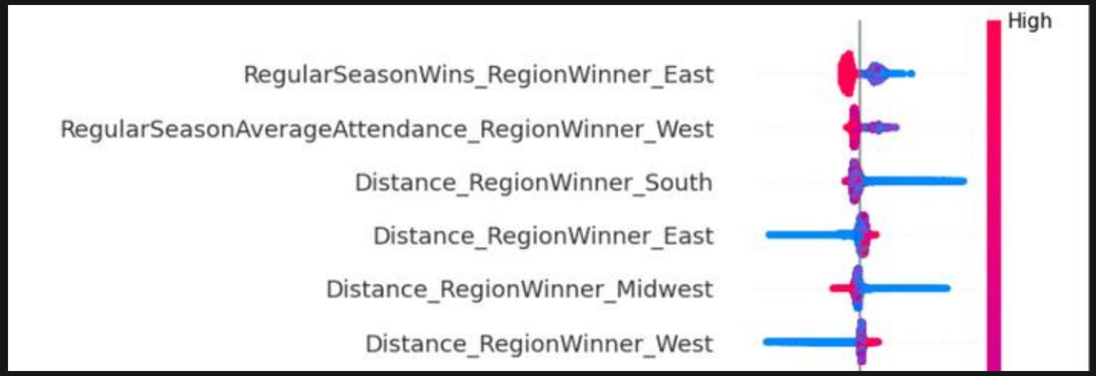
- **Regular Season Wins (East) – The Strongest Factor - Huskies**
- **Regular Season Attendance (West) - Proxy for affinity**
- **Distance Between the Fan & the Institution (West) – Regional Bias**
- **Midwest Attendance - Indirect Impact on Fan Predictions**

Semifinal 2



- **RegionWinner_South_288 – The Biggest Factor - Cougors**
- **Distance Between the Fan & the Institution (South) – Regional Bias**
- **School Size - preference for larger schools.**
- **Distance Between the Fan & the Institution (Midwest) – More Regional Bias**

National Champion

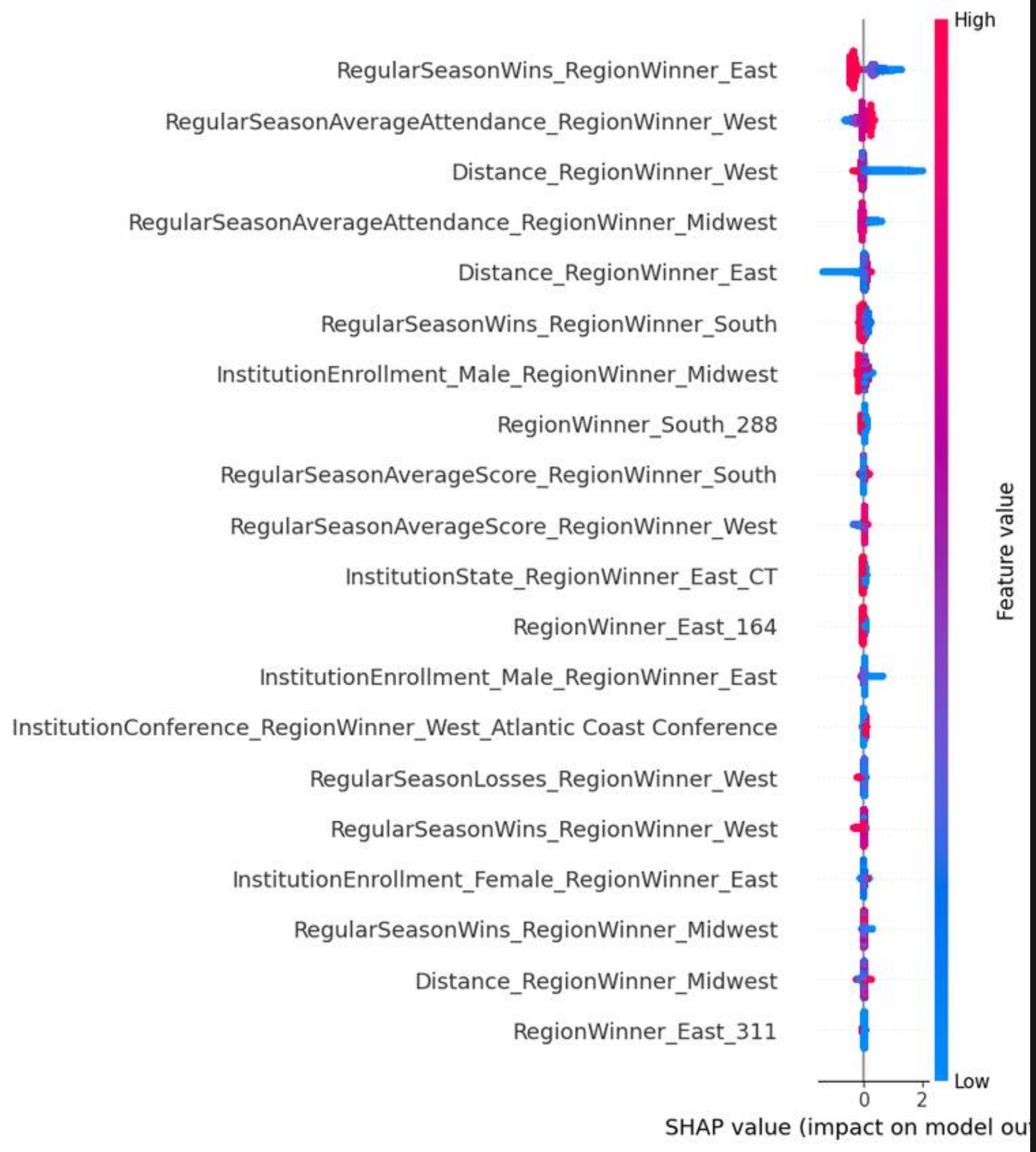


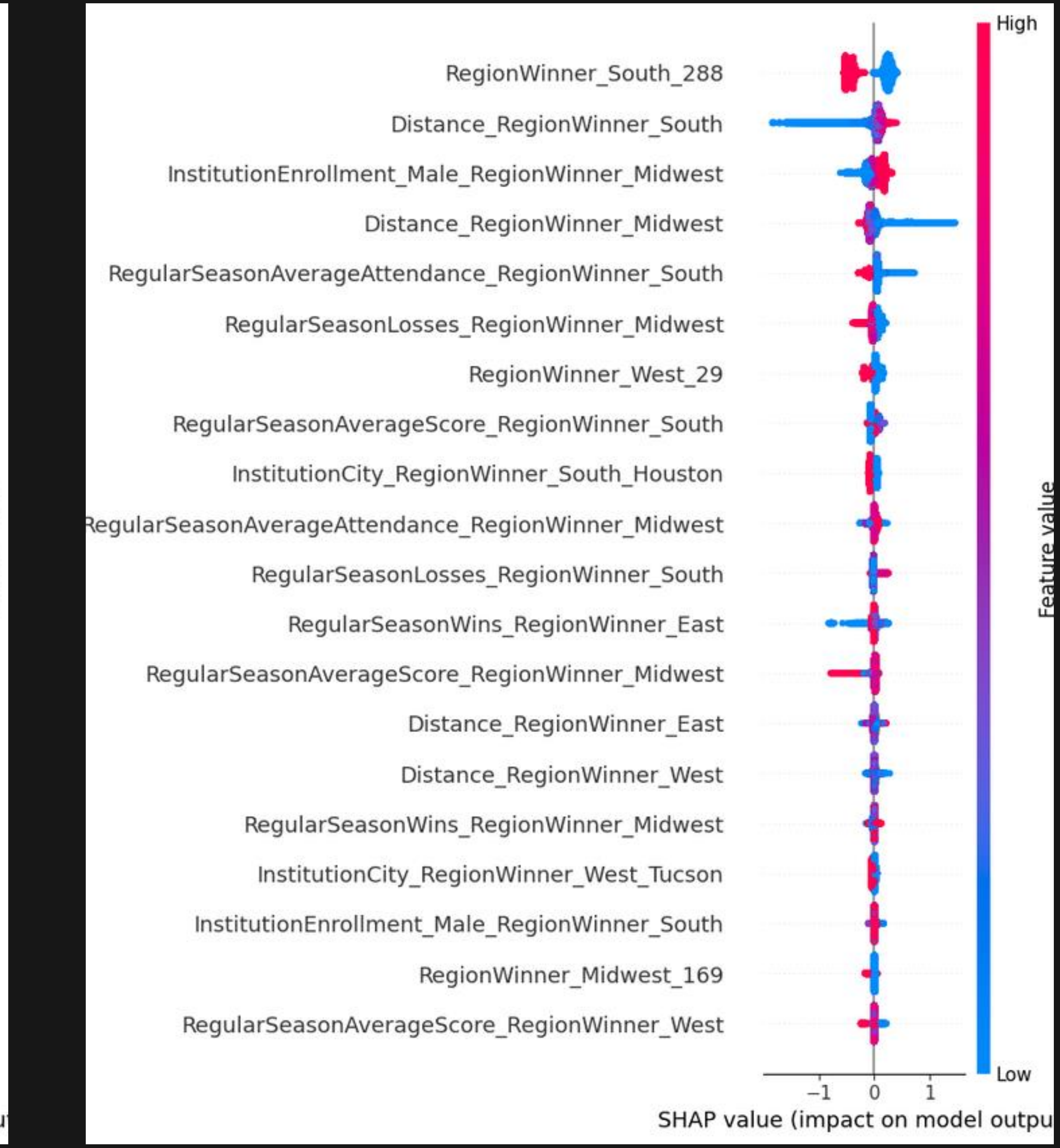
- **4 out of 6 parameters that affect the final are “distance”**
- **Clear indication of how affinity affects fans predicting the bracket**
- **If the team from semifinal-1 is UCONN, fans choose them as National champions in most cases.**

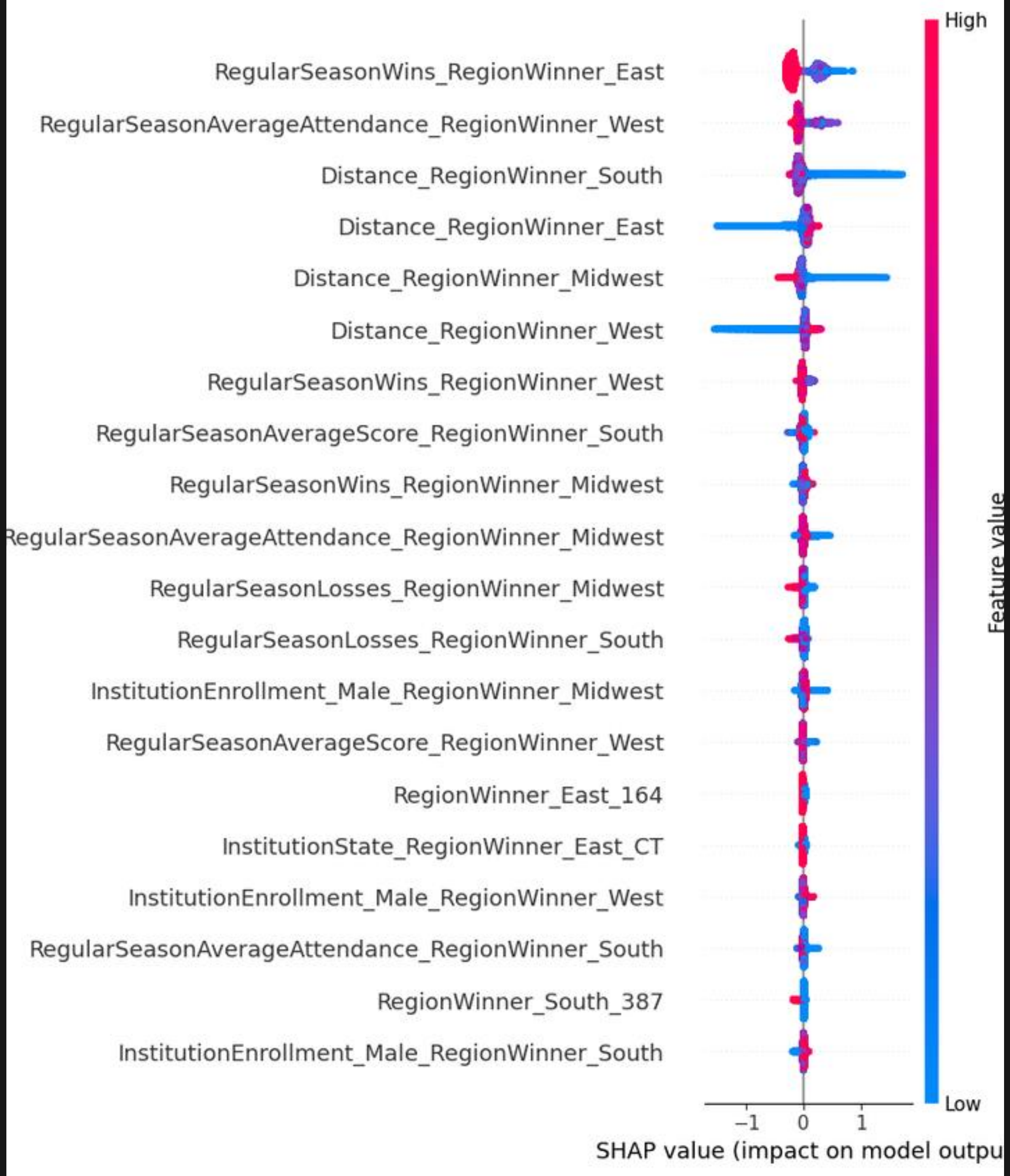
East West Model

South Midwest Model

National Champion Model







What We Recommend

- Yes, distance plays a role in affinity, at least for stronger teams. So we recommend they engage and involve fans during home games

- Texas has high school affinity and higher fan engagement- which makes it a good place to target investments and other promotional events

- Stats play a role!- NCAA can build an interactive or AI-based stats dashboard and tools to increase engagement, as its evident