March Data Crunch Madness 2022

Fordham Sports Analytics Society Team 1

Presented By: Adrian Crisostomo, Paul Gomes, Peter Majors, and Matthew Reese

Part 1: First Steps

First Steps - Provided Data Sources

- Examined Provided Data Frame and Sample Code (2002 - 2021 MM Games)
- Which Fields Did We Think Were The Most Important?
  - Basic Percentage-Based Statistics
  - Offensive and Defensive Efficiencies
  - Coaching /Team History (Regular Season & March Madness)

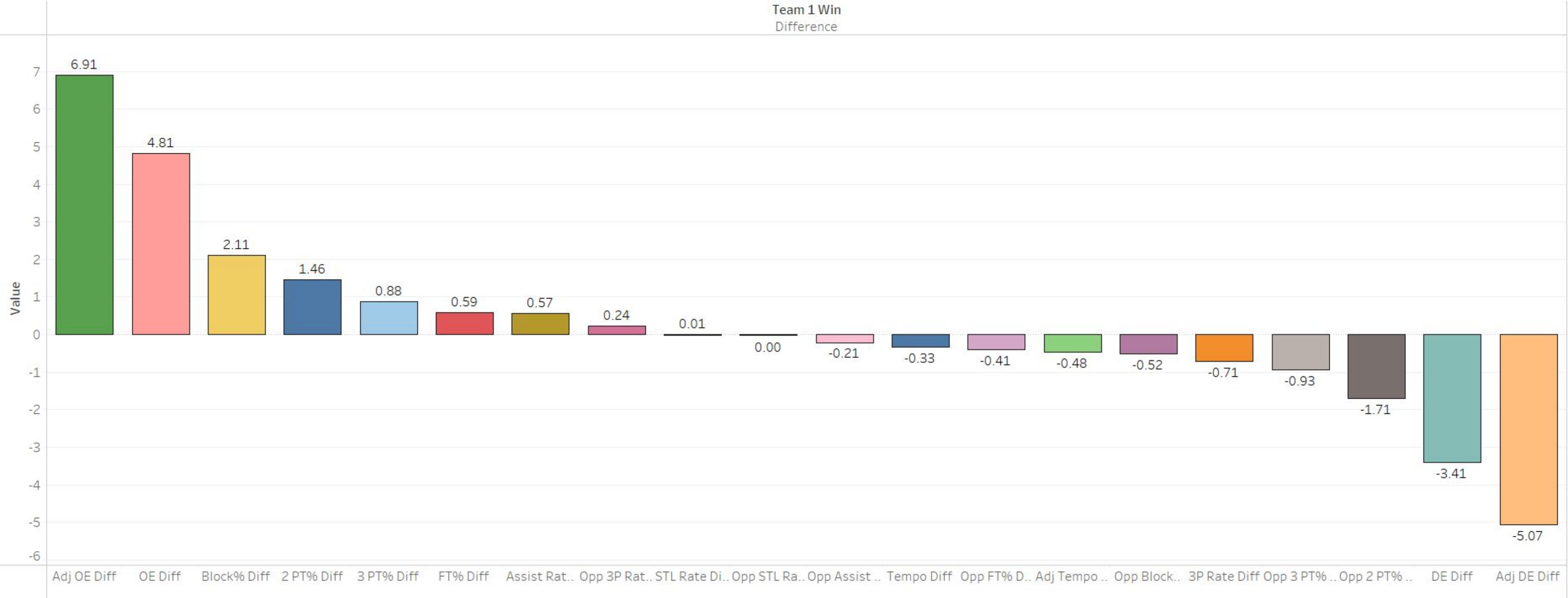
First Steps - Provided Data Sources

- Sample Code:
  - **Pythagorean Win%** = $(\text{adjoe})^{11.5}$ / ($(\text{adjoe})^{11.5}$ + $(\text{adjoe})^{11.5}$)
    - Probability Of Team Winning Based On Quality Of Play
  - **team1_log5** = (pythag_team1 * (pythag_team1 * pythag_team2)) / (pythag_team1 + pythag_team2 - (2 * pythag_team1 * pythag_team2))
    - Probability of Team Winning Based On Competing Pythags

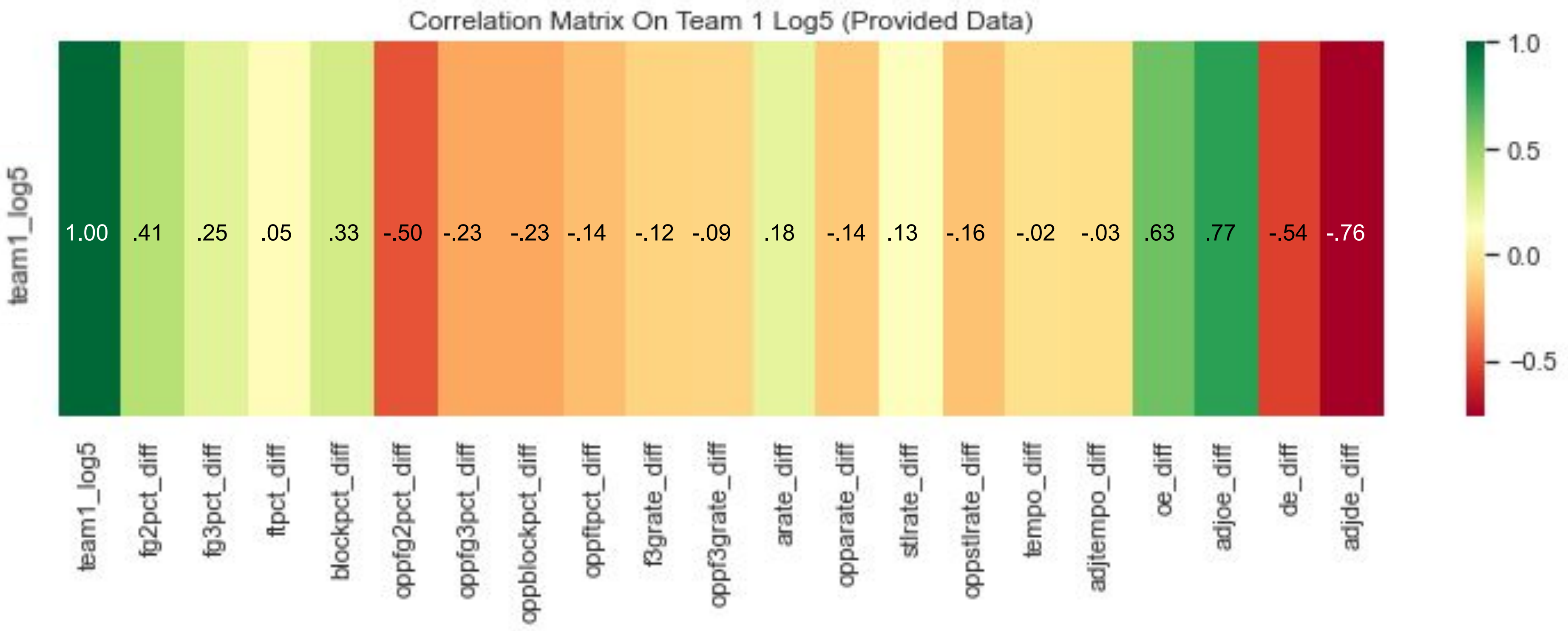
Part 2: Exploratory Analysis

Exploratory Analysis - Mean Differences

Average Difference In Season Statistics Between Winning and Losing Teams (March Madness 2002 - 21)

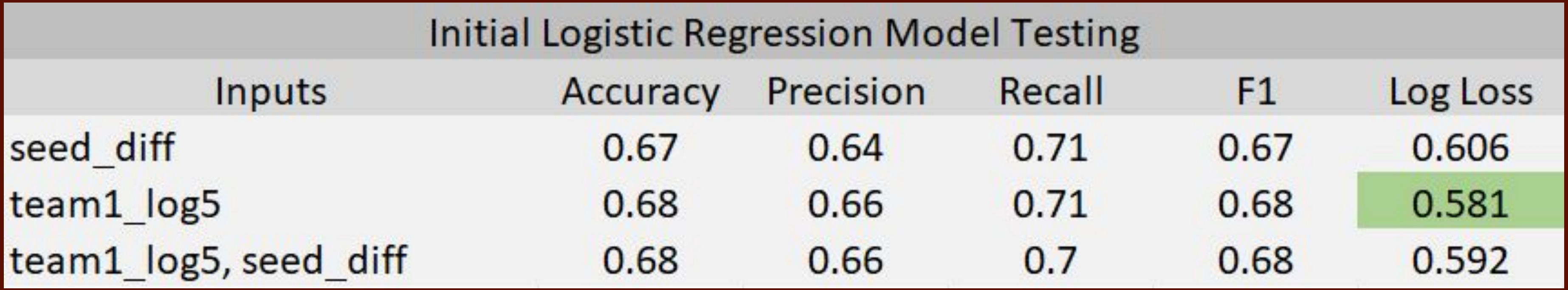


Exploratory Analysis - Correlation Matrix



Correlation Matrix On Team 1 Log5 (Provided Data)

Exploratory Analysis - Initial Tests



Initial Logistic Regression Model Testing

Inputs	Accuracy	Precision	Recall	F1	Log Loss
seed_diff	0.67	0.64	0.71	0.67	0.606
team1_log5	0.68	0.66	0.71	0.68	0.581
team1_log5, seed_diff	0.68	0.66	0.7	0.68	0.592

Team 1 Wins Seed Difference: **3.67**

Team 1 Loses Seed Difference: **-3.42**

Team1 Log5 & Seed Difference Correlation: **-.91**

Part 3: Applying Our Basketball Knowledge!

Basketball Knowledge - External Data

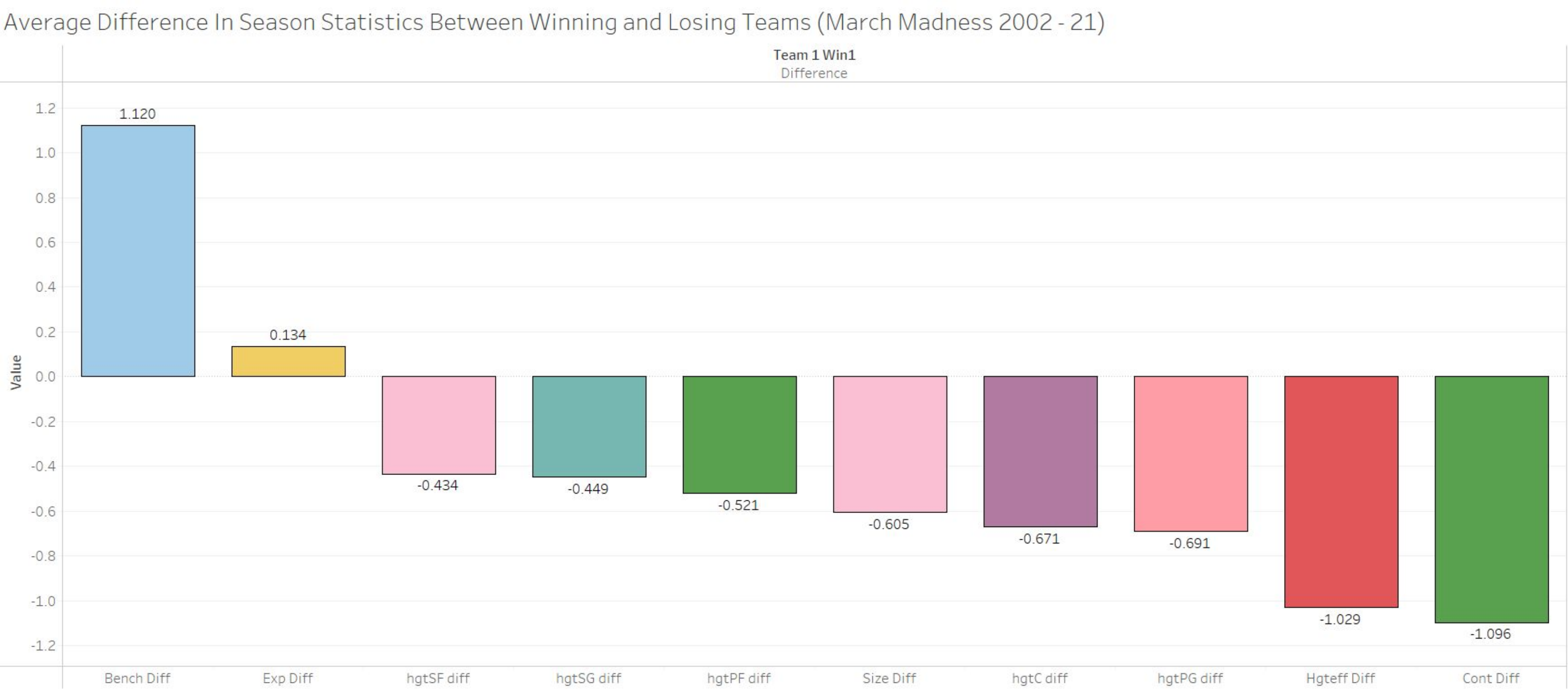
- Utilized kenpom.com, College Basketball's Foremost Data Analytics Resource
- Pulled Data From 2007 to 2021 (Excl. Covid-Cancelled 2020)
    - Reduced March Madness Games From 1246 to 916 (24.6% Decrease)

kenpom.com

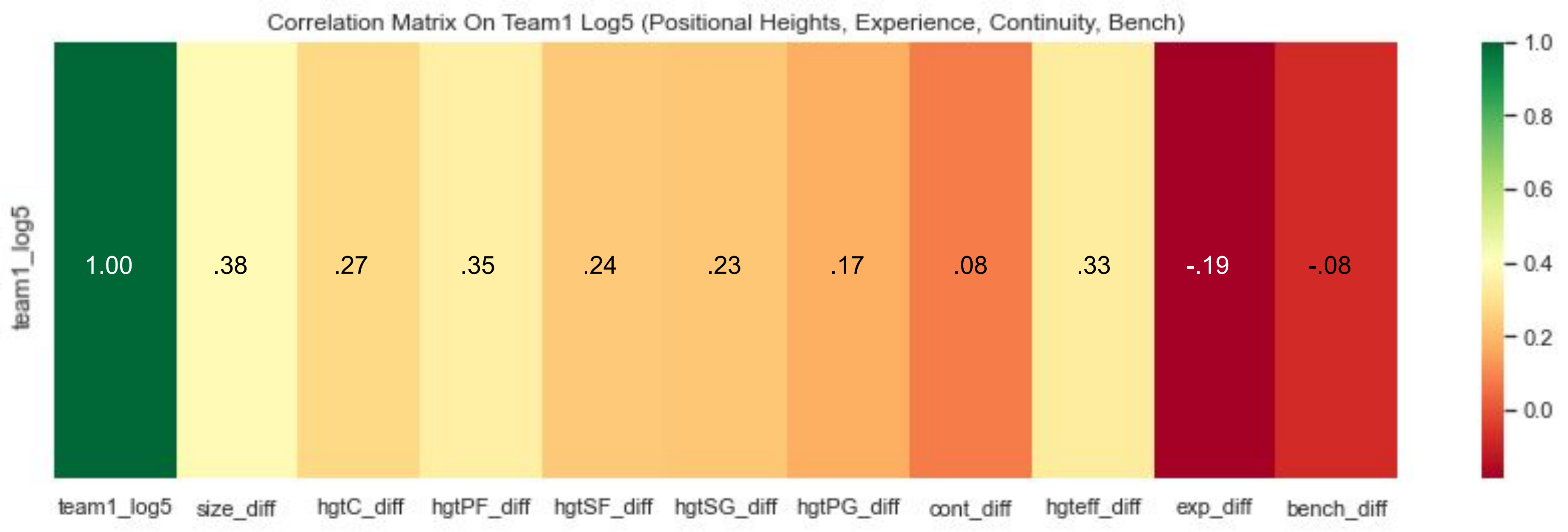
Basketball Knowledge - External Data

- Team Continuity
- Team Experience
- Strength of Bench
- Offensive Rating by Position
- Defensive Rating by Position
- Points Per Game by Position
- Height by Position
- Size by Position
- Home Court Advantage Rating
- Points Favored at Home Court
- Elevation of Home Court
- Other Home Court Metrics …

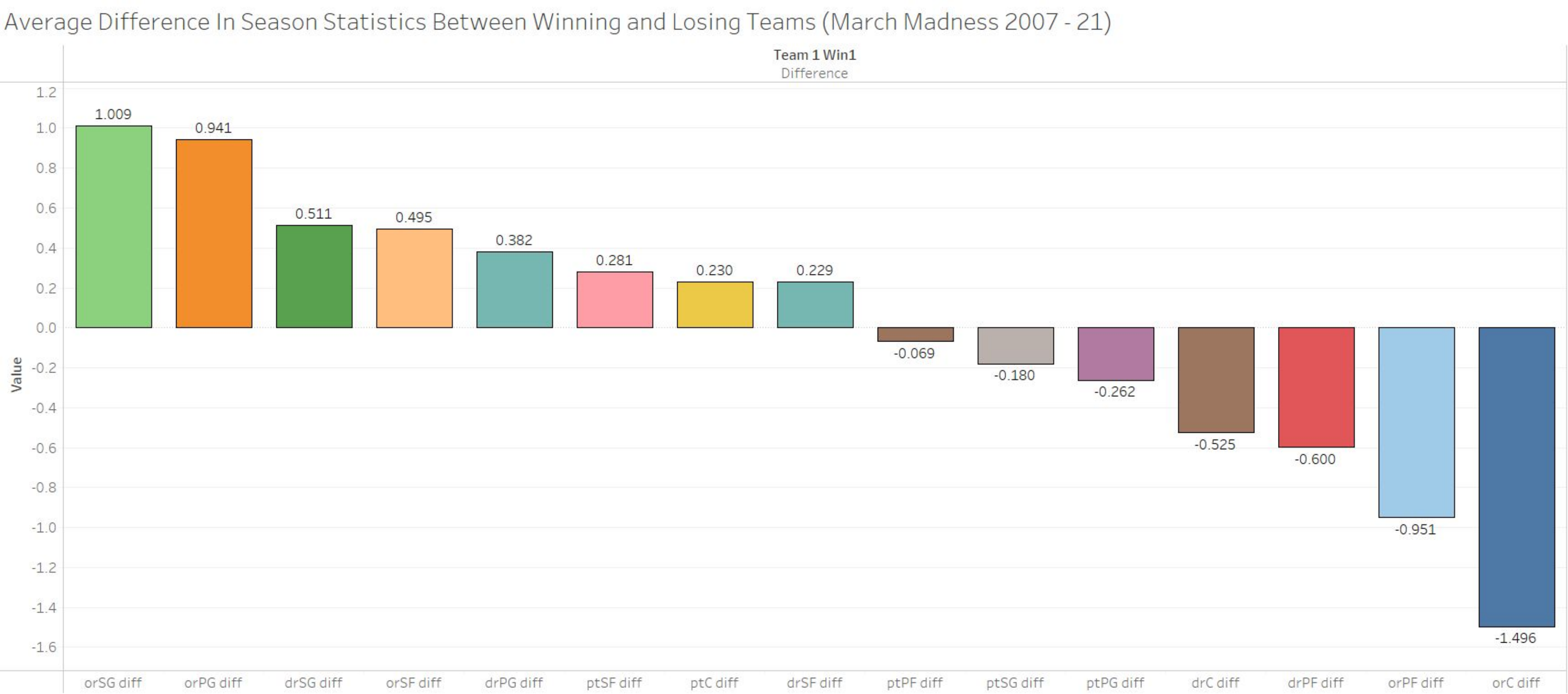
Basketball Knowledge - Position Size & Exp



Basketball Knowledge - Position Size & Exp

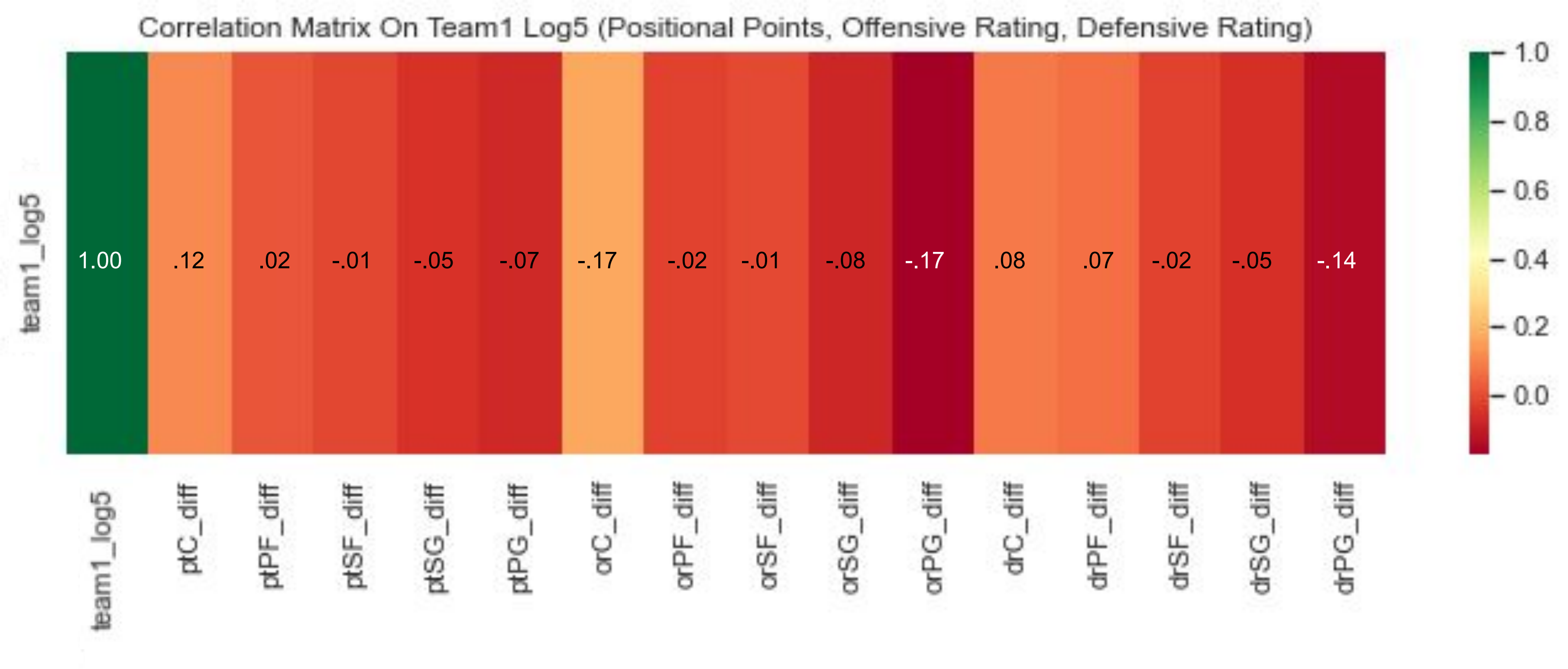


Basketball Knowledge - Position Skill



Average Difference In Season Statistics Between Winning and Losing Teams (March Madness 2007 - 21)

Basketball Knowledge - Position Skill

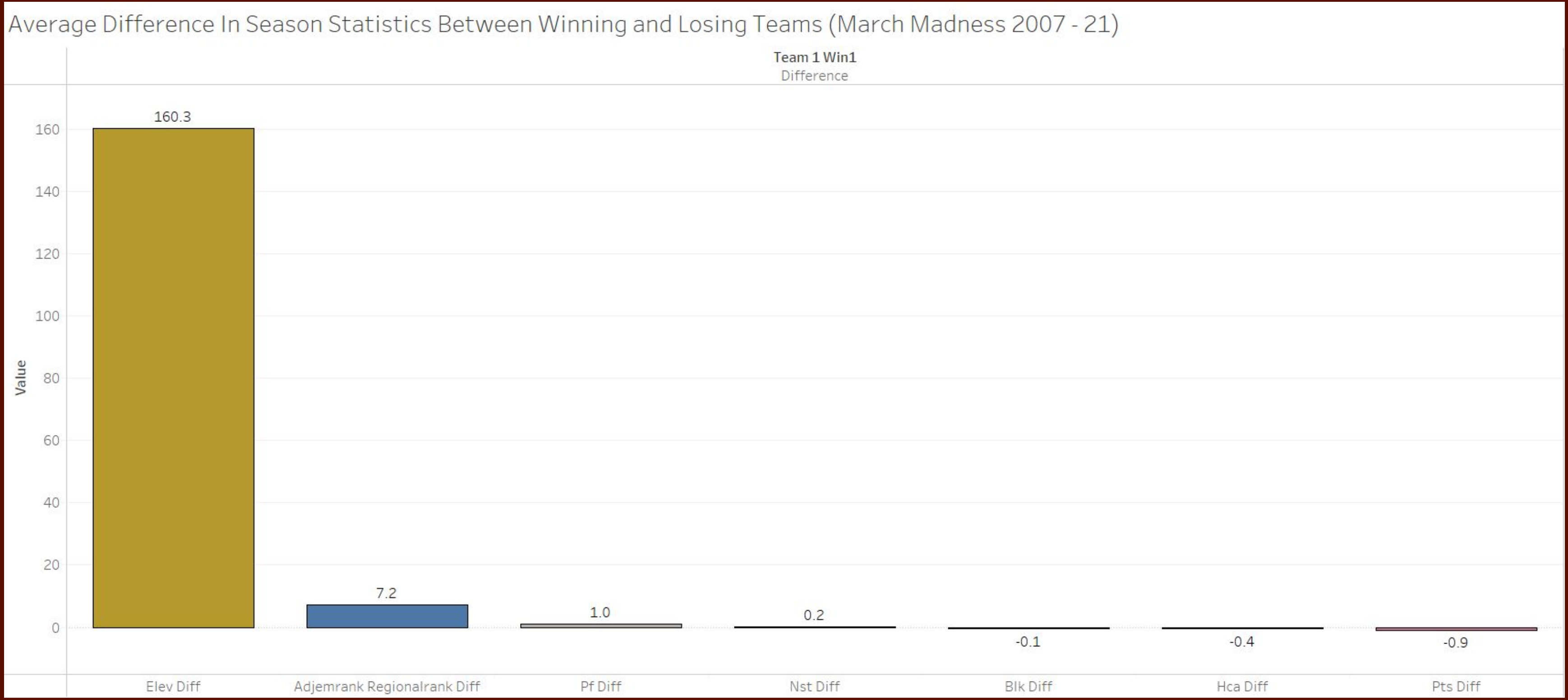


Correlation Matrix On Team1 Log5 (Positional Points, Offensive Rating, Defensive Rating)

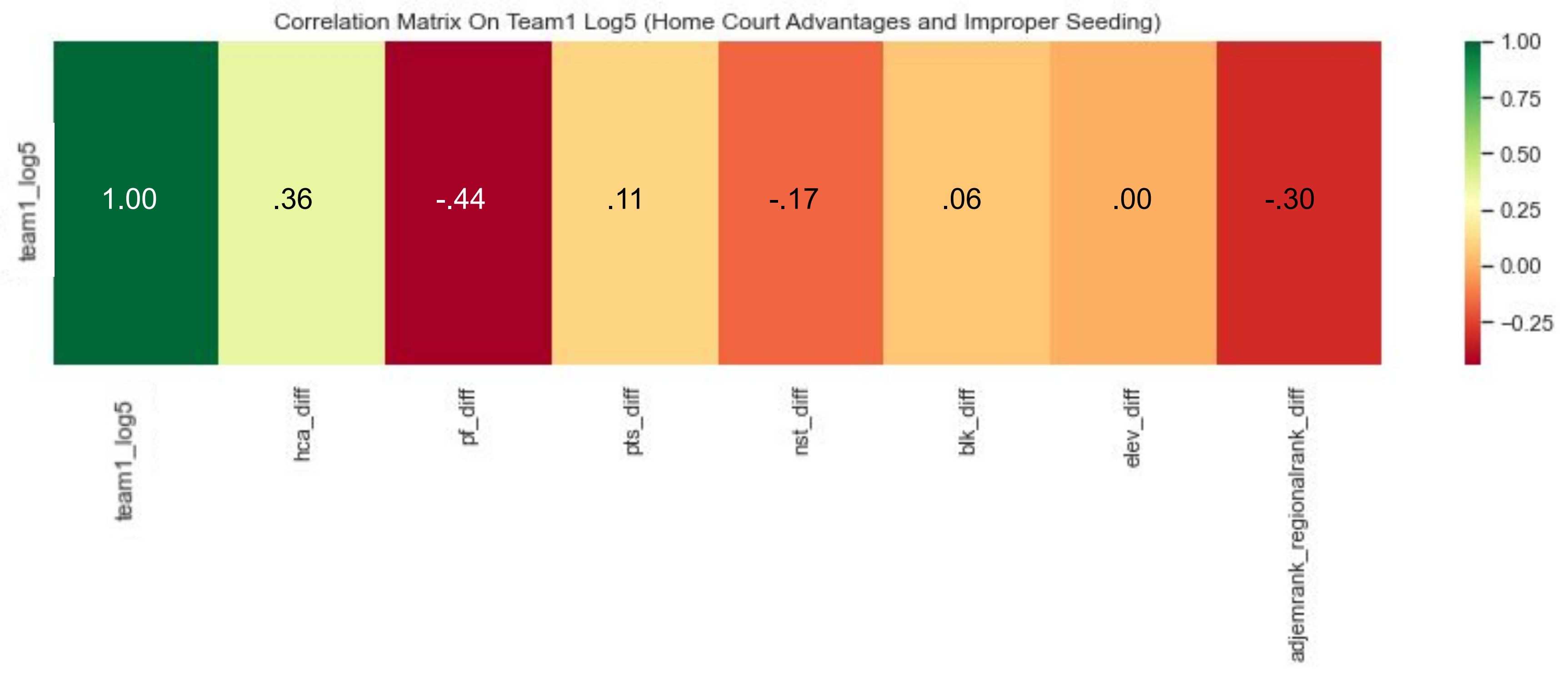
Basketball Knowledge - Improper Seeding

- Seeding Is A Contentious Part of March Madness
  - Can We Assign A Value To The “Over/Underratedness” of a Team Based On Their Seed and a Singular Performance Metric?
- Seed Region Rank *minus* Adj EM Region Rank (Each Team)
  - Higher = Over Inflated Seeding, Lower = Under Inflated Seeding
- Found Differences Between Team 1 and Team 2 (Each Matchup)
  - Higher = Winning Team Seed Inflated, Lower = Winning Team Seed Deflated
  - 'adjemrank_regionalrank_diff'

Basketball Knowledge - Seeding & HCA



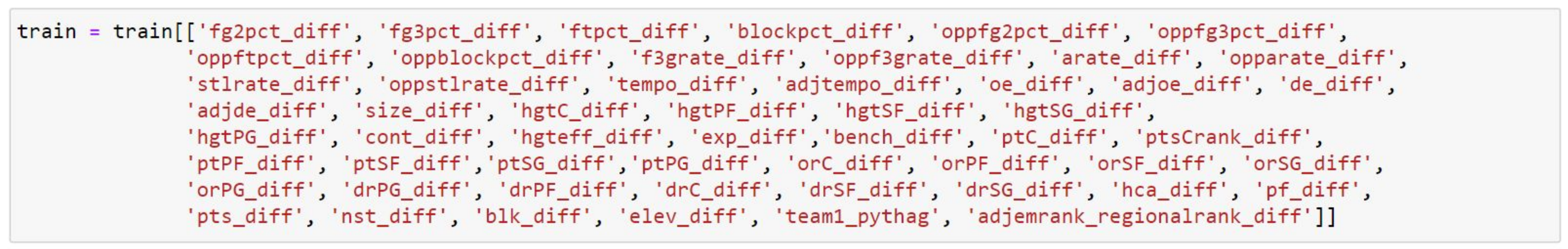
Exploratory Analysis - Correlation Matrix



Part 4: XGBoost

XGBoost - Data Preparation

- Utilized **54** Features
  - Focussed On Team Skill Differences, Intangibles, Uncontrollable Factors
- Test - Train Split: **70/30**
- Training Rows: **641** / Testing Rows: **275**

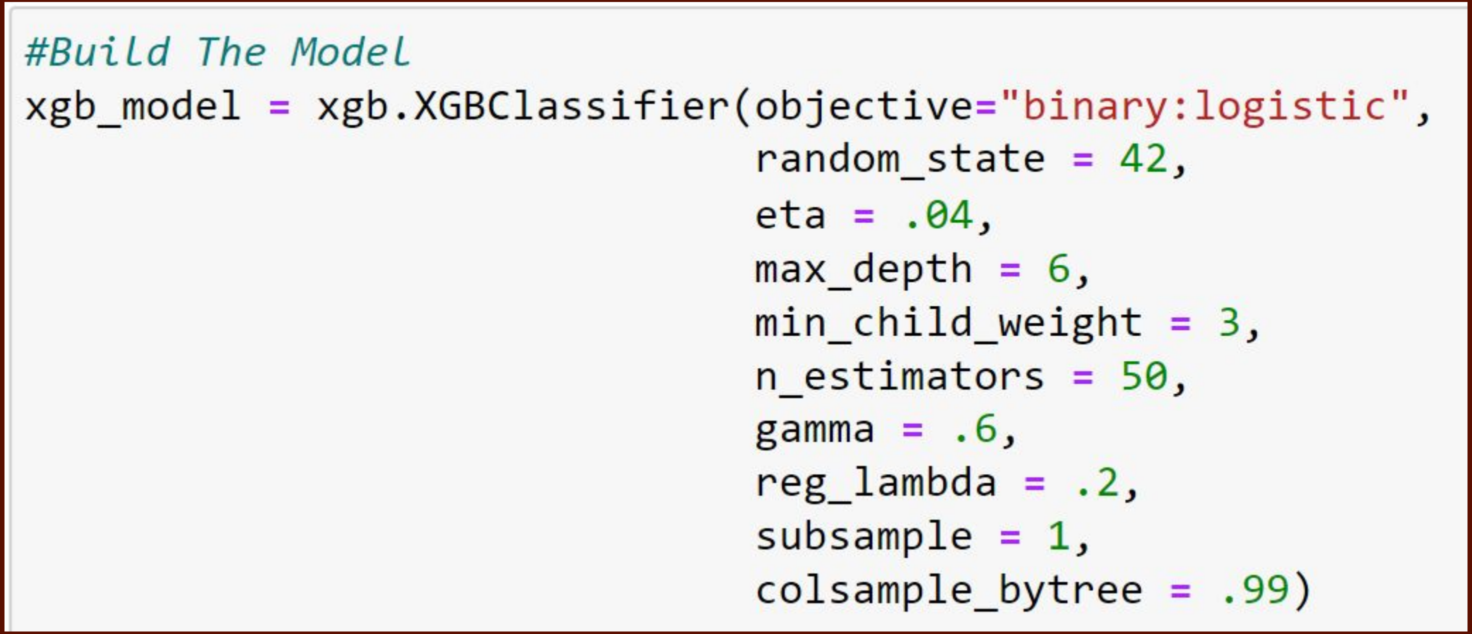


```
train = train[['fg2pct_diff', 'fg3pct_diff', 'ftpct_diff', 'blockpct_diff', 'oppfg2pct_diff', 'oppfg3pct_diff',
               'oppftpct_diff', 'oppblockpct_diff', 'f3grate_diff', 'oppf3grate_diff', 'arate_diff', 'opparate_diff',
               'stlrate_diff', 'oppstlrate_diff', 'tempo_diff', 'adjtempo_diff', 'oe_diff', 'adjoe_diff', 'de_diff',
               'adjde_diff', 'size_diff', 'hgtC_diff', 'hgtPF_diff', 'hgtSF_diff', 'hgtSG_diff',
               'hgtPG_diff', 'cont_diff', 'hgteff_diff', 'exp_diff','bench_diff', 'ptC_diff', 'ptsCrank_diff',
               'ptPF_diff', 'ptSF_diff','ptSG_diff','ptPG_diff', 'orC_diff', 'orPF_diff', 'orSF_diff', 'orSG_diff',
               'orPG_diff', 'drPG_diff', 'drPF_diff', 'drC_diff', 'drSF_diff', 'drSG_diff', 'hca_diff', 'pf_diff',
               'pts_diff', 'nst_diff', 'blk_diff', 'elev_diff', 'team1_pythag', 'adjemrank_regionalrank_diff']]
```

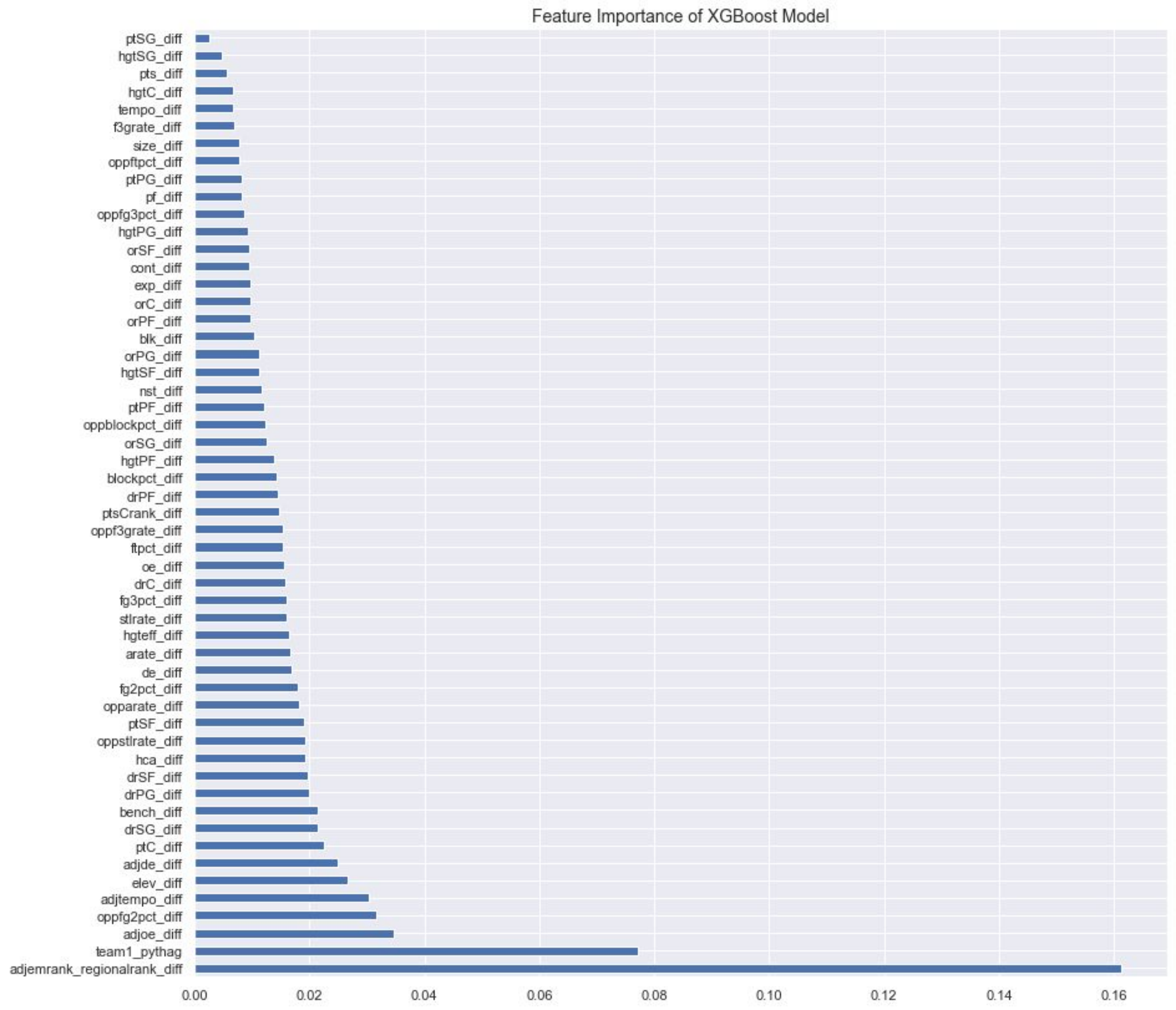
XGBoost - Summary

- Fit An XGBoost Model To Determine Which Features Mattered The Most
- Assigned A “Binary: Logistic” Objective and Evaluated On “Log Loss”
- Utilized GridSearchCV For Hyperparameter Tuning
    - Number of Jobs = **4**
    - Cross Validation = **3**
    - Early Stopping = **5**

XGBoost - Tuned Hyperparameters



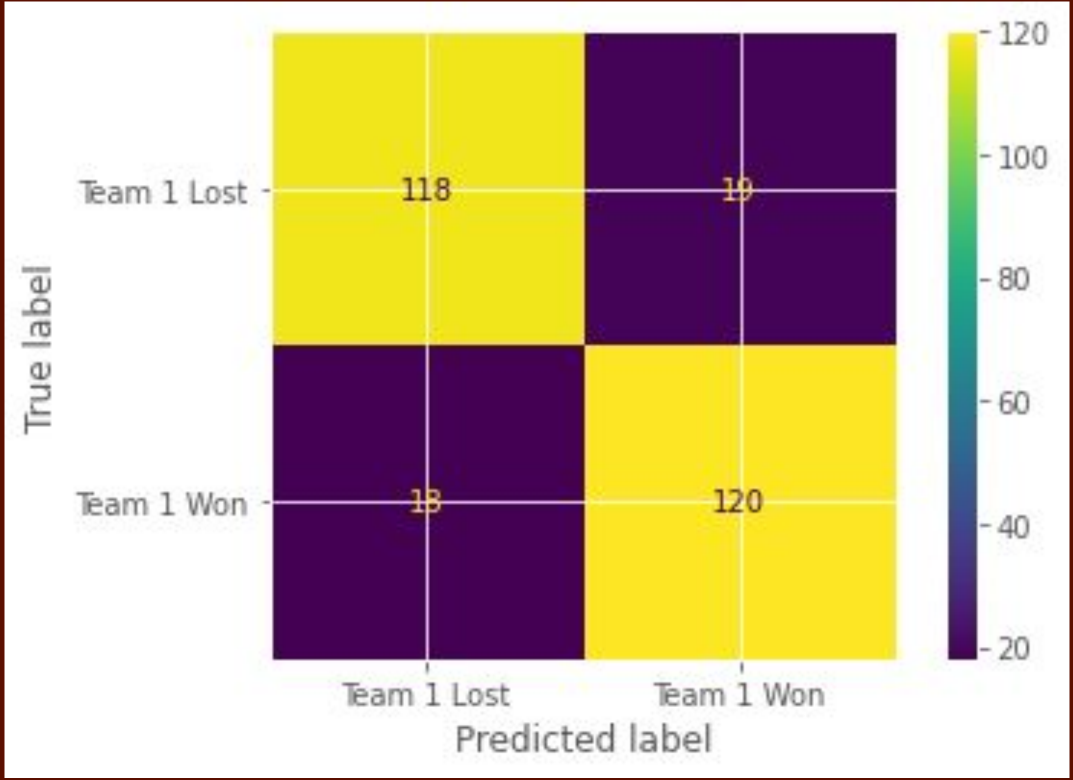
```
#Build The Model
xgb_model = xgb.XGBClassifier(objective="binary:logistic",
                              random_state = 42,
                              eta = .04,
                              max_depth = 6,
                              min_child_weight = 3,
                              n_estimators = 50,
                              gamma = .6,
                              reg_lambda = .2,
                              subsample = 1,
                              colsample_bytree = .99)
```


Feature Importance of XGBoost Model

XGBoost - Results On Test Data

- Log Loss: **.317**
- Accuracy: **84%**
- Precision: **84%**
- Recall: **86%**
- F1 Score: **85%**

XGBoost - Confusion Matrix



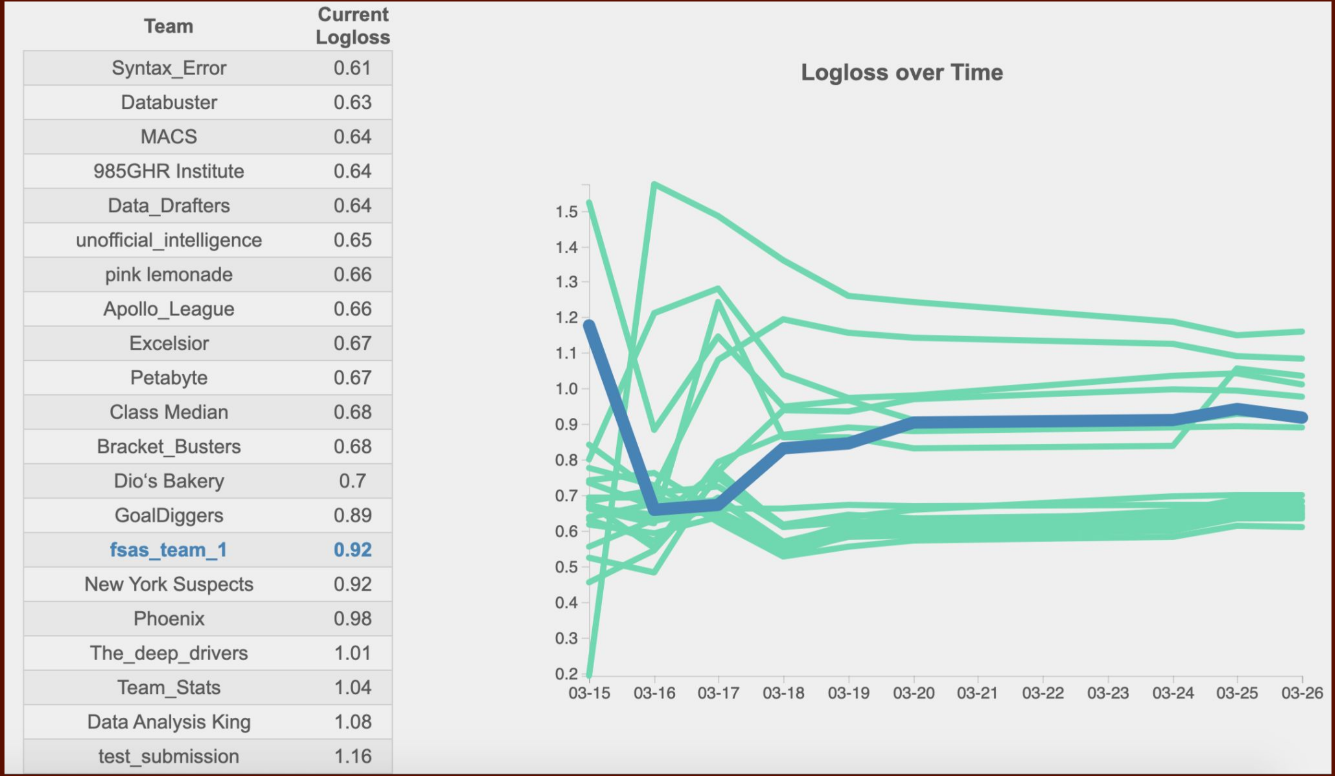
Part 5: Reflections

Sports Analytics Society

Reflections - General Analysis

- 1 Seeds Greatly Undervalued
- Highly Confident In Many Predictions, Not Many ~ 50%
  - Log Loss Hurt By Decisiveness
- Performance Drastically Improves As Games Become More Tightly Contested
- Strength of Schedule Seems To Be Lacking in Model

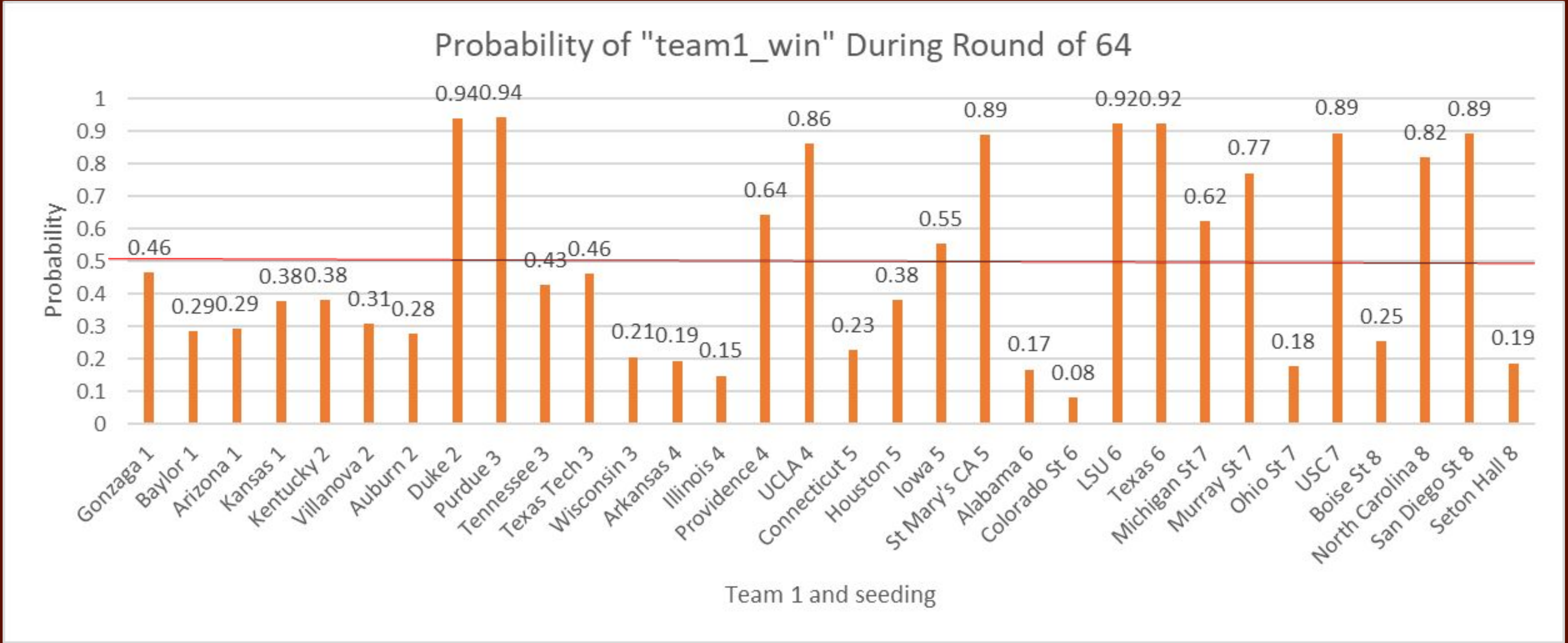
Reflections - Current Performance



Team	Current Logloss
Syntax_Error	0.61
Databuster	0.63
MACS	0.64
985GHR Institute	0.64
Data_Drafters	0.64
unofficial_intelligence	0.65
pink lemonade	0.66
Apollo_League	0.66
Excelsior	0.67
Petabyte	0.67
Class Median	0.68
Bracket_Busters	0.68
Dio's Bakery	0.7
GoalDiggers	0.89
fsas_team_1	**0.92**
New York Suspects	0.92
Phoenix	0.98
The_deep_drivers	1.01
Team_Stats	1.04
Data Analysis King	1.08
test_submission	1.16

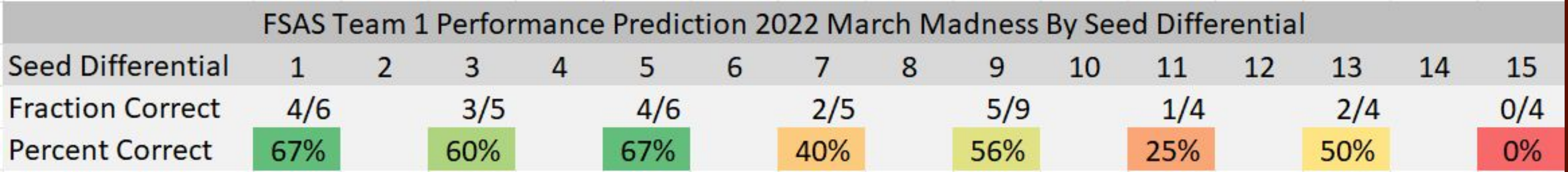
Logloss over Time

Reflections - Higher Seed Winning Prob



Reflections - Performance by Seed Diff

- Better Performance With Tighter Seed Differentials (First 43 Games)
  - Performs Best When Human Intuition is More Uncertain
  - Obvious Pitfalls When Games Seem More Certain



FSAS Team 1 Performance Prediction 2022 March Madness By Seed Differential

Seed Differential	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Fraction Correct	4/6		3/5		4/6		2/5		5/9		1/4		2/4		0/4
Percent Correct	67%		60%		67%		40%		56%		25%		50%		0%

Potential Changes For Future Years

- Scale Metrics from Winners & Losers Using Ratios, Not Differences
  - Or Utilize StandardScaler()
- Reduce The Number Of Features In XGBoost
- Fit an XGBoost Model For Ranges of Seed Differentials
- Include More Opponent-Specific Metrics (Conf, SoS, NCSoS)
- Be More Keen On Overfitting

Thank You!
Any Questions?

Sports Analytics Society