Mining NCAA basketball data to inform March Madness predictions

Team Members

Ishika Patel

ispa0196@colorado.edu
CSCI 5502

Ethan Meyer

etme9299@colorado.edu
CSCI 4502

Agenda

01 **Re-introduction**
Reminder of proposed work

02 **Timeline Check-in**
Status of completed work relative to milestones

03 **Overview of Completed Work**
Summary of conducted analyses

04 **Issues & Risks**
Overview of current roadblocks

Re-introduction

Reminder of proposed work

Reminder of Proposed Work

Given NCAA team-specific statistics and Vegas moneyline odds we wish to **leverage game predictions to inform bets on tournament game outcomes**. We would like to place a specific emphasis on how recent team performance (e.g. team/player streaks, upsets in game-play, general team stats) can inform on future game outcomes.

Timeline Check-in

Status of completed work relative to milestones

Timeline Check-in

Data Cleaning & Integration

Deadline: Oct. 13
Completion: **100%**

Exploratory Data Analysis

Deadline: Oct. 20
Completion: **100%**

Attribute Computation & Subset Selection

Deadline: Nov. 10
Completion: **30%**

Model Creation Examination & Revision

Deadline: Nov. 17
Completion: **10%**

Result Synthesization

Deadline: Dec. 6
Completion: **0%**

Overview of Completed Work

Summary of conducted analyses

Data Cleaning and Integration

Spelling Key

TeamNameSpelling	TeamID
tennessee tech	1399
tennessee-chattanooga	1151
tennessee-martin	1404
tennessee-st	1398
tennessee-state	1398

Rarely exact match

Moneyline Data

Date	VH	Team	ML
1105	N	Maine	200
1105	N	Richmond	-240
1105	V	TennMartin	13000
1105	H	MemphisU	-39000
1106	N	GardnerWebb	-160
1106	N	AlabamaA&M	140

Game Data

Season	DayNum	WTeamID	LTeamID
2008	0	1272	1404
2008	0	1350	1263
2008	1	1205	1105
2008	1	1246	1146
2008	1	1272	1350
2008	1	1404	1263

Complicated parsing

Data Cleaning and Integration

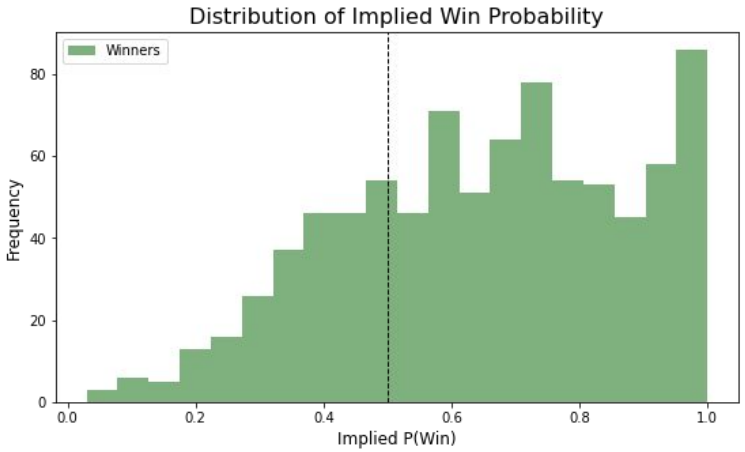
$$P(win) = \begin{cases} \frac{|odds|}{|odds|+100} & \text{if } odds < 0 \\ \frac{100}{odds+100} & \text{if } odds > 0 \end{cases}$$

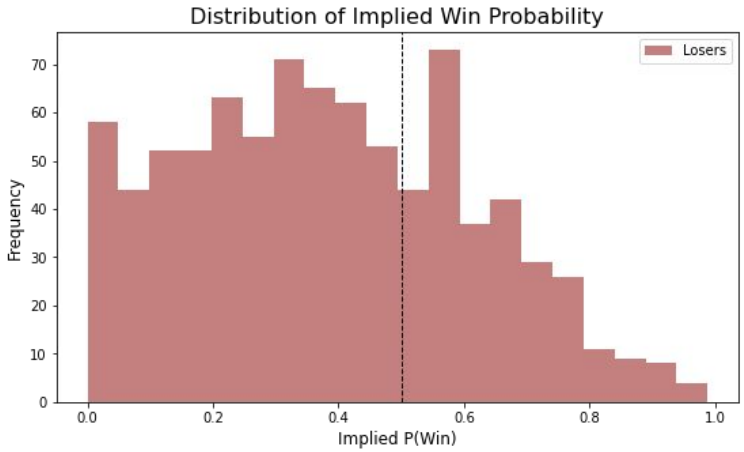
Data Cleaning and Integration

Integrated Data

	Season	DayNum	WTeamID	LTeamID	WMoneyline	LMoneyline	WTeam_Win%	LTeam_Win%
0	2008	134	1291	1164	-375.0	315.0	0.789474	0.240964
1	2008	136	1181	1125	-4500.0	2250.0	0.978261	0.042553
2	2008	136	1242	1340	-5000.0	2500.0	0.980392	0.038462
3	2008	136	1243	1425	140.0	-160.0	0.416667	0.615385
4	2008	136	1266	1246	-275.0	235.0	0.733333	0.298507

Exploratory Data Analysis: Moneyline Data



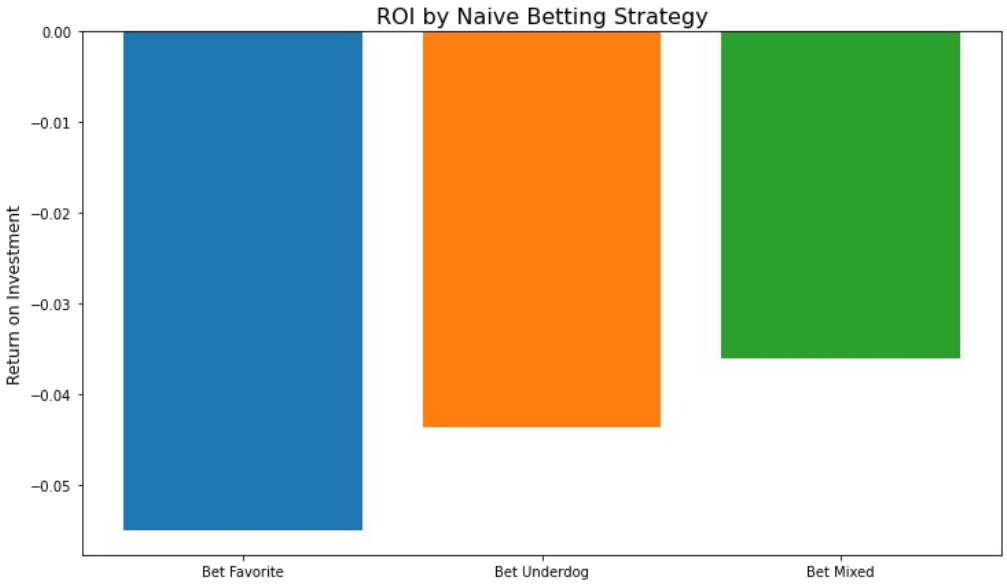


Exploratory Data Analysis: Naive Strategies

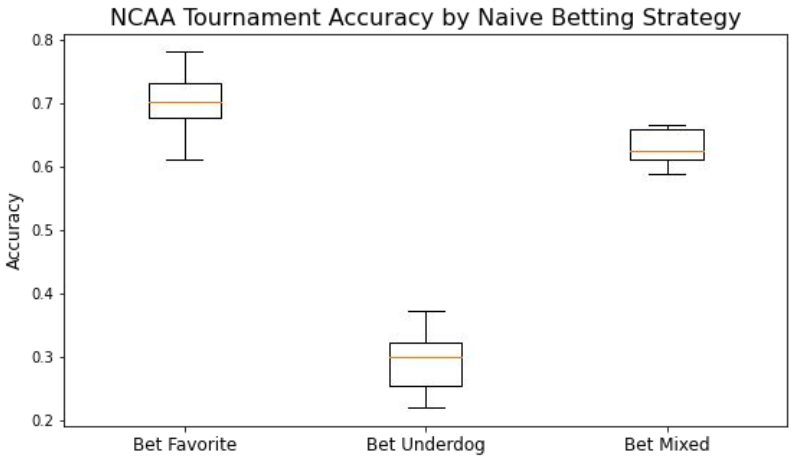
1. **Bet Favorite**
   Always bet on team with ***higher*** implied probability

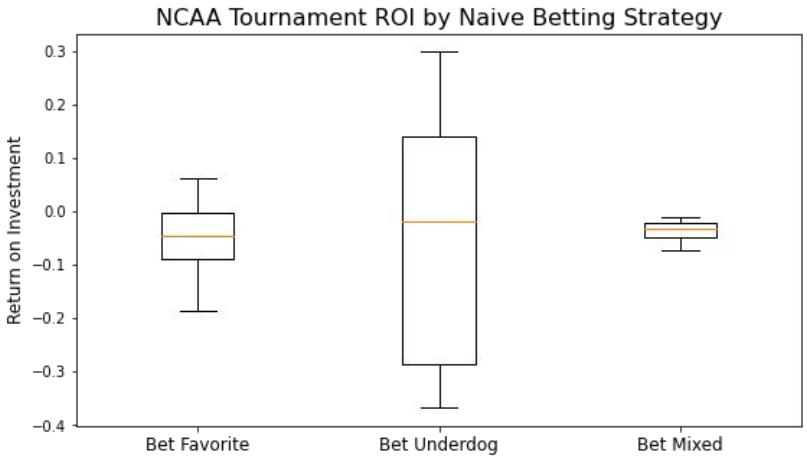
2. **Bet Underdog**
   Always bet on team with ***lower*** implied probability

3. **Bet Mixed**
   Bet in alignment with ***random choice*** based on implied win%

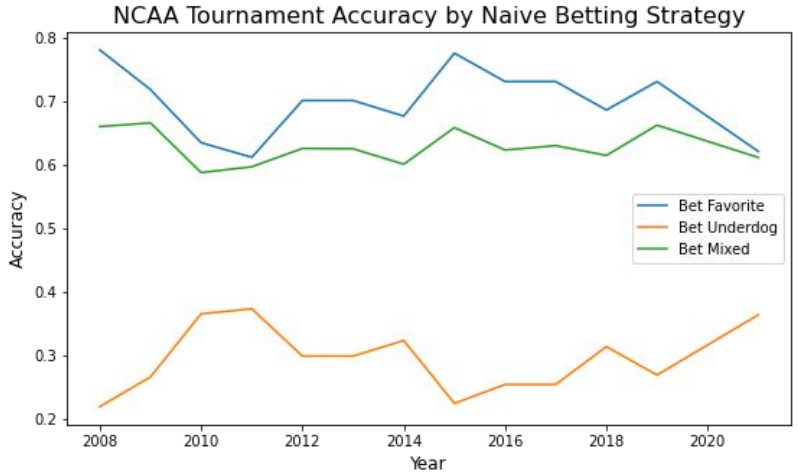


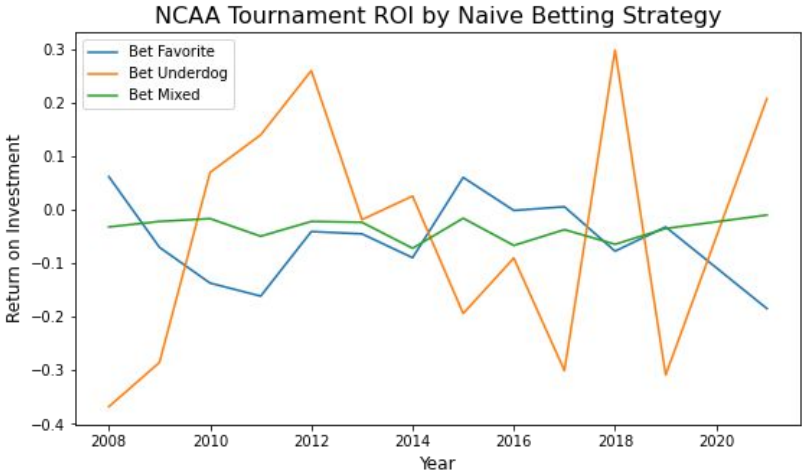
Exploratory Data Analysis: Naive Strategies



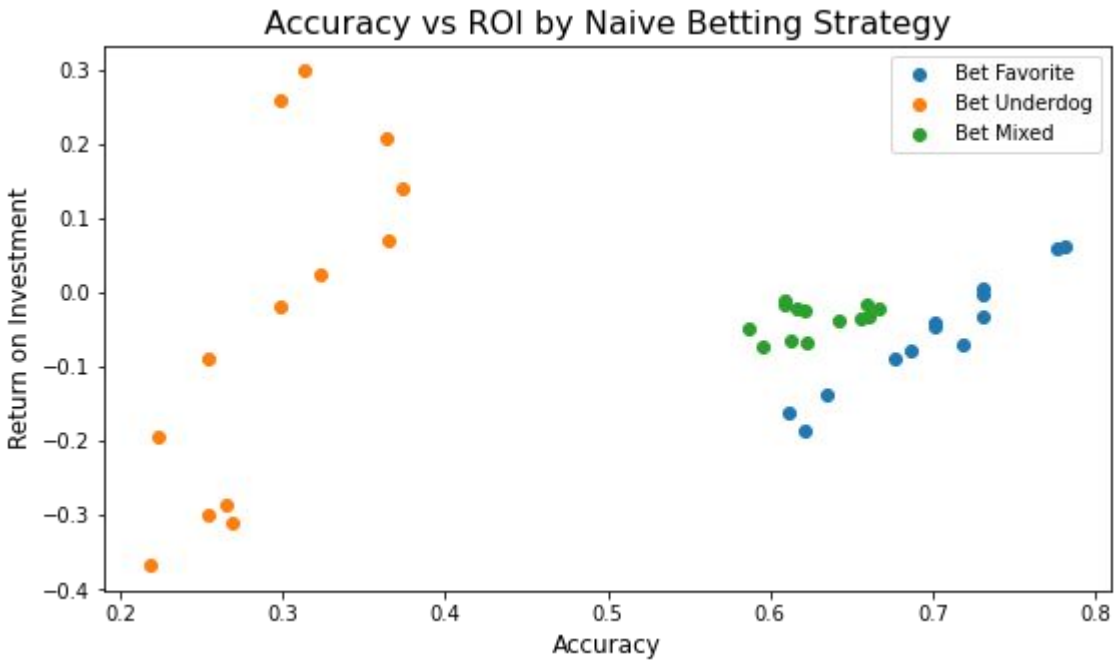


Exploratory Data Analysis: Naive Strategies





Exploratory Data Analysis: Naive Strategies



Attribute Subset Selection

- Dean Oliver (American statistician and assistant coach for the NBA's Washington Wizards) wrote a paper on Four Factors that contribute to a winning basketball team based on box score data an how impactful these are to the win (weighted).

1. **Effective Field Goal Percentage**
   Weighted probability of 2-point and 3-point field goals a team makes

2. **Turnover Percentage**
   Percentage of a team's possessions that end in a turnover

3. **Offensive Rebound Percentage**
   An estimate of the percentage of available offensive rebounds a team grabbed

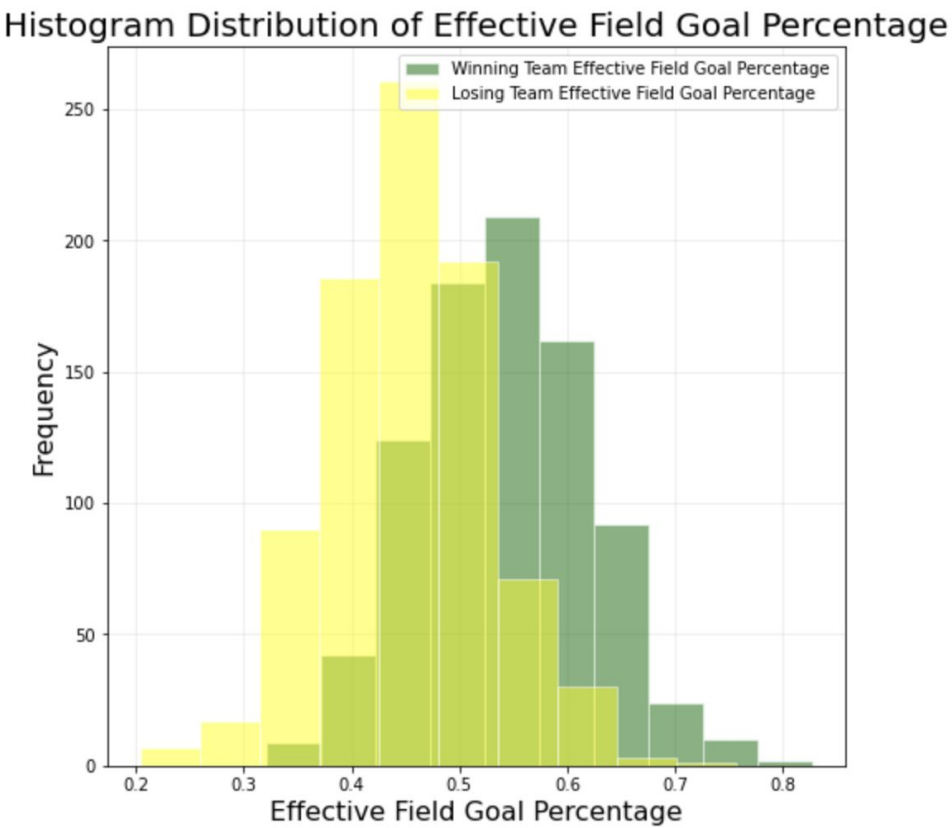
4. **Free Throw Rate**
   How often a team gets to the line and how often they make them

- We have chosen these four factors in the upfront attribute selection as a baseline to encompass the detailed data given by Kaggle

Attribute Computation and Selection

1 Effective Field Goal Percentage

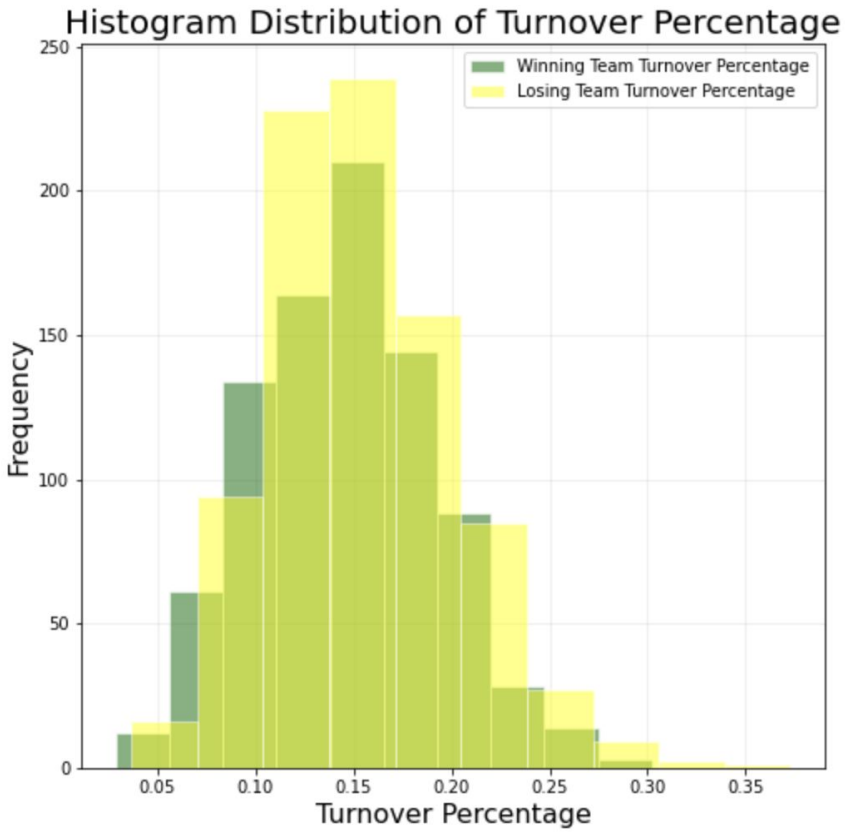
- 40% weight contribution to a winning team
- Leverages Field Goals Made, 3-Pointers Made, Field Goals Attempted
- Winning teams from the data set have a clear greater EFGP



Attribute Computation and Selection

2 Turnover Percentage

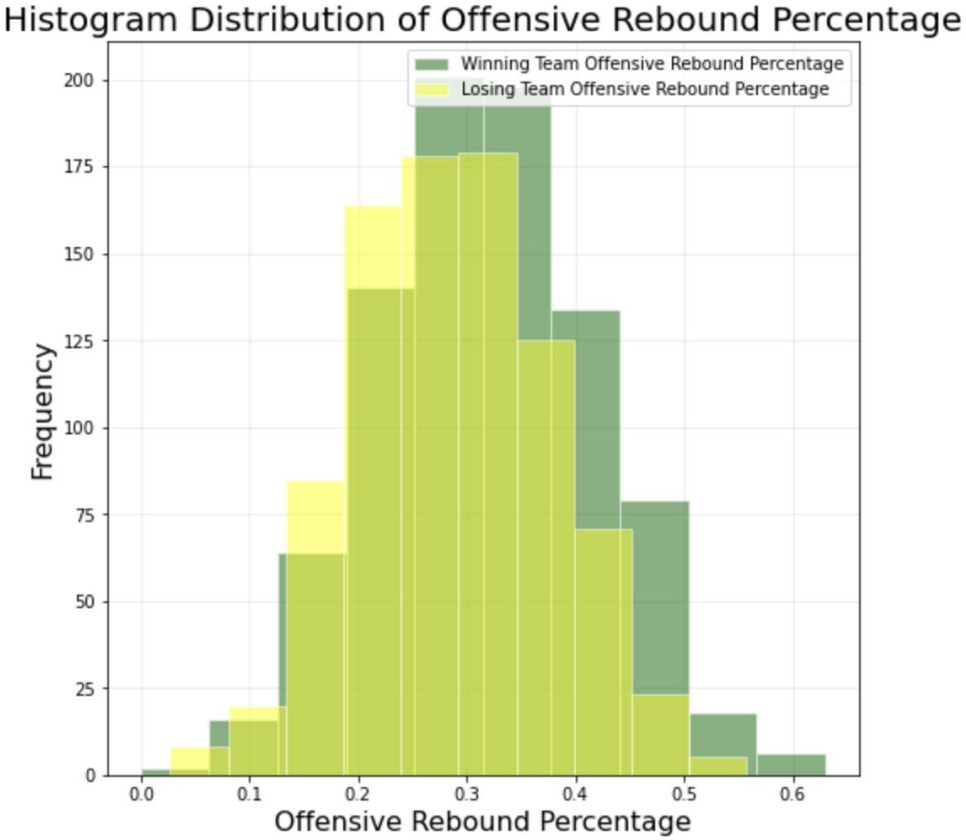
- 25% weight contribution to a winning team
- Leverages Turnover Count, Field Goals Attempted, Three Throws Attempted
- There is greater losing team turnover frequency
- Teams ideally want a low turnover rate



Attribute Computation and Selection

3 Offensive Rebound Percentage

- 20% weight contribution to a winning team
- Leverges Offensive Rebound and Opposing Team's Defensive Rebounds
- Winning teams grab offensive rebounds with slightly greater frequency than losing teams

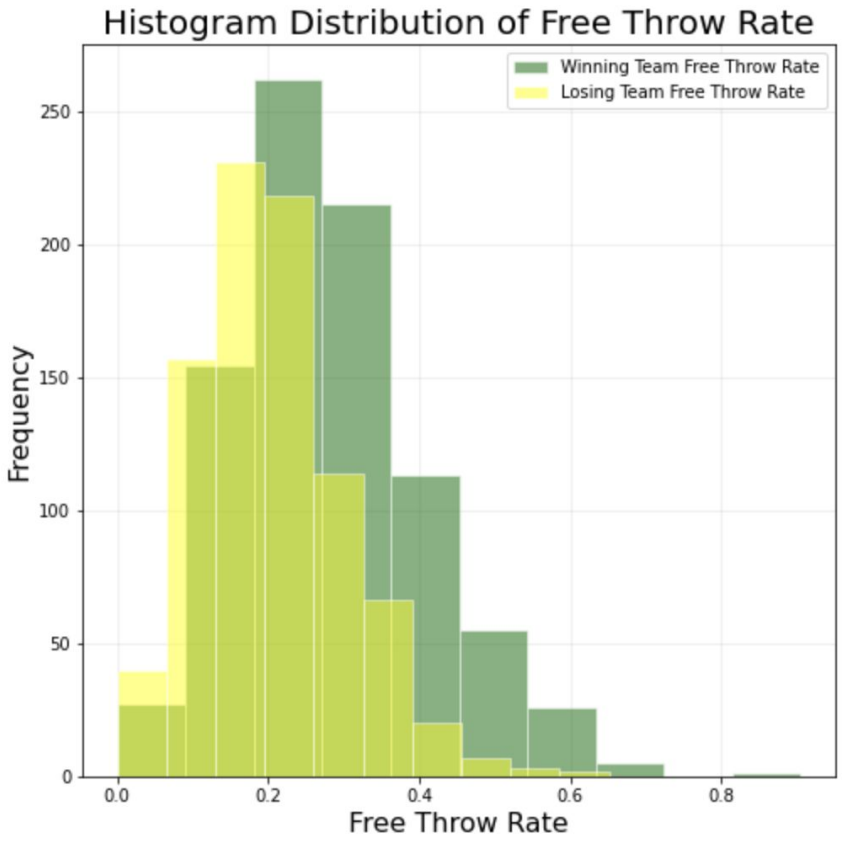


Attribute Computation and Selection

4

Free Throw Rate

- 15% weight contribution to a winning team
- Leverages Free Throws Made and Field Goals Attempted
- Winning take and make free throws at a greater rate than losing teams



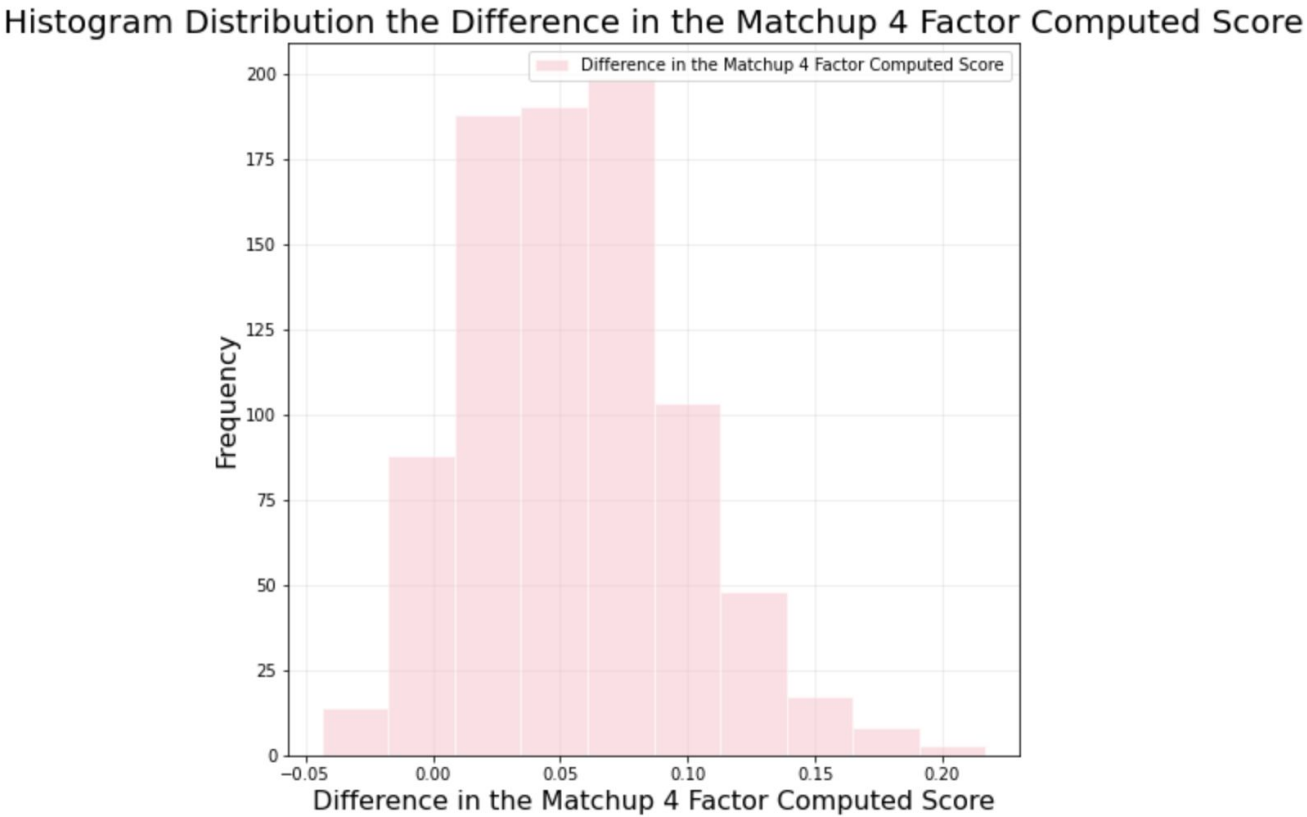
Four Factor Model Creation

- We computed a per game estimate of **each teams four-factor winning score** based on the weights of attribute contribution and computed a **difference** between the winning team and losing team's four-factor score
- Place bets with a similar algorithm

```
if difference >= alpha:
    Bet on winning team
if difference < -alpha:
    Bet on losing team
else:
    Do not bet
```

- Four Factor Model:
  - Average the W_4/L_4 team metrics per team id for an average metric of the team's performance from a training set
  - Use these stats to place bets in matchups on a test set
  - Compare confidence in betting predictions made across models

Four Factor Model Creation



Issues & Risks

Overview of current roadblocks

Issues & Risks

Data Integration

- Unanticipated challenges regarding connecting data sources
- Resolved, but has delayed other aspects of project

Feature Computation

- Underestimated difficulty of navigating Kaggle data structure (very granular)

Defining Betting Strategy

- Not yet certain how to optimize betting strategy based on model's predictions
  - All games? Equal sized bets?
  - Edge? Model's P(win) v.s. Vegas P(win)
- Little available research on this

Model Creation

- Using the same factors on different models vs using different factors in different models

Questions & Feedback