NCAA Bracket Predictive Analysis - Query Queens

Leanna Jeon
Tzu Yun Huang
Yi Shiuan Chiang
Shih Min Lin

PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

Agenda

01 **About us**
Introduction to our team and project goals

02 **Tableau Insights**
Key data visualizations and trends from our analysis

03 **Basic Modeling**
Foundational analysis of bracket predictions

04 **Advanced Model**
Deeper predictive insights

About Us

Leanna Jeon
Department of Management
Purdue University
West Lafayette, USA
jeon139@purdue.edu

Yi Shiuan(Elsa) Chiang
Department of Management
Purdue University
West Lafayette, USA
chian130@purdue.edu

Shih Min(Hazel) Lin
Department of Management
Purdue University
West Lafayette, USA
lin1944@purdue.edu

Tzu Yun Huang
Department of Management
Purdue University
West Lafayette, USA
huan2261@purdue.edu

Tableau Insights

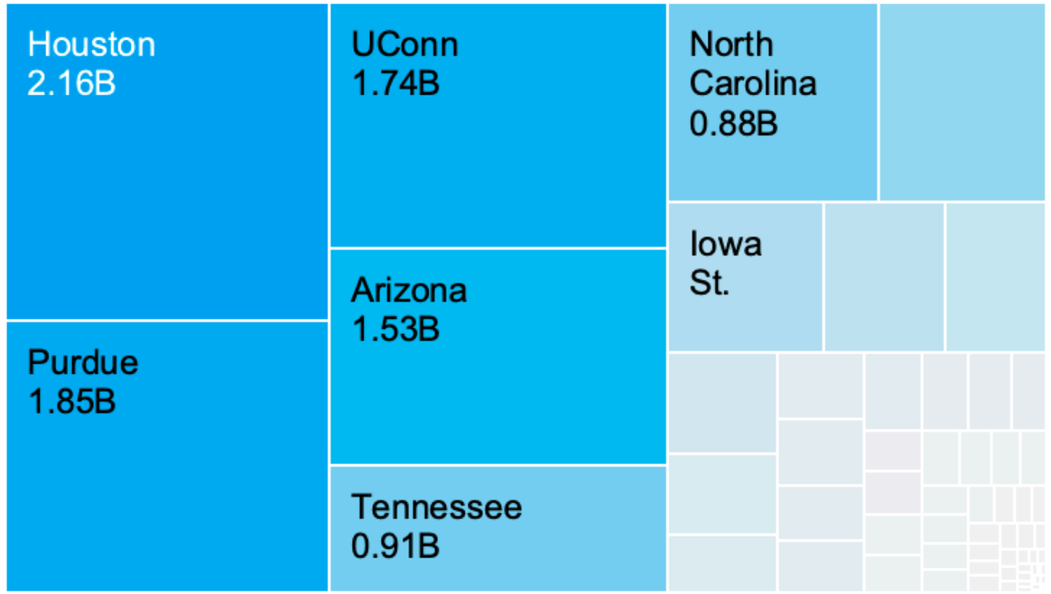
Performance vs. Affiliation in Bracket Picks

What influences their choices: Performance or Loyalty?

- School Size
- Team Performance
- Affiliation

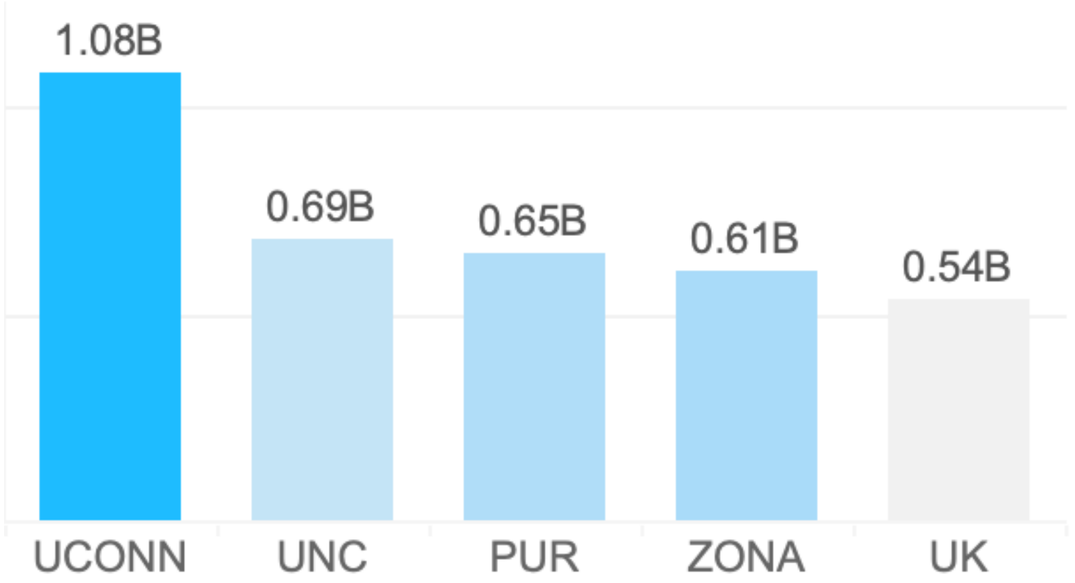
How School Size & Fan Base Impact Bracket Picks

Number of Enrollment by Institution



- Top schools with the most enrollment: **Houston** (2.16B), **Purdue** (1.85B), **UConn** (1.74B), and **Arizona** (1.53B)
- Larger institutions tend to appear in the semifinals or finals more often.

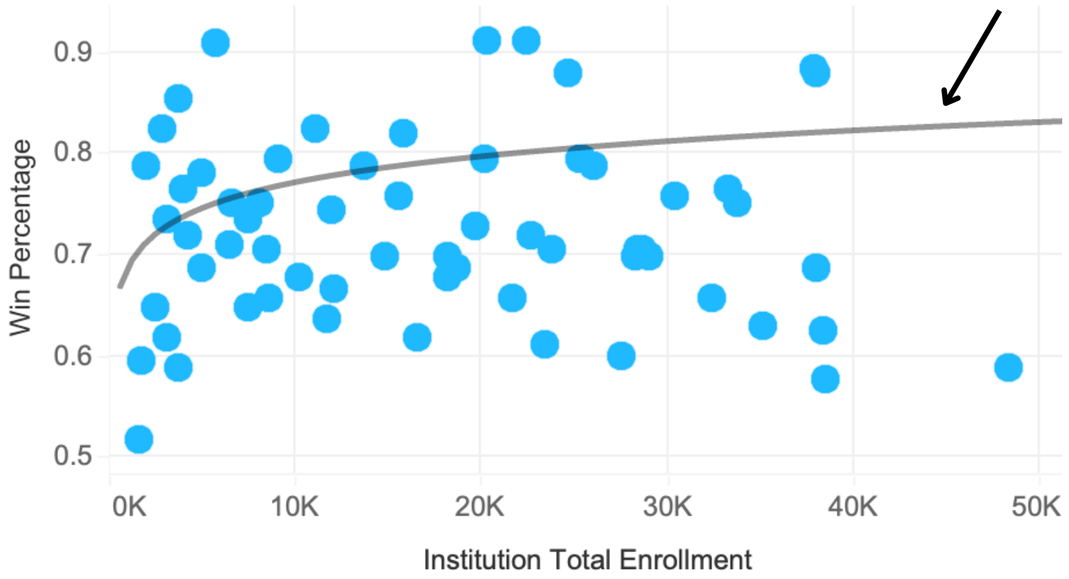
Top 5 Attendence by Institution



- Top 5 institutions with the highest attendance are **UConn**, **UNC**, **Purdue**, **Arizona**, and **Kentucky**.
- More students = More potential fans and alumni attending games.

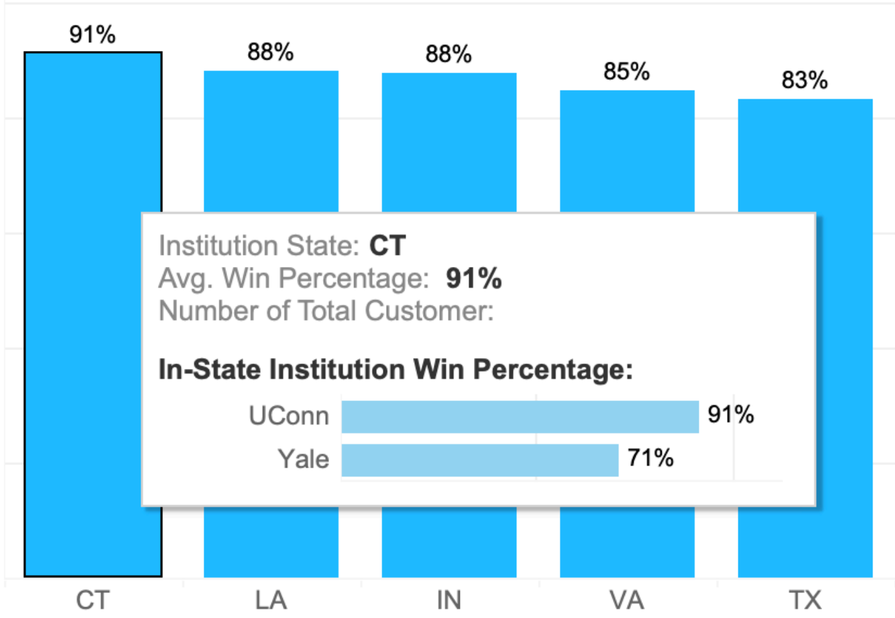
Does School Size Predict Success?

Team Performance vs. Enrollment Size



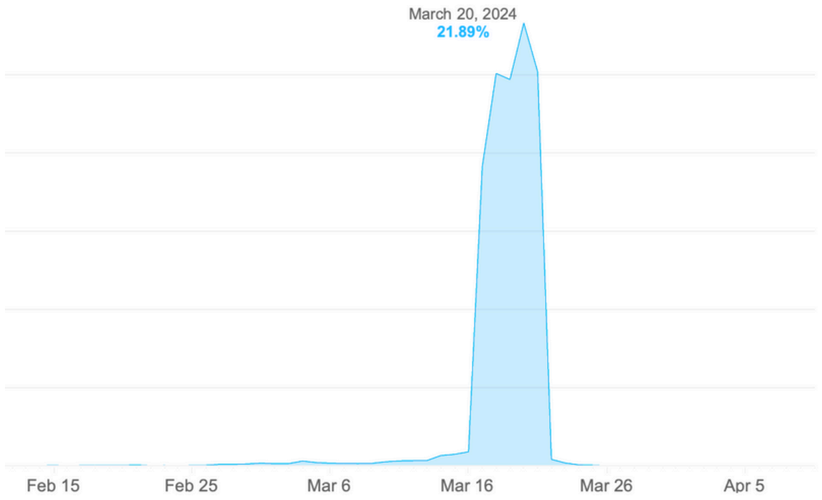
- For **small schools** (<5K), resources significantly impact win percentage.
- But for **large schools**, additional enrollment doesn't strongly correlate with better performance.

Top 5 Win Percentage by State



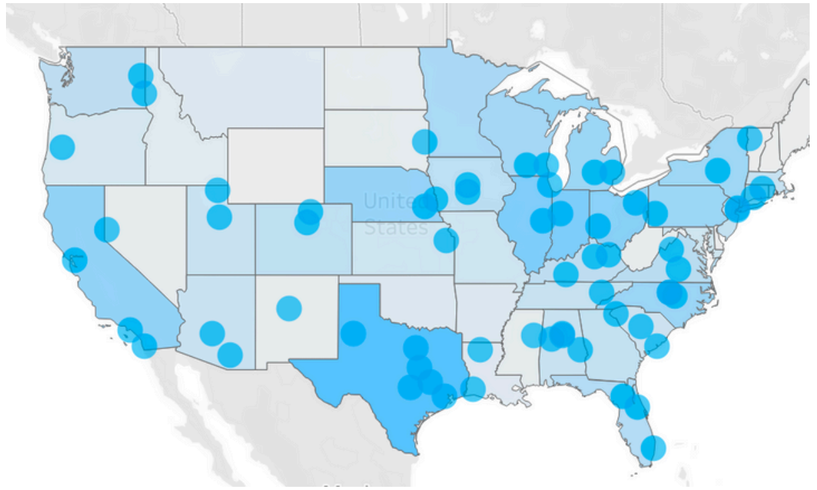
- **CT** (91%), **LA** (88%), and **IN** (88%) have the highest win rates, showing strong basketball traditions in these states.
- **Connecticut's** high percentage is largely driven by **UConn's** success.

What Influences Picks: Performance or Affiliation?



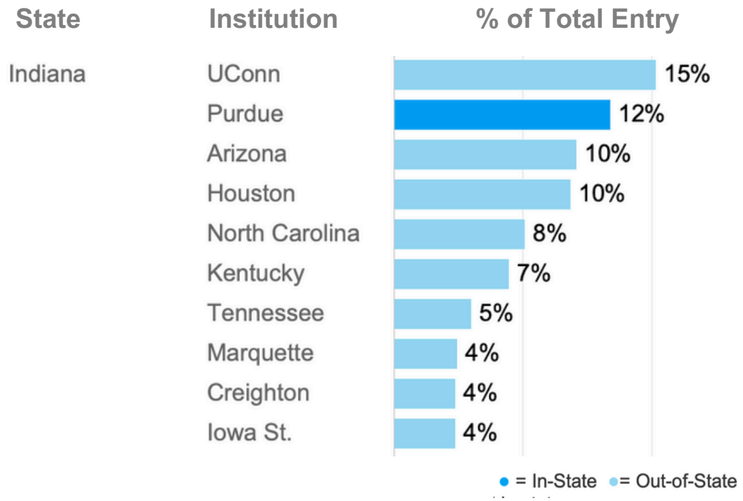
1.Bracket Submission Trends

- March 20 saw 21.89% of all bracket submissions.
- Most fans wait until the last moment to make picks.



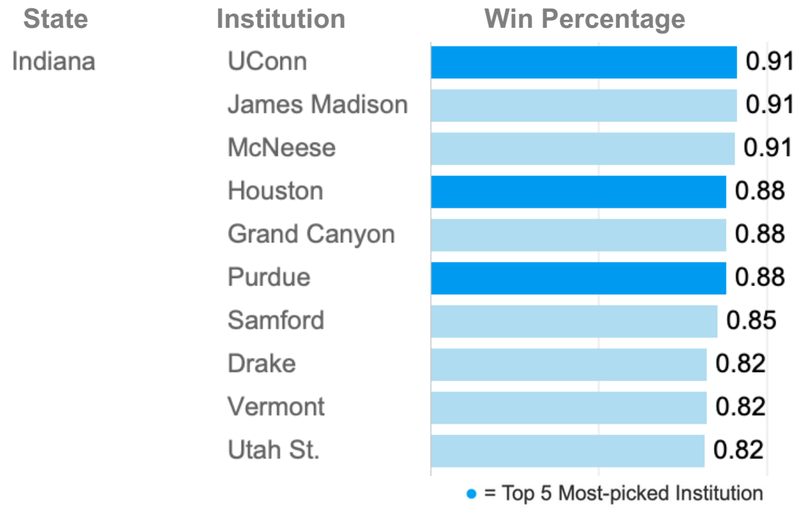
2.Map

Texas (6.7% total bracket participation) has a high number of participants, likely due to Houston's presence.



3.Team-Picked by Institution

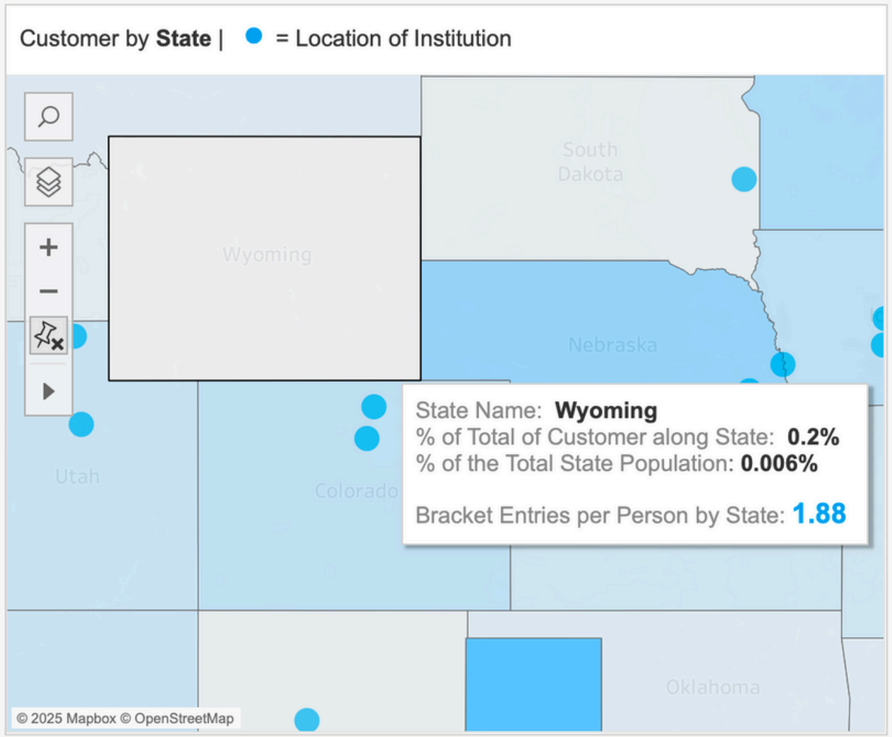
- Shows which teams are most picked within each state.
- Purdue ranks high in its home state, but UConn (15%) is still #1.

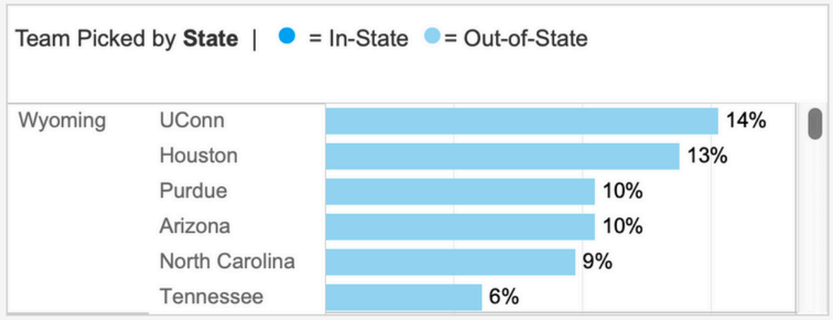


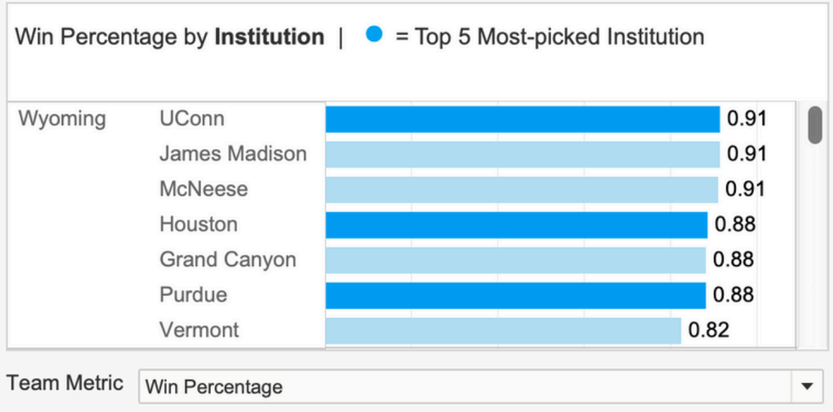
4.Win Percentage by Institution

UConn, **Houston**, and **Purdue** are among the top teams in win percentage and also rank in the top 5 most-picked teams in brackets.

States without any institutions— Wyoming





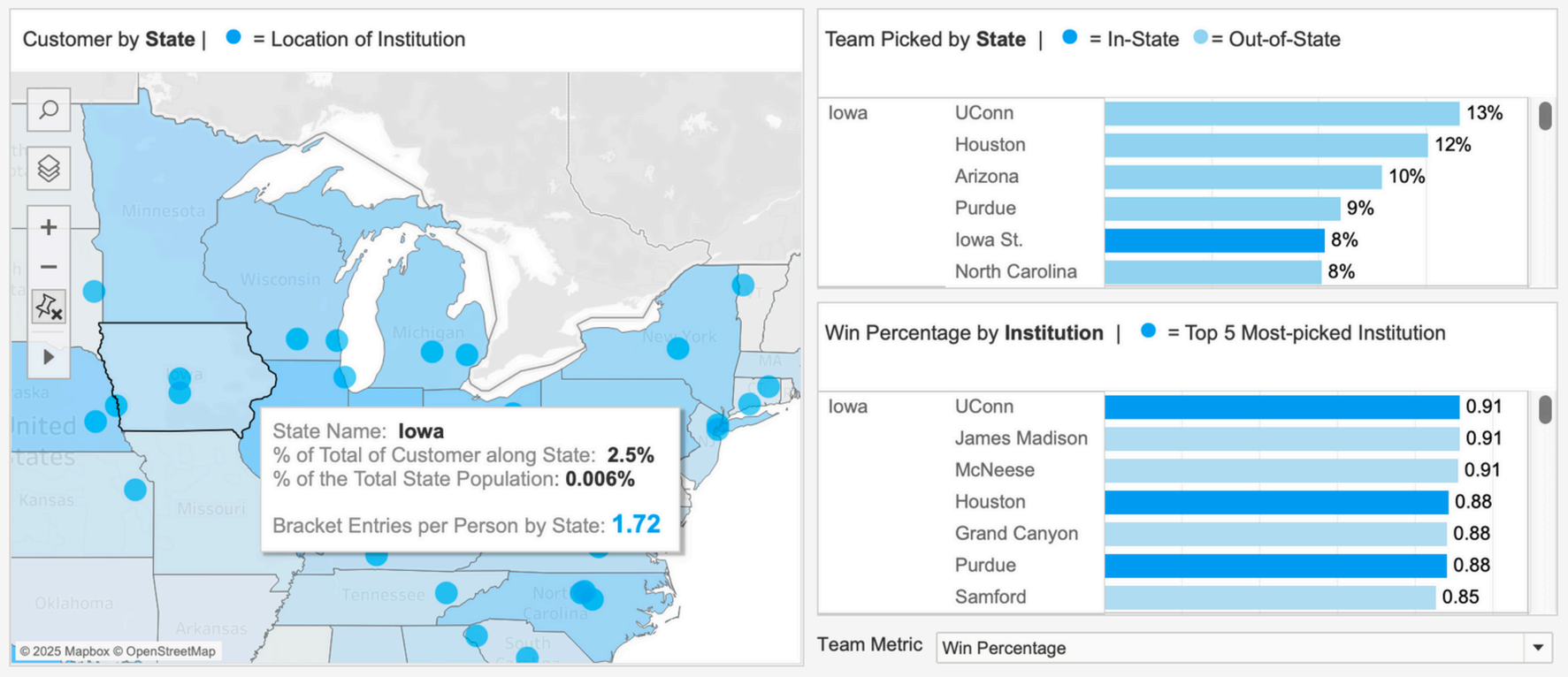


All of the most-picked teams are out-of-state.

Among the top 7 high-performing schools, 3 are within Wyoming's top 5 most-picked teams.

Performance is the primary factor influencing team selection, as there is no local affiliation.

States without strong institutions— Iowa: Iowa St.



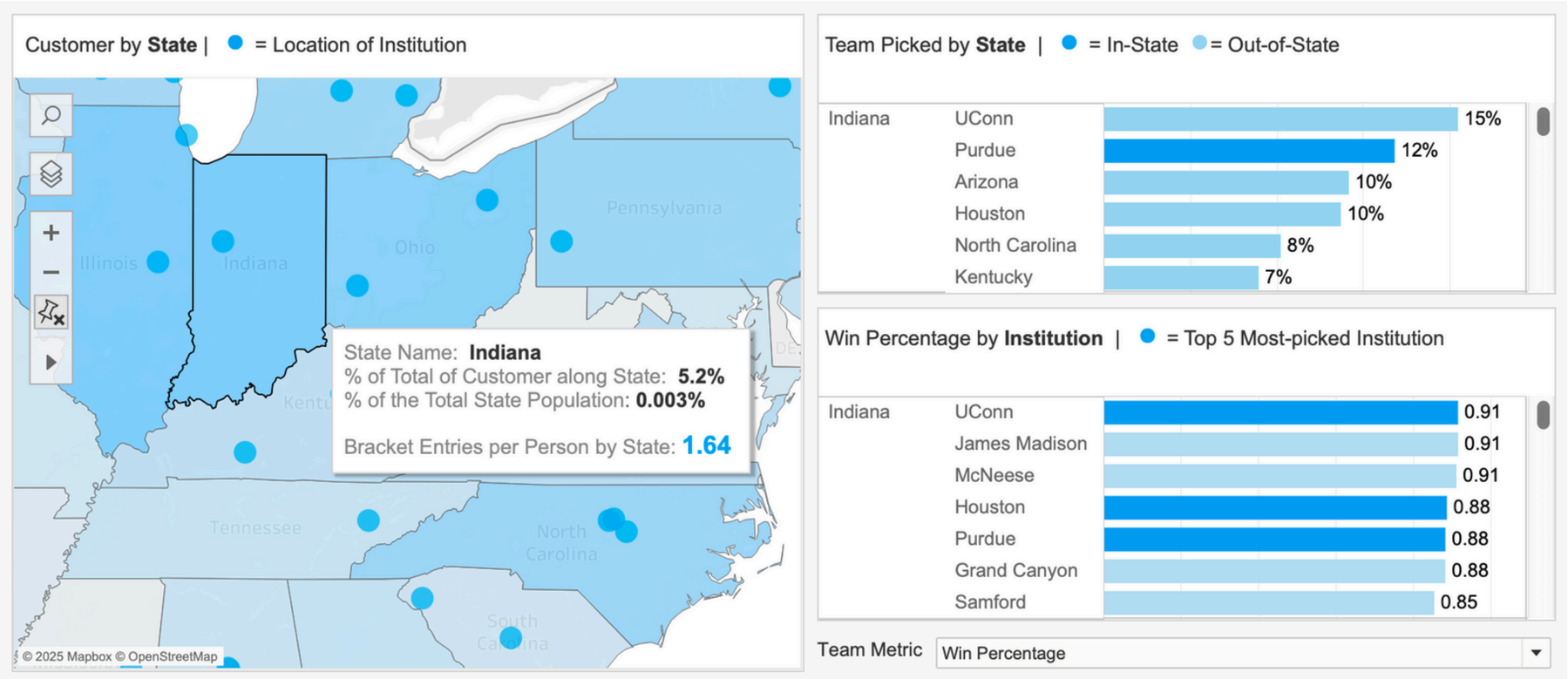
One of the most-picked teams is in-state.

Among the top 7 high-performing schools, 3 are within Iowa's top 5 most-picked teams.

In states without strong schools, affiliation has a certain level of influence, but performance remains the primary factor in team selection.

States with strong institutions— Indiana: Purdue

In states with high-performing schools, bracket challenge participation tends to be higher.

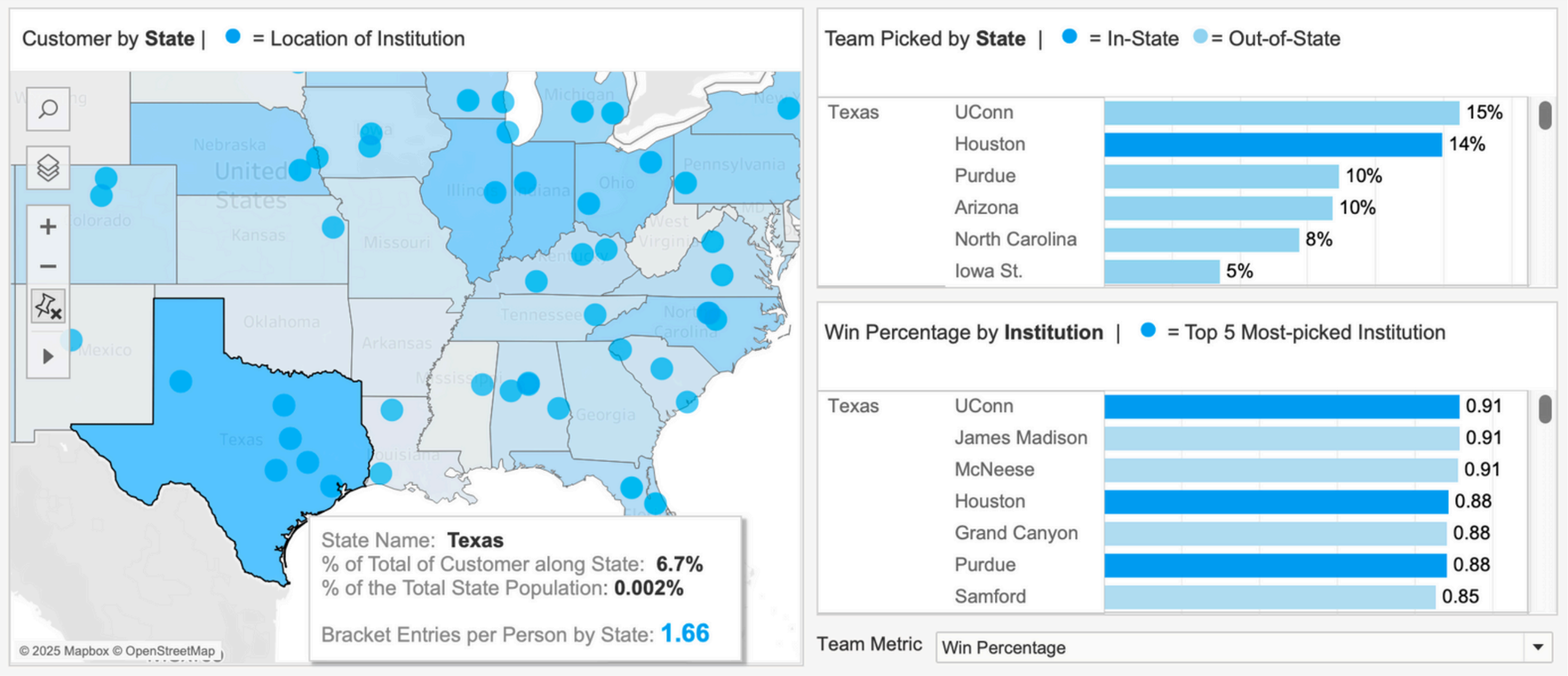


One of the most-picked teams is in-state.

Among the top 7 high-performing schools, 3 are within Indiana's top 5 most-picked teams.

States with strong institutions— Texas: Houston

In states with high-performing schools,
bracket challenge participation tends to be higher.



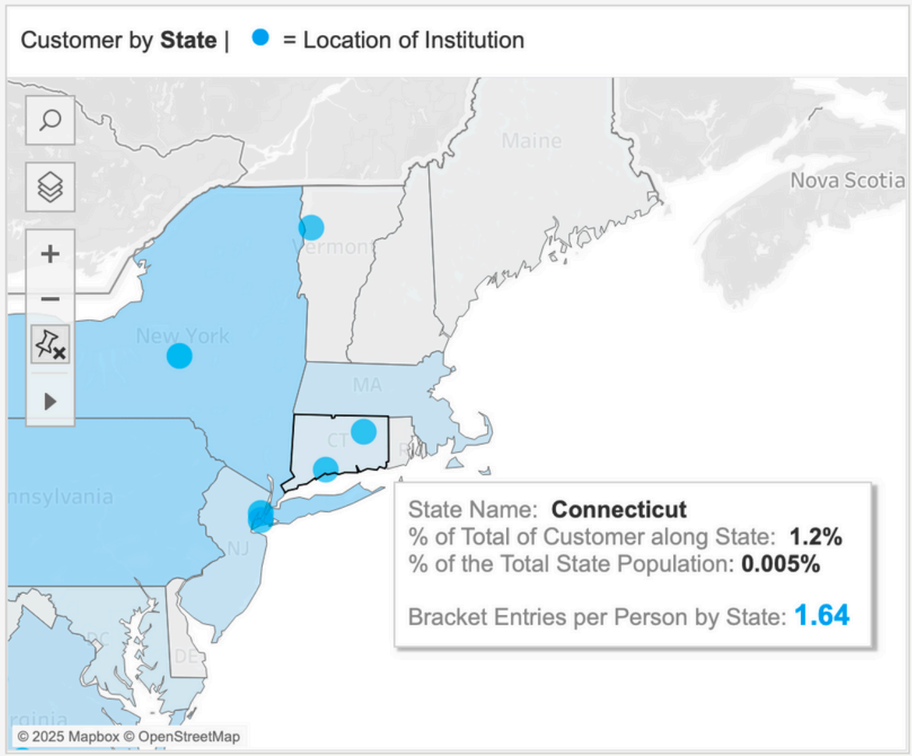
One of the most-picked teams is in-state.

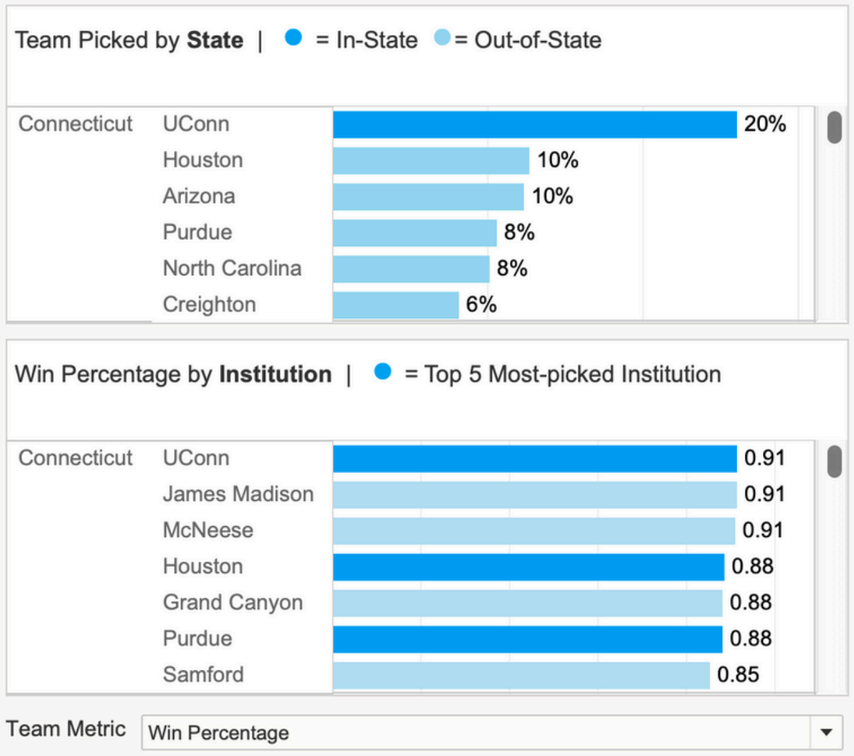
Among the top 7 high-performing schools, 3 are within Texas's top 5 most-picked teams.

In states with strong schools, performance remains the primary factor influencing team selection. However, highly ranked in-state schools tend to have a higher pick percentage within their state.

States with strong institutions— Connecticut: UConn

Although Connecticut's overall participation seems lower, its participant-to-population ratio is higher than in Indiana and Texas.





UConn is the top-picked team in Connecticut, with 20% of participants selecting them.

Among the top 7 high-performing schools, 3 are within Connecticut's top 5 most-picked teams.

The pick rates for Houston and Purdue are lower than in other states, indicating stronger local affiliation in Connecticut.

Conclusion

Key Insights

- Larger schools boost engagement, leading to more bracket submissions.
- Team selection is primarily influenced by team performance.
- Affiliation remains a key factor in decision-making, especially in states with strong teams.

Bracket Strategy

- Team Performance
- Affiliation & Loyalty
- School Size

Based on these insights, we developed our predictive model.

BASIC MODELING

Data Preprocessing

Feature Engineering

Win Percentage (Wins/Total Games)

Tenure (This Year – The year the institution joined NCAA)

Merge institution

Region Choices

West
East
South
Midwest

Institution ID

Institution Data

Average Score
Tenure
Win Percentage
Institution Enrollment

Modeling Methodology

Why 3 separate models?

Performance

- Win Percentage
- Average Score
- Tenure
- Average Attendance

Geographic

- Latitude
- Longitude
- DMA Code

Scale

- Institute Enrollment(Total)
- Institute Enrollment(Male)
- Institute Enrollment(Female)

Diverse characteristics improve ensemble model effectiveness

Basic Model

XGBOOST Ensemble Model

SOFT VOTING

Performance + Geographic + Scale

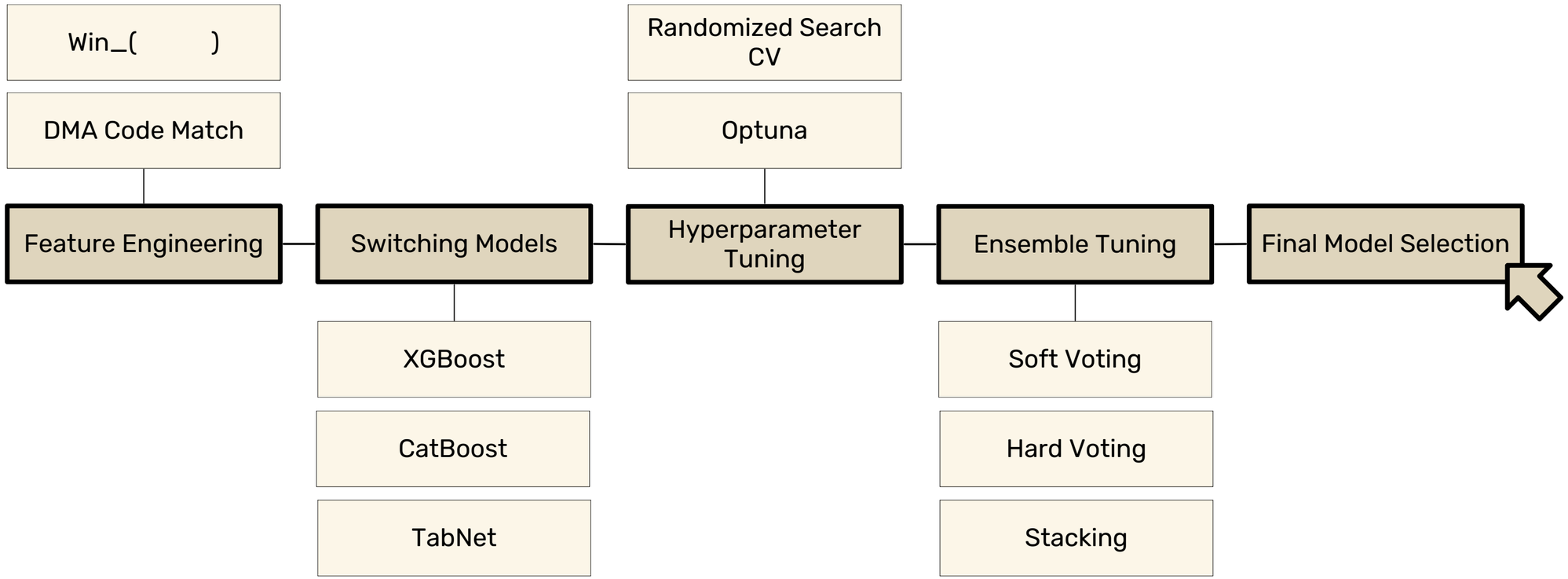
Parameters

N Iteration : 100 **Learning Rate : 0.05** **Max Depth : 6**

ADVANCED MODELING

Advanced Model

Approach Overview



Feature Engineering

Enhance Feature Impact

01 If Customer DMA Code is same with Institution DMA Code or not ?

02 Does Win_Percentage have interactions with other team performance features?

DMA Code Match

- DMA_East_Match
- DMA_West_Match
- DMA_South_Match
- DMA_Midwest_Match
- DMA_Total_Match

Win_( )

- Win_Tenure: Win_Percentage * NCAA_Tenure
- Win_Attendance: Win_Percentage * RegularSeasonAverageAttendance
- Win_Score: Win_Percentage * RegularSeasonAverageScore

Hyperparameter Tuning

RandomizedSearchCV

- Selects random **‘integer’** hyperparameter combinations from a predefined range (manual specification of parameter)
- **No adaptive tuning**: Does not learn from previous trials
- **Works well with smaller search spaces** but inefficient for large search spaces, using with 5 cross-validation

#integer #simple #crossvalidation

Optuna

- Select and explore random **‘float’** hyper-parameter from a predefined range
- **Adaptively explore** best hyperparameters: Learns **from previous trials** to narrow the search space efficiently
- **More efficient for complex** and high dimensional search spaces
- Define **optimization function** to find optimal

#float #complex #optimization

Best Model

XGBoost Ensemble Model

OPTUNA TUNIG

Performance Geographic Scale

HARD VOTING

Ensemble

Parameters Search Scope

N-Iteration : 500~1500 Learning Rate : 0.01~0.1 Max Depth : 4~10

Voting Weight

Performance : 2 Geographic : 3 Institution Scale : 1

Feature Importance

Top 20 Most Important Features in Best Ensemble

- The three aspects of features are **evenly distributed** in top 20 list
- In terms of dominance, feature importance follows the order:

1. **geographic**
2. **performance**
3. **institution scale**

Results

- **Accuracy (NationalChampion): 0.4624**
- **Accuracy (SemiFinal East/West): 0.6915**
- **Accuracy (SemiFinal South/Midwest): 0.6412**

ex) Optimal Parameters for National Champion Bracket Prediction

Team Performance XGBoost Model		
N-estimators: 1287	Learning Rate: 0.0342	Max Depth: 8
Geographical Factor XGBoost Model		
N-estimators: 1436	Learning Rate: 0.0388	Max Depth: 6
Institution Scale XGBoost Model		
N-estimators: 1079	Learning Rate: 0.0303	Max Depth: 6

Key Takeaways

Previous Assumption from Tableau Dashboard

- Larger schools boost engagement, leading to more bracket submissions.
- Team selection is primarily influenced by team performance.
- Affiliation remains a key factor in decision-making, especially in states with strong teams.

Final Conclusion

- Multiple model experiments revealed that the three aspects - geographical, performance, and scale - contribute in a well-balanced manner to the predictions.
- Addition to school affiliation, geographical affinity is also a key factor.

Thank You

Look Forward to Further Discussion!

Appendix

Code Chunk
Detailed Results

Parameter Optimization

```python
def objective(trial, X_train, y_train, X_val, y_val, start_time, total_trials):
    "XGBoost Hyperparameter Optimization Function with Optuna"
    params = {
        'objective': 'multi:softmax',
        'num_class': len(np.unique(y_train)),
        'learning_rate': trial.suggest_float("learning_rate", 0.01, 0.1),
        'max_depth': trial.suggest_int("max_depth", 4, 10),
        'n_estimators': trial.suggest_int("n_estimators", 500, 1500),
        'subsample': trial.suggest_float("subsample", 0.5, 1.0),
        'colsample_bytree': trial.suggest_float("colsample_bytree", 0.5, 1.0),
        'gamma': trial.suggest_float("gamma", 0.0, 0.5),
        'reg_lambda': trial.suggest_float("reg_lambda", 1.0, 10.0),
        'reg_alpha': trial.suggest_float("reg_alpha", 0.0, 5.0),
        'random_state': 42
    }
```

Optimal Parameter Values

Optimal Parameters			
Team Performance XGBoost Model			
National Champion	N-estimators: 1287	Learning Rate: 0.0342	Max Depth: 8
Semi Final E/W	N-estimators: 560	Learning Rate: 0.0238	Max Depth: 5
Semi Final S/M	N-estimators: 543	Learning Rate: 0.0111	Max Depth: 10
Geographical Factor XGBoost Model			
National Champion	N-estimators: 1436	Learning Rate: 0.0388	Max Depth: 6
Semi Final E/W	N-estimators: 792	Learning Rate: 0.0855	Max Depth: 7
Semi Final S/M	N-estimators: 1035	Learning Rate: 0.0749	Max Depth: 4
Institution Scale XGBoost Model			
National Champion	N-estimators: 1079	Learning Rate: 0.0303	Max Depth: 6
Semi Final E/W	N-estimators: 507	Learning Rate: 0.0213	Max Depth: 5
Semi Final S/M	N-estimators: 1417	Learning Rate: 0.0610	Max Depth: 4

Detailed Feature Importance