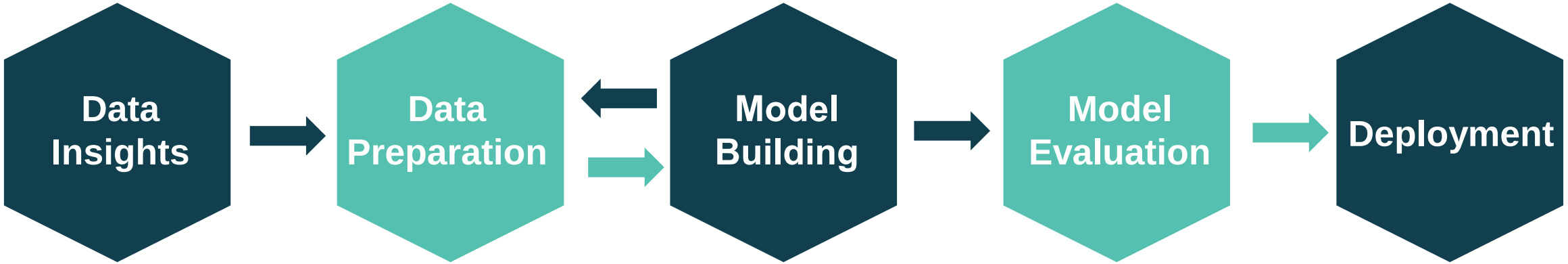
404FOUND

Hao Lu / Haohan Wang / Jiahao Wu / Hui Yuan

Project Introduction

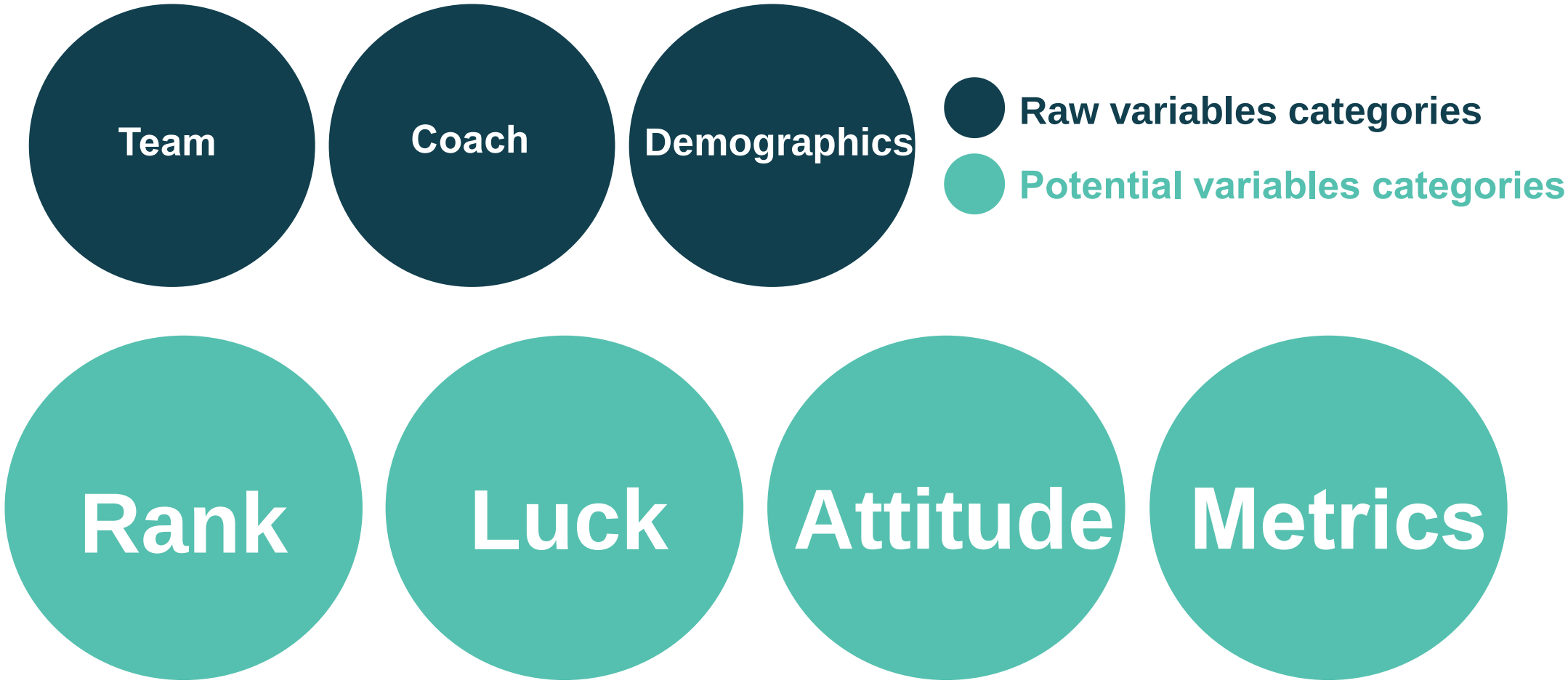
MARCH DATA CRUNCH MADNESS

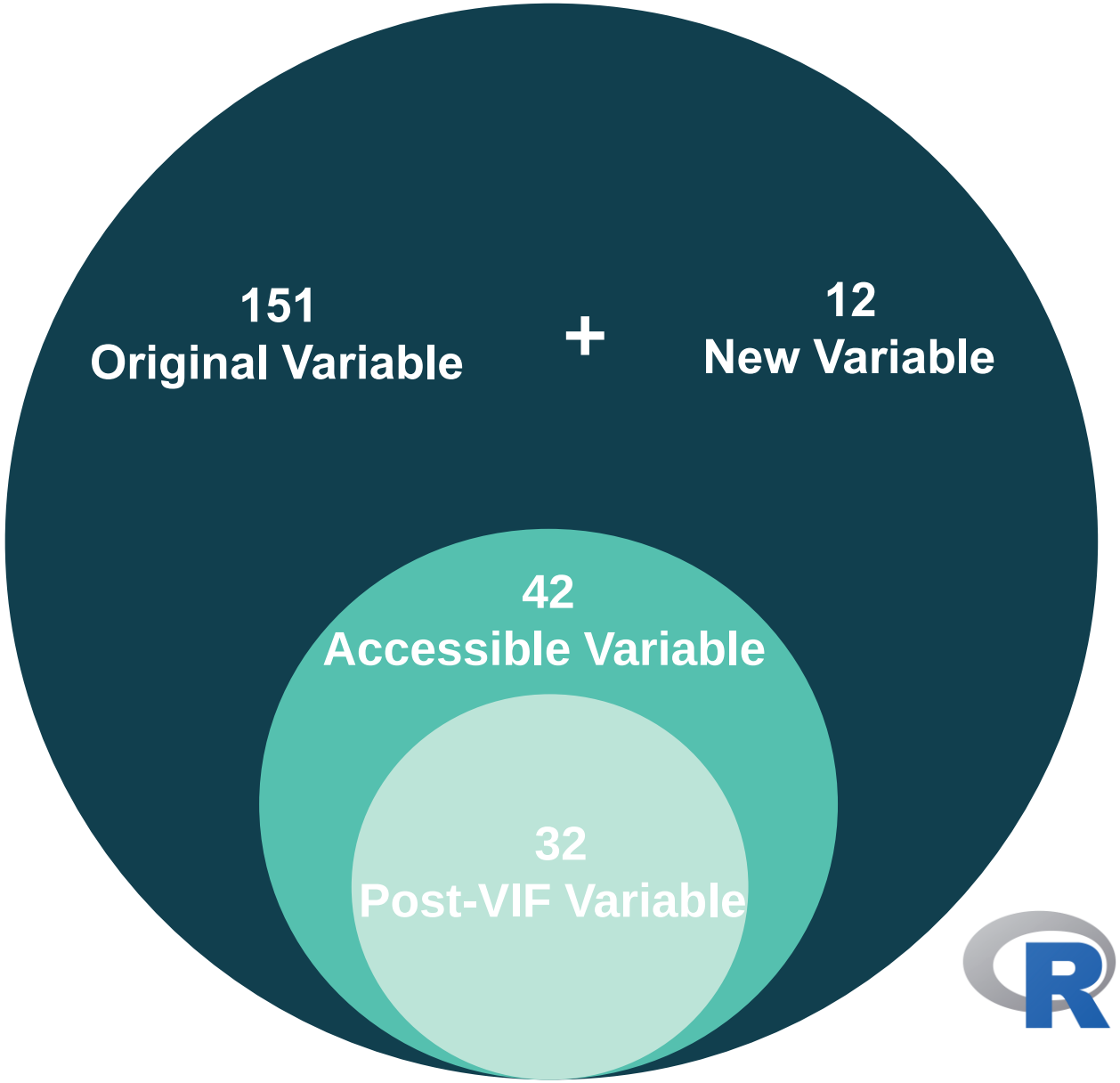
The goal of this project is to predict NCAA tournament bracket. In order to comprehensively build a model, we applied different types of features via famous NCAA websites such as Basket reference, Colley Rankings, Kenpom and selected them by using VIF method. With a 74.6% accurate model ensembled by 5 machine learning models, the probabilities and results of all possible matches are predicted, and the bracket is finalized. Features related to the team's ability play the most important role in predicting our model. For further improvement, applying some other relative features like the average arm length or height of the whole team.



Initial Thoughts

MARCH DATA CRUNCH MADNESS




151
Original Variable
+
12
New Variable
42
Accessible Variable
32
Post-VIF Variable

New Variables

MARCH DATA CRUNCH MADNESS

Simple Rating System
- ✓ Average Point Differential
- ✓ Strength of Schedule

W-L%
- ✓ Winning Percentage

Colley Rating
- ✓ Dr. Colley Ranking System

Strength of Schedule
- ✓ Opponent winning %
- ✓ Opp-opp winning %

Basketball Four Factors
- ✓ 40% Shooting (0.4*eFG%)
- ✓ 25% Turnovers (0.25*TOV%)
- ✓ 20% Rebounding (0.2*ORB%)
- ✓ 15% Free Throws (0.15*FT%)

Luck

Kenpom Basketball Analysis

Deloitte. NCAA®

NCAA March Data

Collinearity Detection

MARCH DATA CRUNCH MADNESS

Variance inflation factor (VIF)

- Measuring the ratio of variance in a statistical linear model
- Variance of an estimated regression coefficient is increased

Standard

- VIF values above 10 imply variables with strong collinearity

Conclusion

- Our team remove these features with strong collinearity

Feature Engineering

MARCH DATA CRUNCH MADNESS

3
- Providing a better linear feature
- Can reduce overall dimensions

2
- Binary classification requires linear features
- Increasing the accuracy and modeling speed

1
- Samples are in game scale & target is probability of team1 win
- Variables should consider the result of both team1 and team2
- **For example:** diff_fg2pct = [team1_fg2pct] — [team2_fg2pct]

Feature Selection

MARCH DATA CRUNCH MADNESS

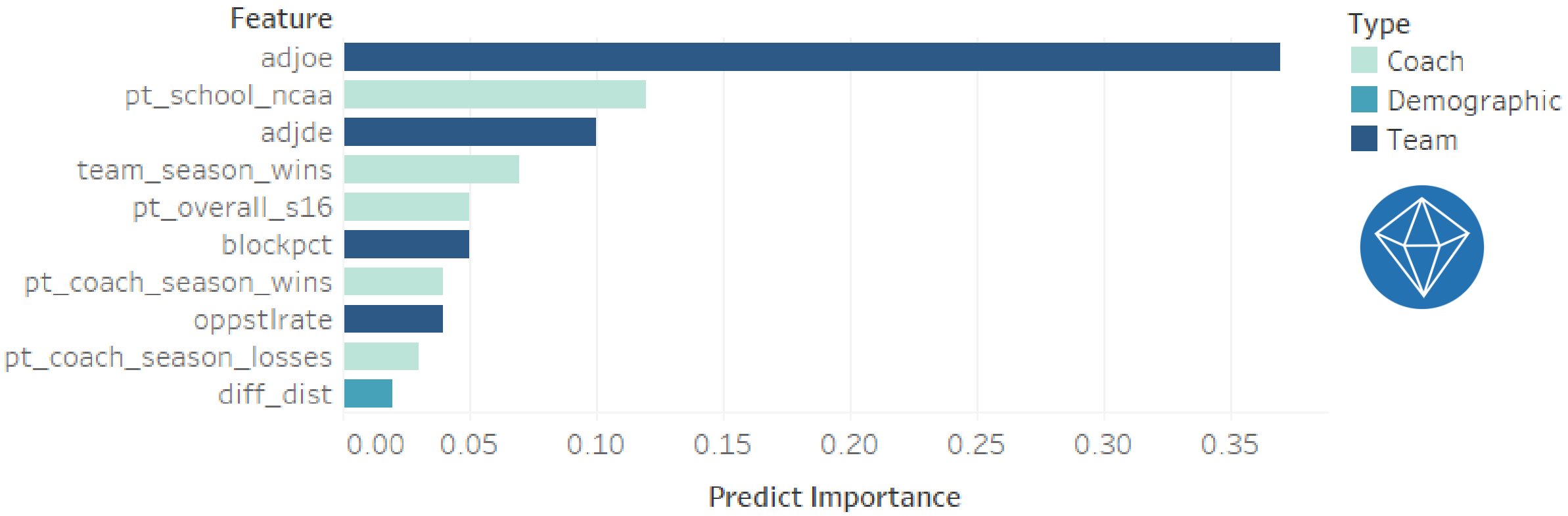


Figure 1. Variables importance – Decision Tree

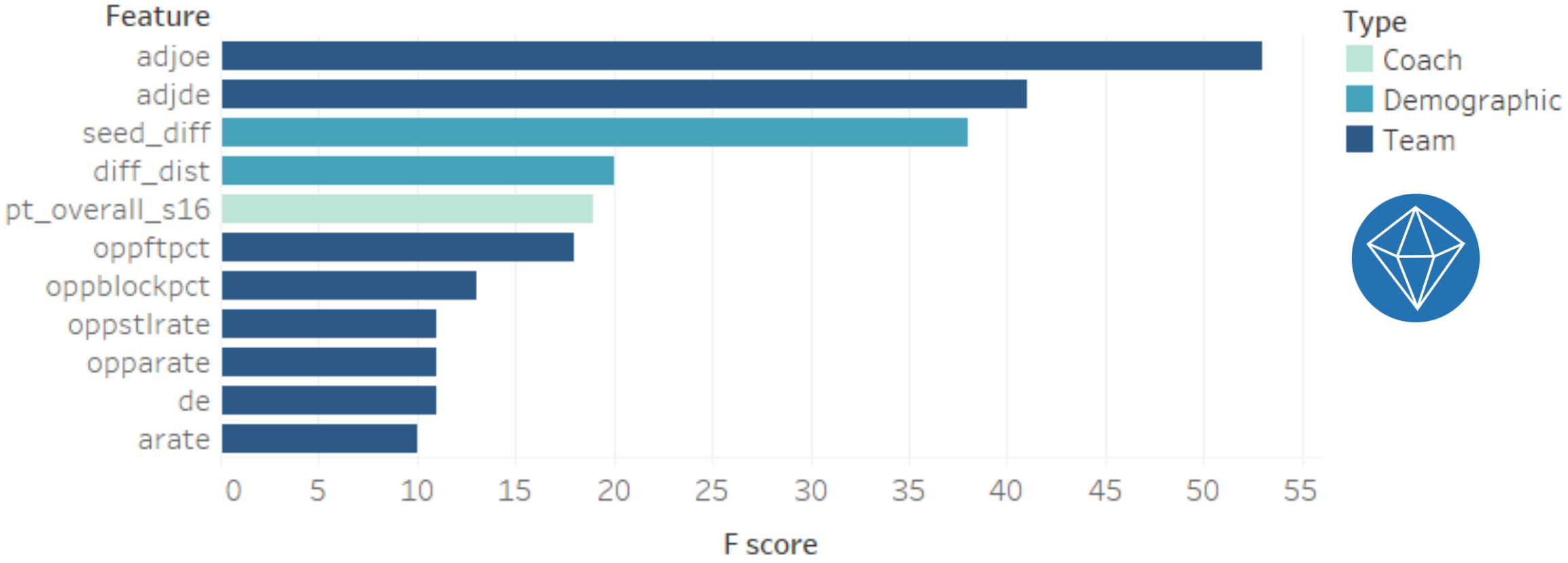


Figure 2. Variables importance - XGBoost

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Crunch Madness

Model Comparisons

MARCH DATA CRUNCH MADNESS

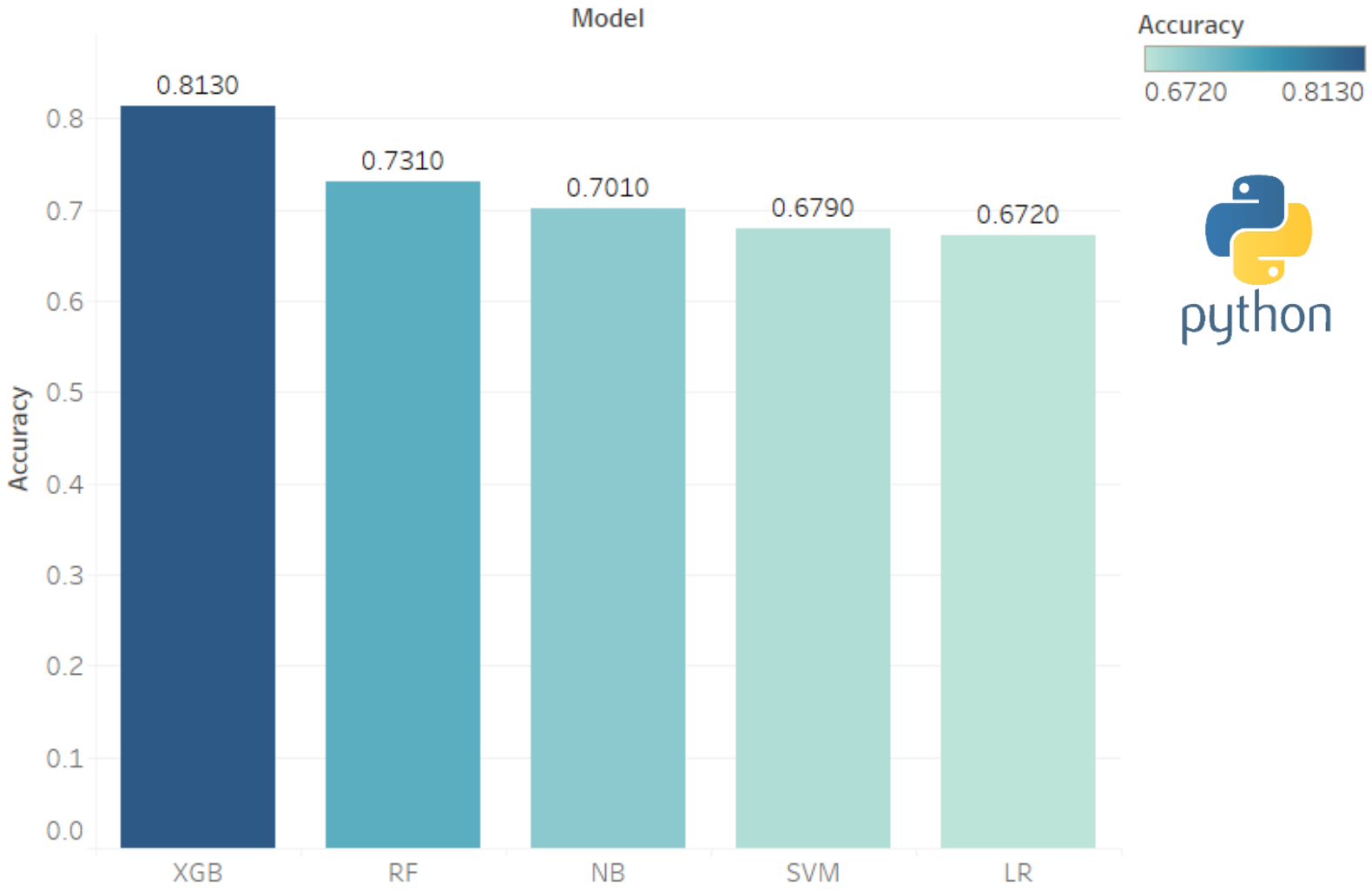


Figure 3. Final model accuracies

Accuracy by Round

MARCH DATA CRUNCH MADNESS

- **For the 1st and 2nd round:**
  - ✓ Games between No.8 seed and No.9 seed are the most difficult to predict.
  - ✓ For all top 3 teams in terms of seed ranking, they defeat opponent teams.

- **For the 4th round:**
  - ✓ High risk to predict if analyst only considers the team's ability and seed.
  - ✓ Must consider other factors (pressure, injury and championship mentality)

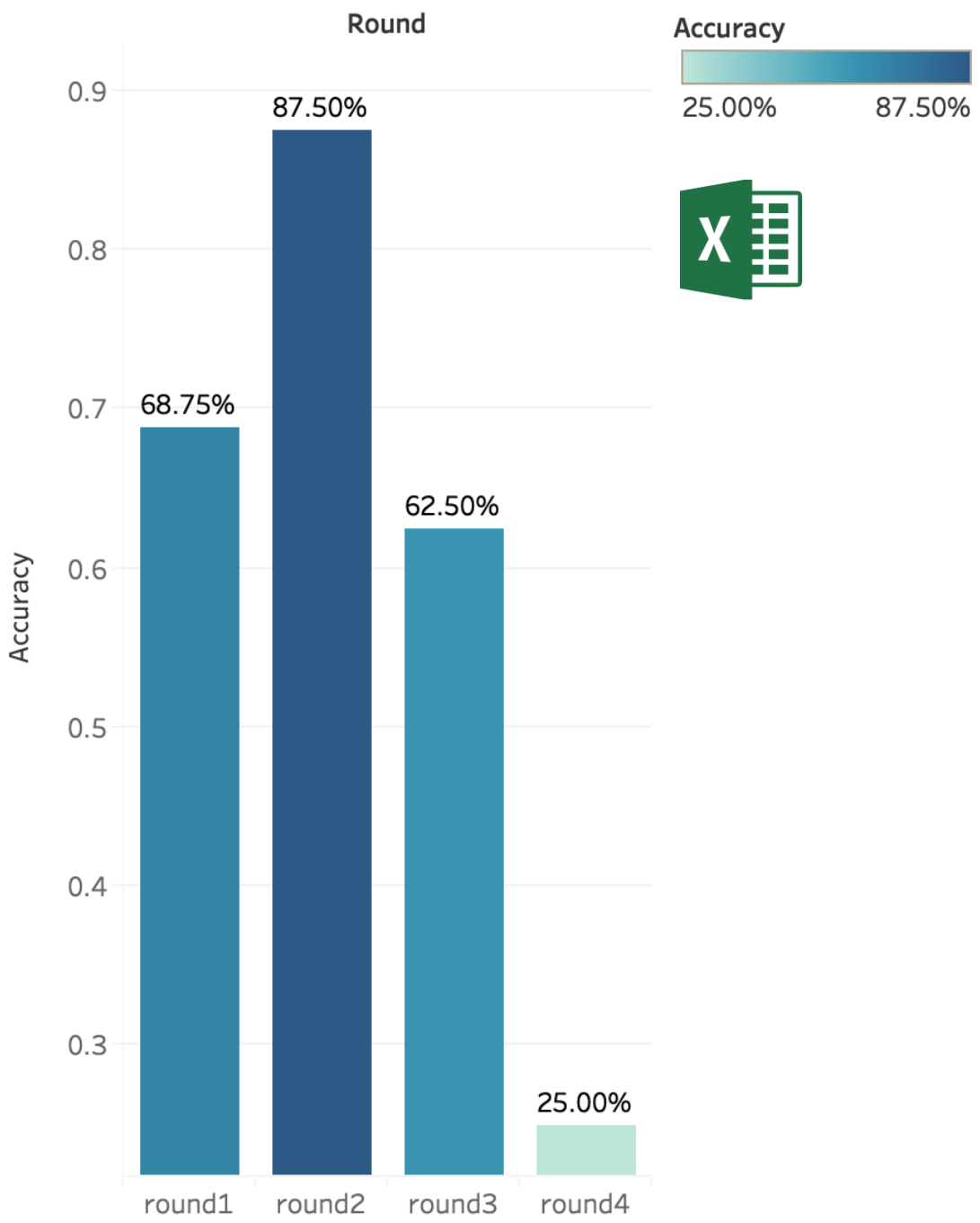


Figure 4. Accuracy by Round

404FOUND

Hao Lu / Haohan Wang / Jiahao Wu / Hui Yuan

Project Conclusion

MARCH DATA CRUNCH MADNESS

- ✓ Among five predictive models, XGBoost has the highest predictor accuracy, followed by Random Forest and Naïve Bayes.
- ✓ All these three models have more accurate prediction than Support Vector Machine and Logistic Regression, which means classification methods are better than regression-based classification methods.
- ✓ That is to say, in our case, classification models can better predict outcome of sports game than regression-based classification models do.

However, since this is a binary classification problem, we shouldn't only focus on accuracy. They are many other evaluation indexes we need to consider, such as logloss and ROC.

Therefore, we run an ensemble to reduce the error rate by collecting all conducted independent hypotheses and combine their predictions.

Project Application

MARCH DATA CRUNCH MADNESS

For coach:

- ✓Adjust coaching strategy respectively based on the predicted game result and probability of win
- ✓Bring extra attentions to the top five variables from the variable importance predictors of win

For managerial level:

- ✓Execute a reward & punishment mechanism according to team's expected position in tournament

Project Evaluation

MARCH DATA CRUNCH MADNESS

Model optimization:

- ✓ Conducting **Gradient Descent Algorithm** to optimize our models.
- ✓ Models typically has a cost function evaluate particular set of parameters.
- ✓ It can help us to find weights for parameters, minimizing the cost function.

New features:

- ✓ Adding more variables to make our dataset even more comprehensive.
- ✓ Focusing on variables related to players' physical fitness and experience
- ✓ E.g. average height of players, average arm length and average years pro.


1 DUKE
16 NCCU/NDSU
DUKE
8 VCU
9 UCF
VCU
DUKE
5 MISSISSIPPI ST
12 LIBERTY
MISSISSIPPI ST
4 VIRGINIA TECH
13 SAINT LOUIS
VIRGINIA TECH
VIRGINIA TECH
EAST
6 MARYLAND
11 BELMONT/TEMPLE
MARYLAND
3 LSU
14 YALE
LSU
LSU
7 LOUISVILLE
10 MINNESOTA
LOUISVILLE
2 MICHIGAN ST
15 BRADLEY
MICHIGAN ST
MICH. ST.
MICH. ST.
1 GONZAGA
16 FDU/PV A&M
MICHIGAN ST
8 SYRACUSE
9 BAYLOR
SYRACUSE
GONZAGA
GONZAGA
5 MARQUETTE
12 MURRAY ST
MARQUETTE
4 FLORIDA ST
13 VERMONT
FLORIDA ST
FLORIDA ST
WEST
6 BUFFALO
11 ASU/ST JOHN'S
BUFFALO
3 TEXAS TECH
14 N. KENTUCKY
TEXAS TECH
TEXAS TECH
7 NEVADA
10 FLORIDA
NEVADA
2 MICHIGAN
15 MONTANA
MICHIGAN
MICHIGAN
NATIONAL CHAMPIONSHIP
March 19-20
FIRST FOUR
Dayton
16 NC CENTRAL
16 N DAKOTA ST
11 BELMONT
11 TEMPLE
16 FDU
16 PV A&M
11 ARIZONA ST
11 ST. JOHN'S
VIRGINIA
VIRGINIA
VIRGINIA
1 VIRGINIA
16 GARDNER-WEBB
OKLAHOMA
8 OLE MISS
9 OKLAHOMA
WISCONSIN
WISCONSIN
5 WISCONSIN
12 OREGON
KANSAS ST
4 KANSAS ST
13 UC IRVINE
SOUTH
PURDUE
VILLANOVA
6 VILLANOVA
11 SAINT MARY'S
PURDUE
3 PURDUE
14 OLD DOMINION
TENNESSEE
TENNESSEE
7 CINCINNATI
10 IOWA
TENNESSEE
2 TENNESSEE
15 COLGATE
N. CAROLINA
N. CAROLINA
1 NORTH CAROLINA
16 IONA
UTAH ST
8 UTAH ST
9 WASHINGTON
KANSAS
AUBURN
5 AUBURN
12 NEW MEXICO ST
KANSAS
4 KANSAS
13 NORTHEASTERN
MIDWEST
HOUSTON
IOWA ST
5 IOWA ST
11 OHIO ST
HOUSTON
3 HOUSTON
14 GEORGIA ST
KENTUCKY
KENTUCKY
7 WOFFORD
10 SETON HALL
KENTUCKY
2 KENTUCKY
15 ABILENE CHRISTIAN