NBA Game Score Difference Prediction

Paul Jacob

Business Problem

Problem: Retail sportsbook bettors want to bet on games and turn a profit.

Question: Can we provide guidance on which games to bet on and which games to avoid?

Solution: Provide the data point of a good NBA game score difference prediction model for sportsbook bettors.

Overview

- Create important features that give insight into NBA game score difference.

- Select a good model for NBA game score difference prediction.

- Use insights in the NBA sportsbook spread market by viewing price and spread to gain an advantage.

Data Collection

- Collect data from Kaggle and nba_api
- Read data into a Jupyter Notebook for Data Wrangling

Data Wrangling: Cleaning Up

- Edit player naming differences in the injury list to match the inactive list:

    'Ron Artest' → 'Metta World Peace'
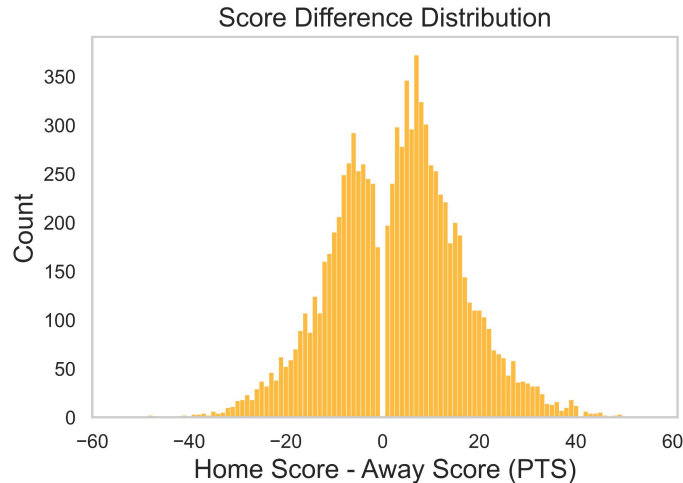   'Marcus Morris' → 'Marcus Morris Sr.'
    'J.J. Redick' → 'JJ Redick'
    'Jose Barea' → 'J.J. Barea'

- Edit Other:

    'Golden State' → 'Oakland'

Exploratory Data Analysis: Target Distribution

- Score Difference = Home Score - Away Score.
- Score Difference is later compared with sportsbook Spread1.

Score Difference Distribution

Feature Engineering: Overview

| Feature Category	Explanation
Box score stats	Points, 3-point field goal attempts, personal fouls, field goal percentage
Advanced box score stats	Net rating, player impact estimate, effective field goal percentage, true shooting percentage
Lost contribution	Player did not play
Schedule	Strength of schedule
Location	Travel distance

Feature Engineering: Lost Contribution

- Team Mean **Effective Field Goal Percentage** Lost Contribution Season Max
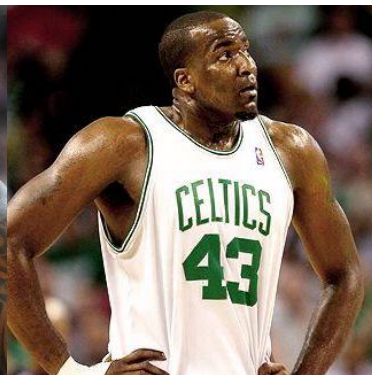

2010-11 Season

inactive — Luke Harangody
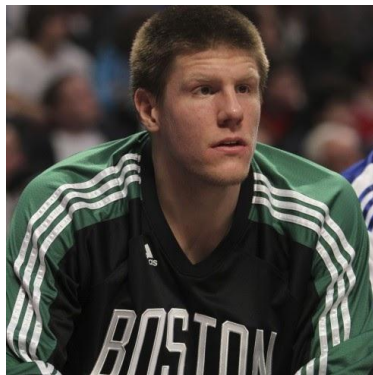
injured — Shaquille O'Neal

inactive — Kendrick Perkins

Team Mean Effective Field Goal Percentage Lost Contribution =

$$\frac{1.0 + 0.57 + 0.63}{3} = .73$$

Random Forest: Modeling & Feature Selection

- Modeling
  - Cross Validate with NBA Team Box Score Features to get best hyperparameters.
  - Perform Random Forest training on smaller datasets following creation of features.

- Feature Selection
  - Take top 50 feature importances of each sub-dataset by Random Forest Mean Decrease in Impurity (MDI).
  - Combine sub-datasets to make 100 features to learn from.
  - Further reduce to 89 by eliminating very high correlations among features.
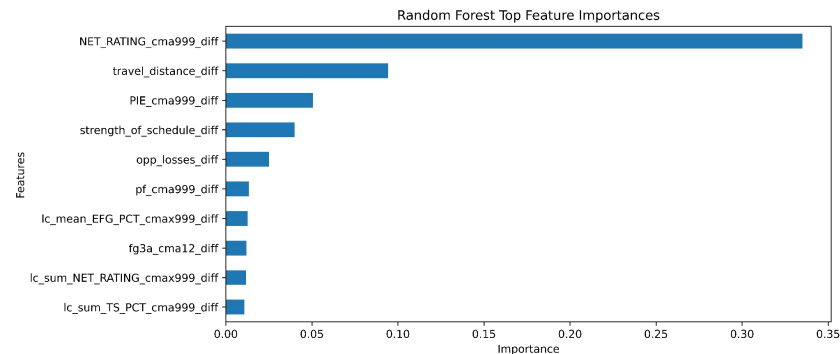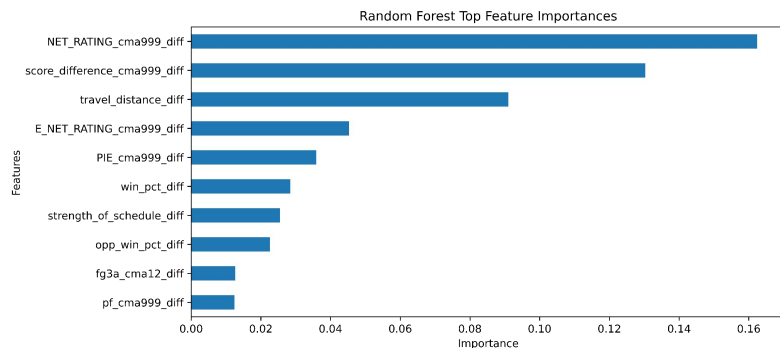
Model Performance

Model Metrics

	MAE	MAPE
Random Forest	9.683	112.866
Gradient Boosting	9.684	113.362
Linear Regression	9.687	111.457

# Betting Policy: Learn Heuristic

- Learn Heuristic:
    - 1. Run betting policy combinations on train:
        - Prediction confidence interval: 1%, 2%, …, 100%.
        - Absolute value of spread lower limit: 0, .5, …, 22.
        - Price break even upper limit: 0.504, 0.507, ..., max.
    - 2. Take learnings from 1. and run a top betting policy combination on each train and validation.
    - 3. Repeat 2. for convergence to heuristic methodology.


Learn and Test Heuristic
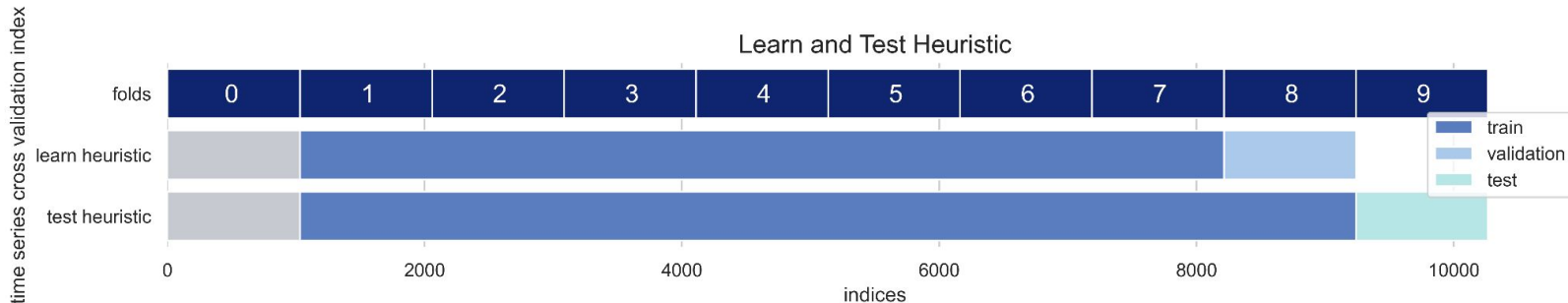
Betting Policy: Test Heuristic

- Test Heuristic:
    - 1. Run betting policy combinations on train
    - 2. Apply heuristic methodology from learn heuristic to get betting policy:
        - Score Difference Confidence Interval: 27%
        - Spread Absolute Value Lower Limit: 9.5
        - Price Break Even Upper Limit: 0.5098 and 0.5122

Learn and Test Heuristic
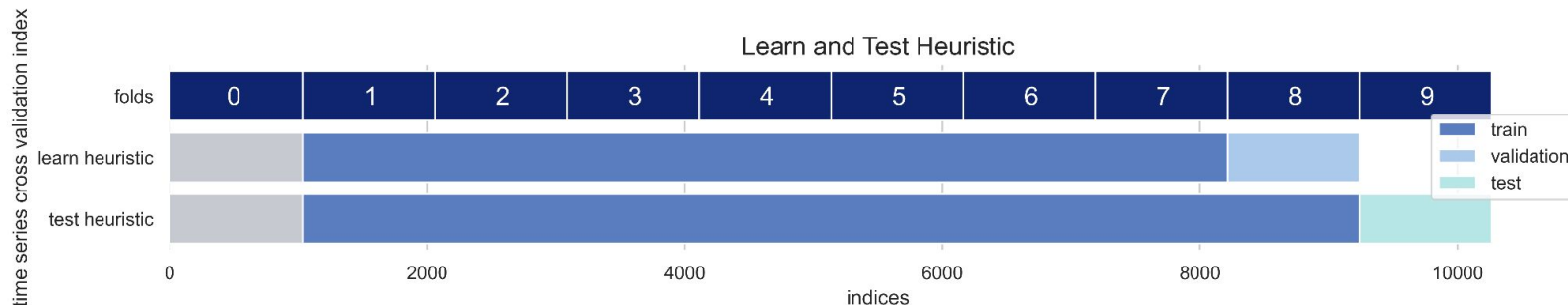
Results

2017-18 Season
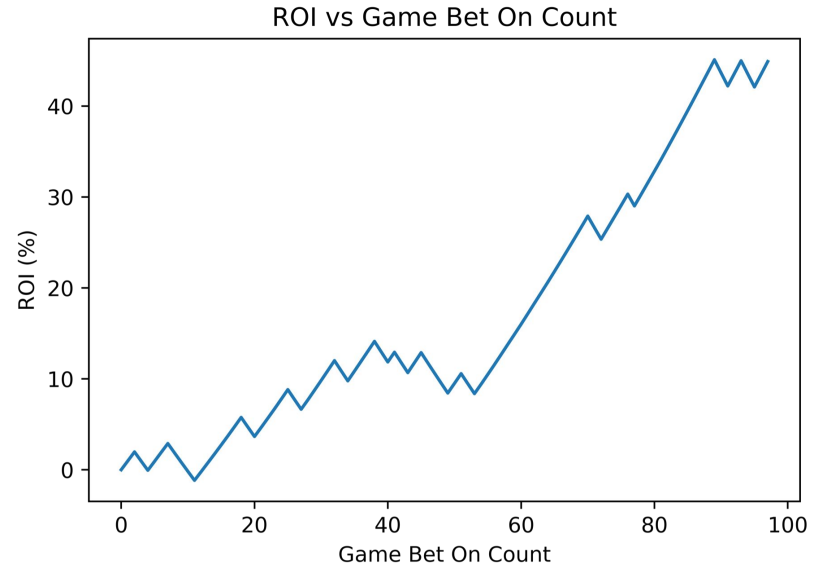
Conclusion

- Using a good score difference prediction model and betting policy, positive net profits are feasible.
- Future Work: Collect more recent NBA data and further test our model

Questions

Appendix

Exploratory Data Analysis: Away-Home Win Percentage

- Away teams of the 250-500 mile travel distance range lose the most frequently on average.
- Long distance games in the NBA are infrequent and see the most variation in win percentage from season to season

Travel Distance Interval vs. Win Percentage
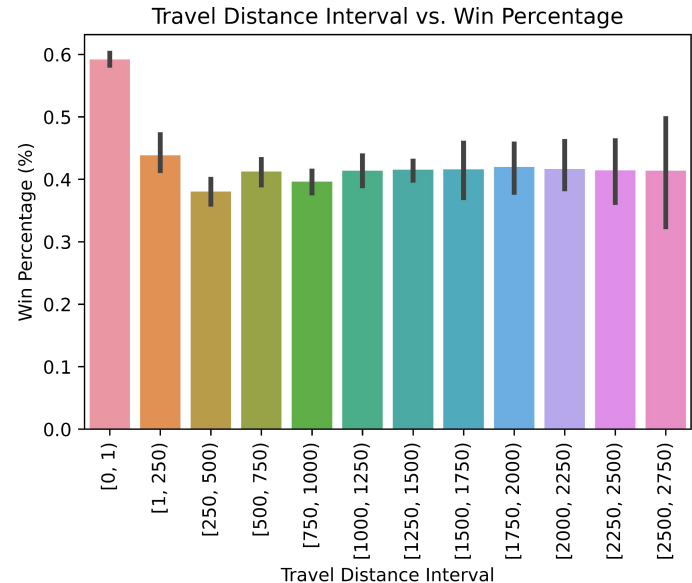
Exploratory Data Analysis: Away vs. Home Win Percentage

- Away teams win 41% of the time on average.
- Home teams win 59% of the time on average.

Away and Home Win Percentage
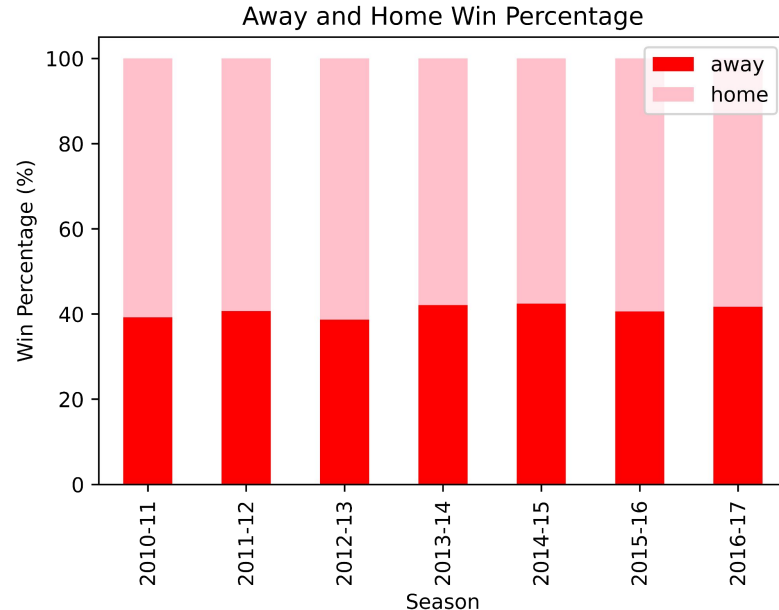
Lost Contribution

- Team Sum **Net Rating** Lost

  Contribution Season Max

2017-18 Season

Team Sum Net Rating Lost
Contribution =

$4.2 + (-1.7) + (-5.1) +$

$4.9 + 250.0 + (-7.4) = 244.9$

Raul Neto
inactive & injured

Dante Exum
inactive & injured

Rudy Gobert
injured

Tony Bradley
inactive

Naz Mitrou-Long
benched

Jonas Jerebko
benched

Lost Contribution

- Team Player Max **Assist To Turnover Ratio** Lost Contribution Season Max
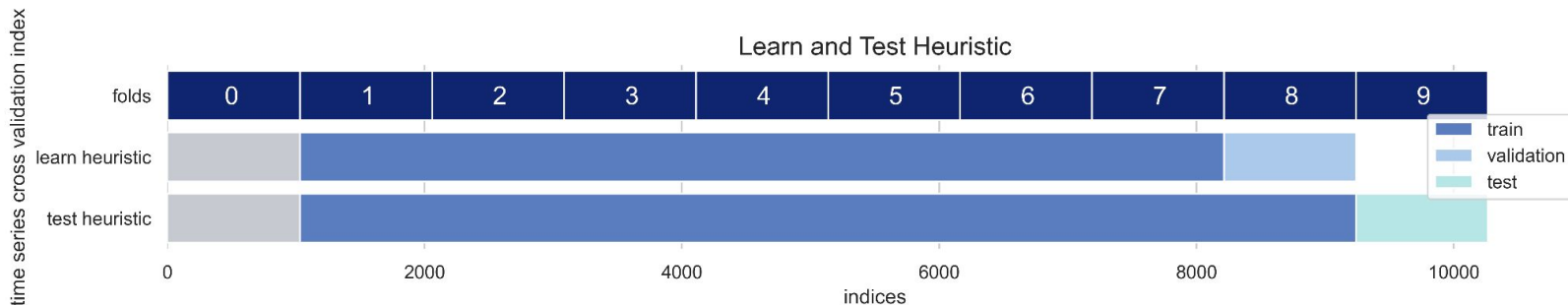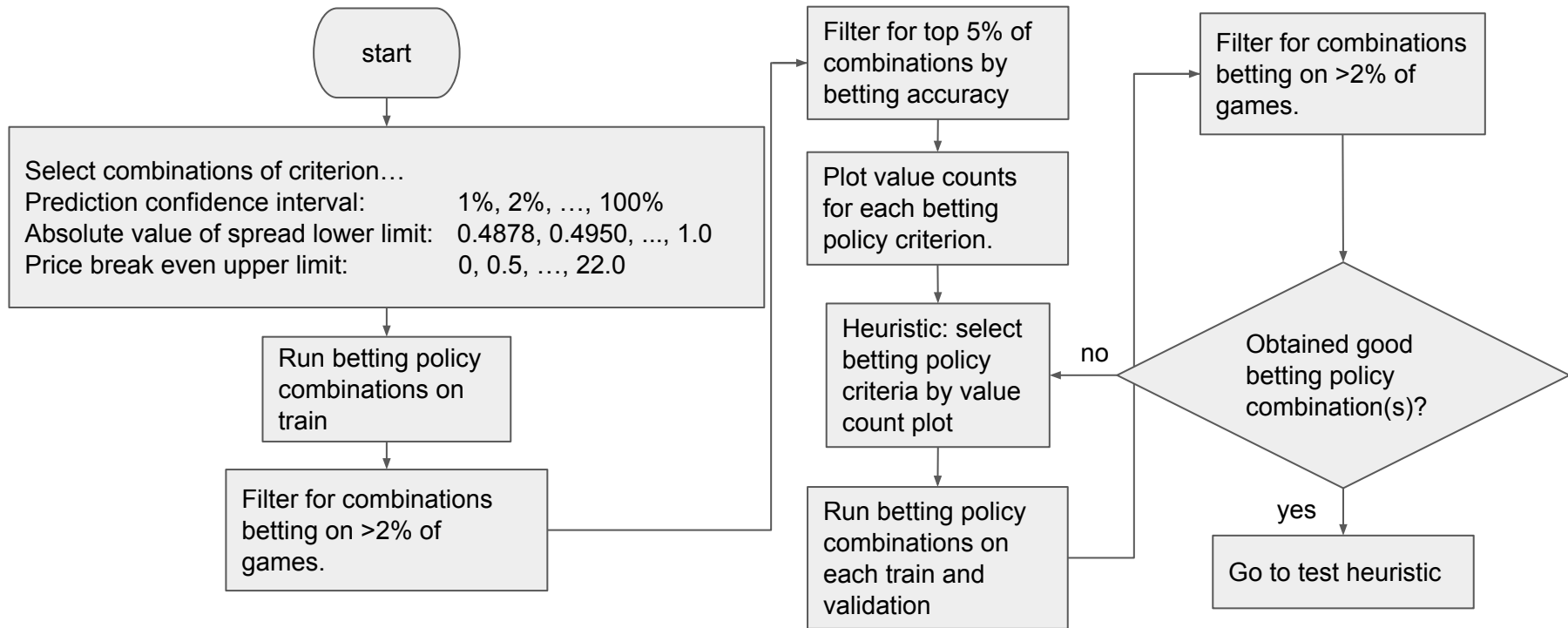
2017-18 Season

Chris Paul
inactive & injured

$$= 10.0$$

Betting Policy: Learn Bettor's Advantage

- Create generalization for betting policy using either a machine learning model or heuristic.
- Learn Heuristic
  - 1. Select combinations of criterion:
    - Prediction confidence interval: 1%, 2%, …, 100%.
    - Absolute value of spread lower limit: 0.4878, 0.4950, ..., max.
    - Price break even upper limit: 0, 0.5, …, 22.0.
  - 2. Run betting policy combinations on train.
  - 3. Select top 5% of betting policy combinations by win percentage with a >2% of games bet on.
  - 4. Select betting policy by criterion value count plot.
  - 5. Run betting policy combination on each train and validation.
  - 6. Confirm betting policy has acceptable accuracy and >2% of games bet on, otherwise return to 4.
- Test Heuristic
  - 1. Learn Heuristic 1-4.
  - 2. Run betting policy combination on train.
  - 3. Confirm betting policy has acceptable accuracy and >2% of games bet on, otherwise return to 1. Learn Heuristic 4.
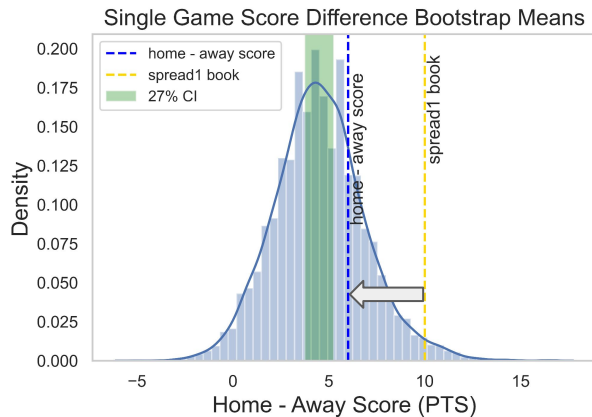 - 4. Run betting policy combination on test.

Betting Policy: Learn Heuristic

start

Select combinations of criterion…
Prediction confidence interval: 1%, 2%, …, 100%
Absolute value of spread lower limit: 0.4878, 0.4950, ..., 1.0
Price break even upper limit: 0, 0.5, …, 22.0

Run betting policy combinations on train

Filter for combinations betting on >2% of games.

Filter for top 5% of combinations by betting accuracy

Plot value counts for each betting policy criterion.

Heuristic: select betting policy criteria by value count plot

Run betting policy combinations on each train and validation

Filter for combinations betting on >2% of games.

Obtained good betting policy combination(s)?
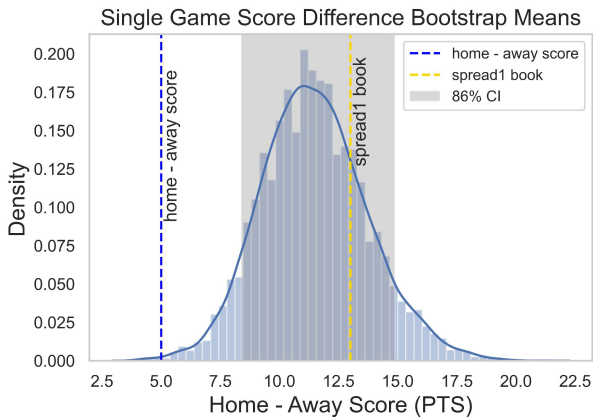
no

yes

Go to test heuristic

Betting Policy Criterion: Prediction Confidence Interval

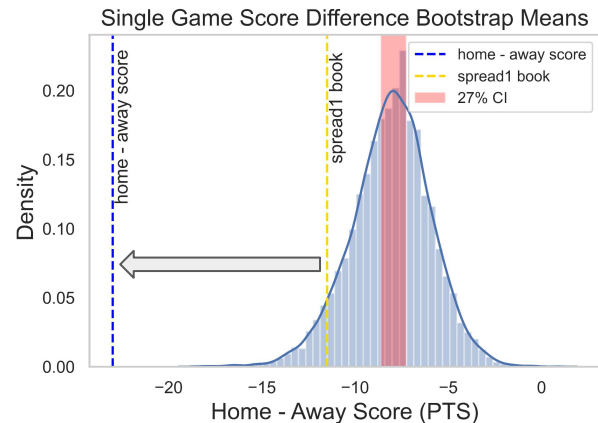● Avoid games where sportsbook spread and predicted score difference are too similar.

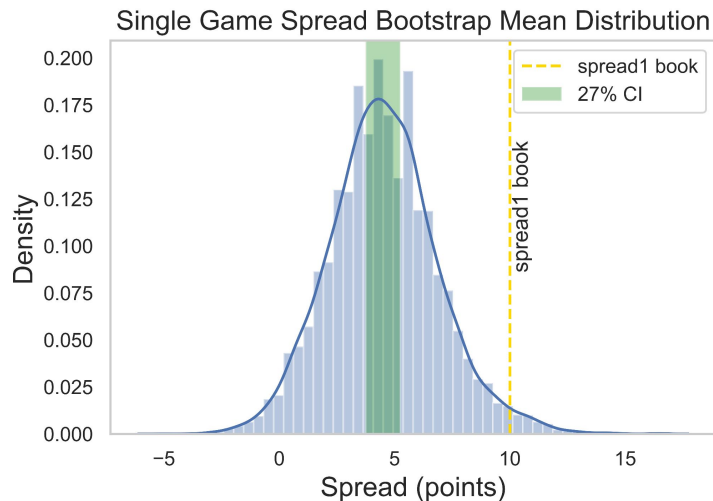Correct Side of Book Spread | No Bet B/c Interval Contains Book Spread | Wrong Side of Book Spread
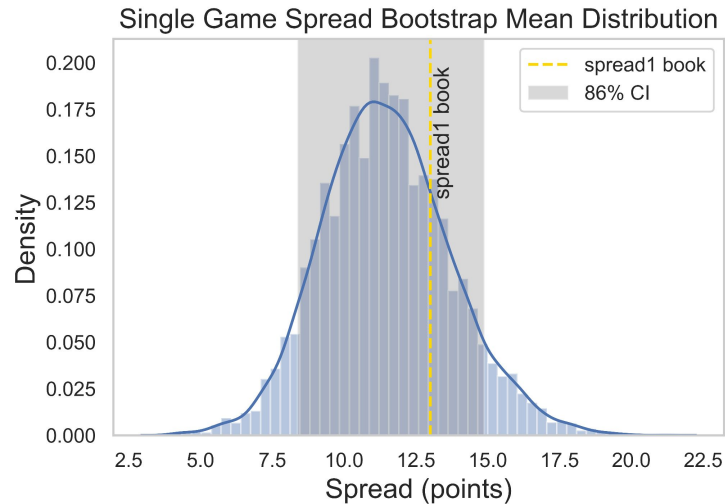
Betting Policy Criterion: Prediction Confidence Interval
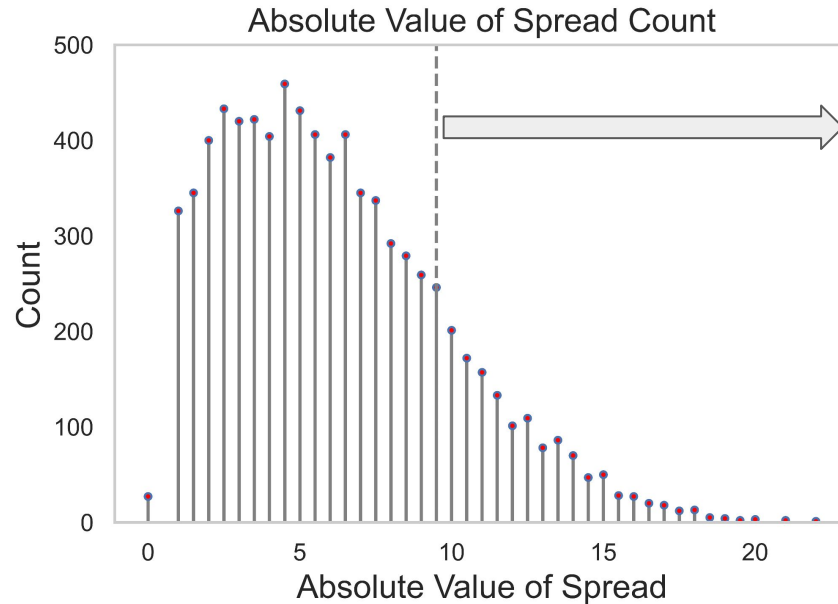
Yes Bet B/c Interval
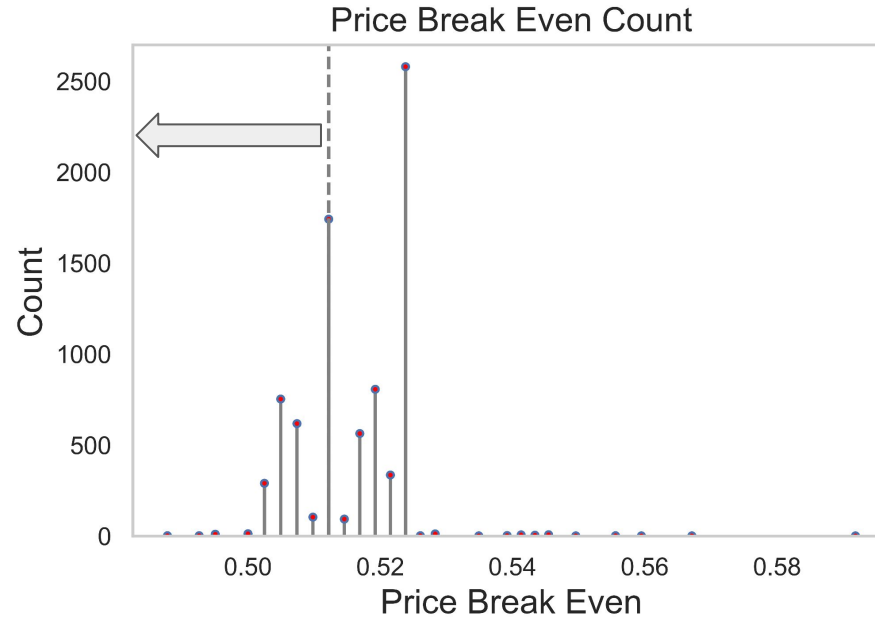Outside Book
Spread

No Bet B/c
Interval Contains
Book Spread

Single Game Spread Bootstrap Mean Distribution

Single Game Spread Bootstrap Mean Distribution

Betting Policy Criterion: Absolute Value of Spread Lower Limit

● Seek games with a spread-represented large favorite

Betting Policy Criterion: Price Break Even Upper Limit

- Avoid games with low payouts

Test: ROI Confidence Interval

- 1% of account balance per bet.
- Expect a ≥22% ROI 90% of the time.

2017-18 Season

ROI Distribution