ATP Match Statistics

FINDING THE WINNING FORMULA

“Winning isn’t everything, it’s the only thing.

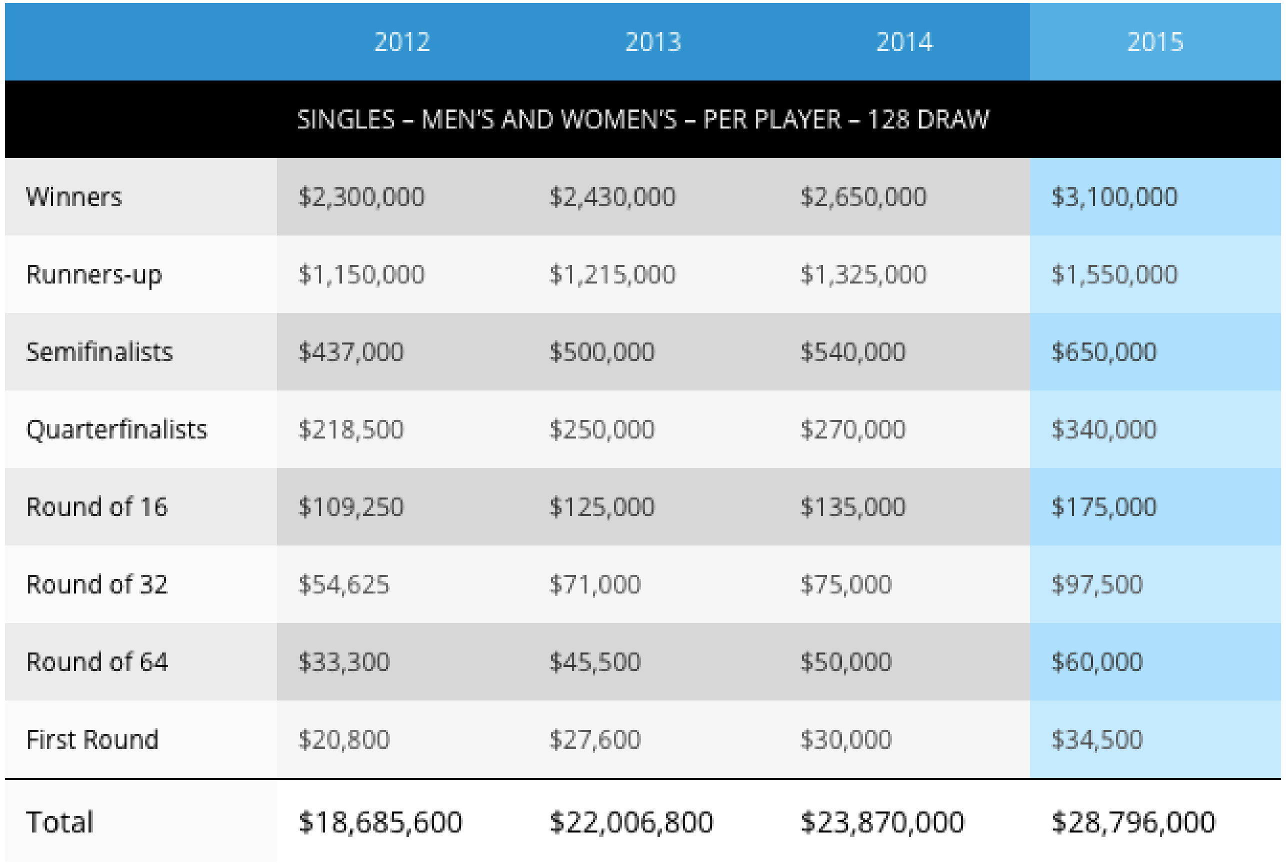
Vince Lombardi

Tennis coaching is expen$ive.

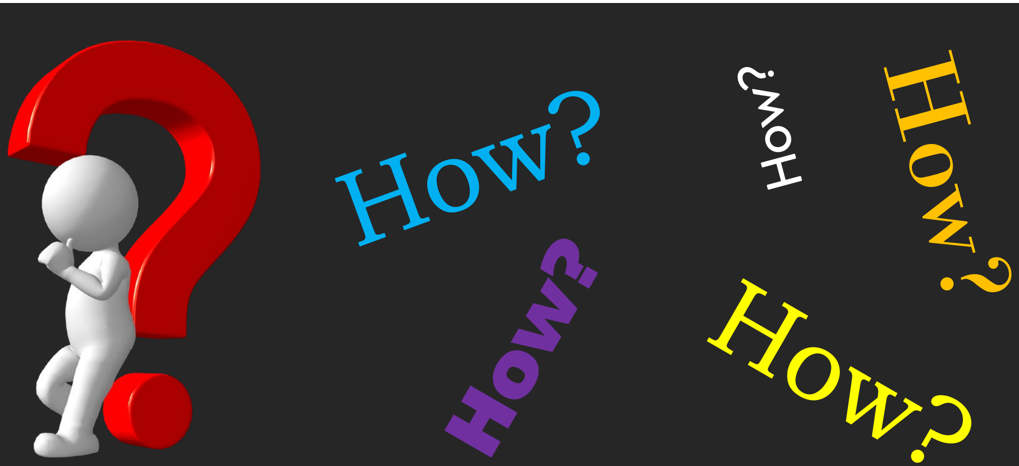
50% OF THE 14,000 PROFESSIONAL TENNIS PLAYERS WIN $0 PRIZE MONEY, YET CAN SPEND $50,000, $100,000…OR MORE… IN COACHING & TRAVEL EXPENSES!

Winning Pays.

WINNERS RECEIVE TWICE AS MUCH AS THEIR 2ND PLACE OPPONENT. FIRST ROUND LOSERS RECEIVE ABOUT 2% OF WHAT THE WINNER RECEIVES.



	2012	2013	2014	2015
SINGLES – MEN'S AND WOMEN'S – PER PLAYER – 128 DRAW				
Winners	$2,300,000	$2,430,000	$2,650,000	$3,100,000
Runners-up	$1,150,000	$1,215,000	$1,325,000	$1,550,000
Semifinalists	$437,000	$500,000	$540,000	$650,000
Quarterfinalists	$218,500	$250,000	$270,000	$340,000
Round of 16	$109,250	$125,000	$135,000	$175,000
Round of 32	$54,625	$71,000	$75,000	$97,500
Round of 64	$33,300	$45,500	$50,000	$60,000
First Round	$20,800	$27,600	$30,000	$34,500
Total	$18,685,600	$22,006,800	$23,870,000	$28,796,000



WHAT STATISTICS MATTER?

WHAT TO PRACTICE?

WHAT TO WORK ON?

WHAT WILL IMPROVE CHANCES OF WINNING?

Develop Machine Learning Model to Predict Match Winners

USING ATP MATCH STATISTICS:

1. IDENTIFY WHICH FEATURES (MATCH STATISTICS) ARE THE MOST MEANINGFUL.

2. EXPLORE RELATIONSHIP BETWEEN MATCH STATISTICS AND FOR MATCH WINNERS AND LOSERS.

3. CREATE MACHINE LEARNING MODEL TO PREDICT MATCH WINNERS.

The Data.

OVERVIEW.

- ATP World Tour Website as distributed on datahub.io
- 53 unindexed CSV files broken into 5 different categories
- Project data used was from 1991 to 2016
- 93,359 match scores in 2,054 tournaments
- All csv files contain uncategorized data.

The Data.

INTERESTING STATISTICS.

Match Averages

- Length: 1 h 44 m 36 s
- Sets played: 2.54
- Games played: 24.6
- Points played: 157

The Data.

INTERESTING STATISTICS.

Winner Averages

- Ranked higher 65.5% of the time.
- Wins 94.3% of all points played.
- Wins 10.9% more points than loser.
- **Wins 0.58 more points per game.**

Even though the winner wins more points, the margin of victory per game, on average, is very small.

Predict Matches

THE CHALLENGE.

Since the Winner wins 94.3% of the points, it should be easy to say, “Just win more points.” Because of the unique scoring and structure of tennis, that’s not very meaningful, especially since there is no way to simply practice “winning points”.

What we need to know is HOW to win points. This will allow for effective practice sessions to improve the areas that will improve our ability to win more points, and thus win more matches.

Predict Matches

THE FEATURES.

Relevant Engineered Features

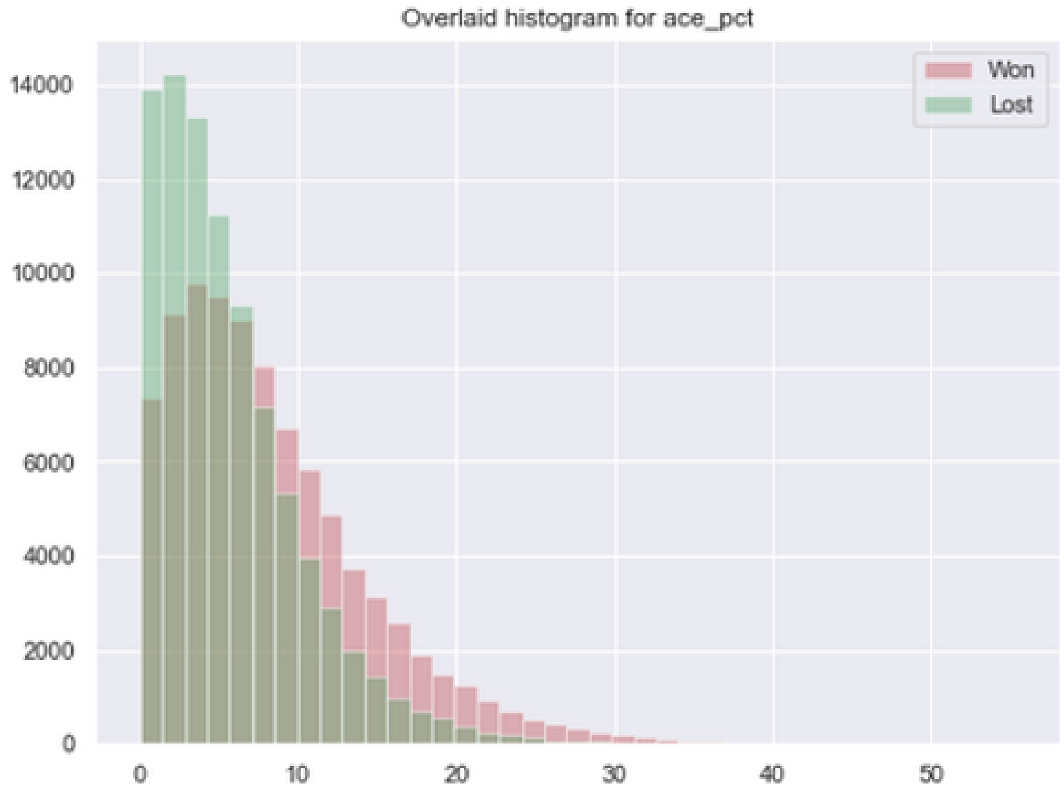
There was substantial overfitting from the initial dataset. We engineered features in order to address this by converting the original features to match percentages. This reduced overfitting and allowed us to develop a better baseline model.

- Percent of Service Aces to All Serves
- Percent of Service Double Faults
- Percent of First Serves In
- Percent of First Serve Points Won
- Percent of Second Serve Points Won
- Percent of Total Serve Points Won
- Percent of Return Service Points Won
- Percent of Break Points Converted

Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS



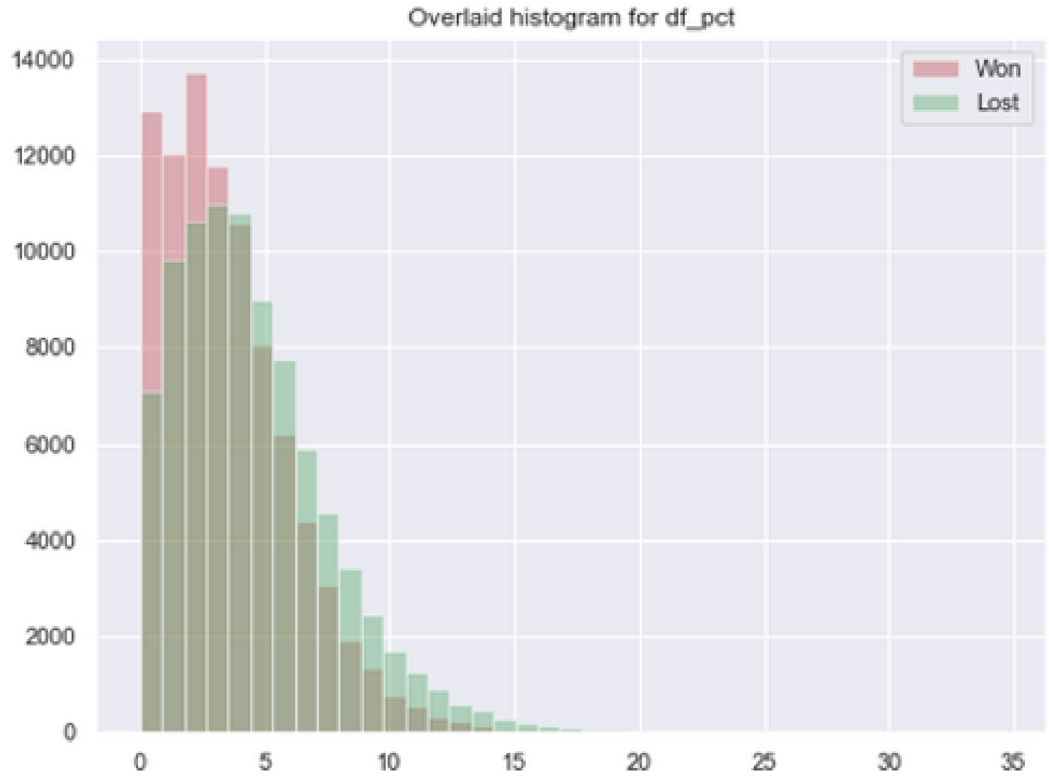
```
Wins Mean:
8.340683934072723
No-Wins Mean:
5.599402213071454
Mean Diff:
2.741281721001269
H0 Diff: 0

p: 0.0
CI: [-0.05431697
0.05277014]
ME:
0.053402597518318703
```

Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS



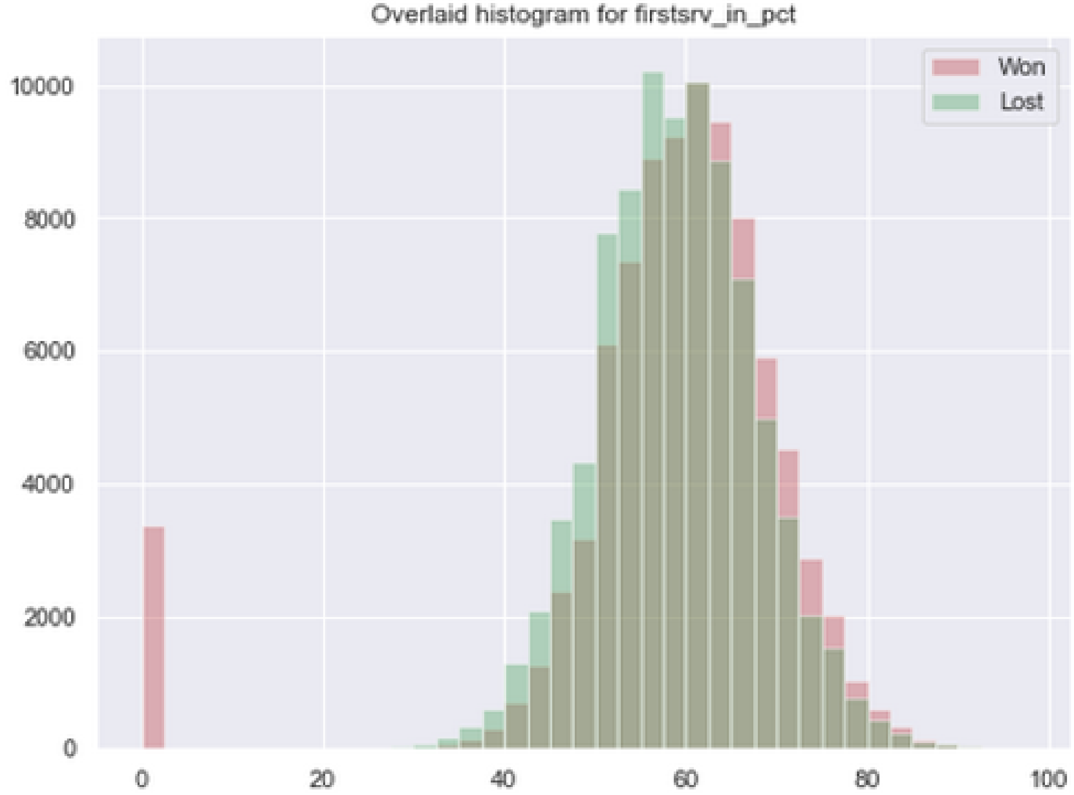
```
Wins Mean:
3.471326834959865
No-Wins Mean:
4.511129954525473
Mean Diff: -
1.0398031195656081
H0 Diff: 0

p: 0.0
CI: [-0.02795768
0.02747843]
ME: 0.0275627659420
```

Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS



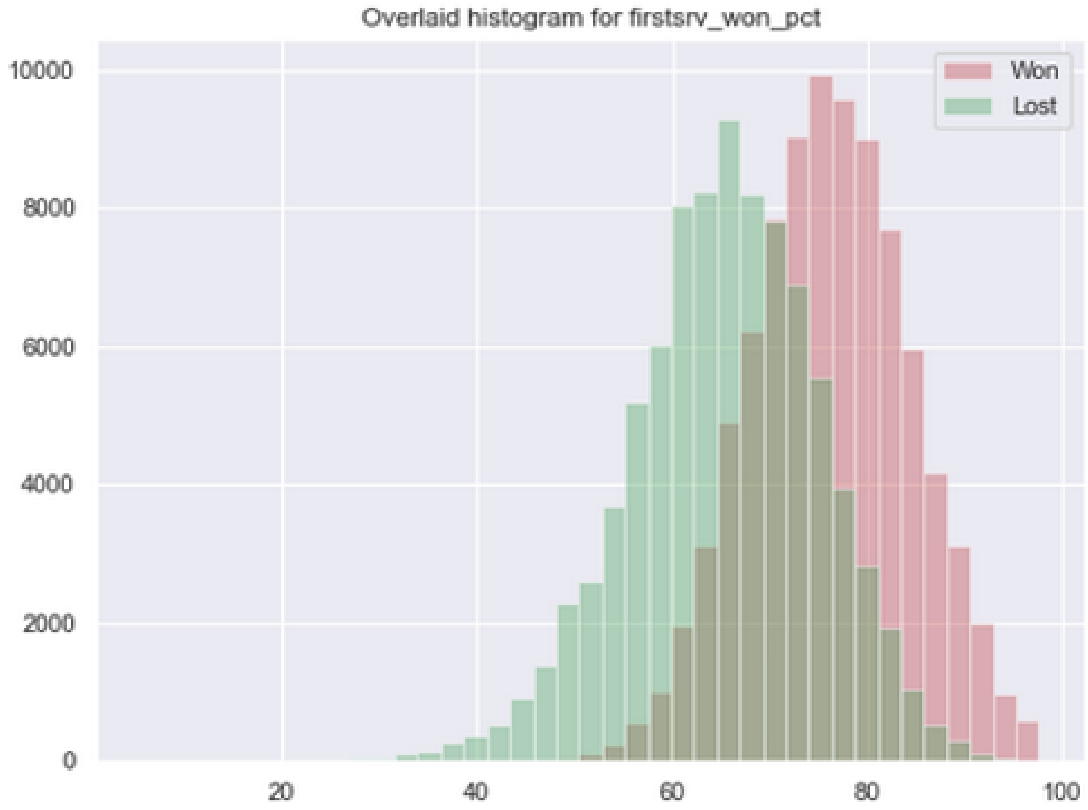
```
Wins Mean:
58.3512523156482
No-Wins Mean:
58.8646494896052
Mean Diff: -
1.0398031195656081
H0 Diff: 0

p: 0.0
CI: [-0.11174051
0.10695299]
ME:
0.10793662133669706
```

Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS



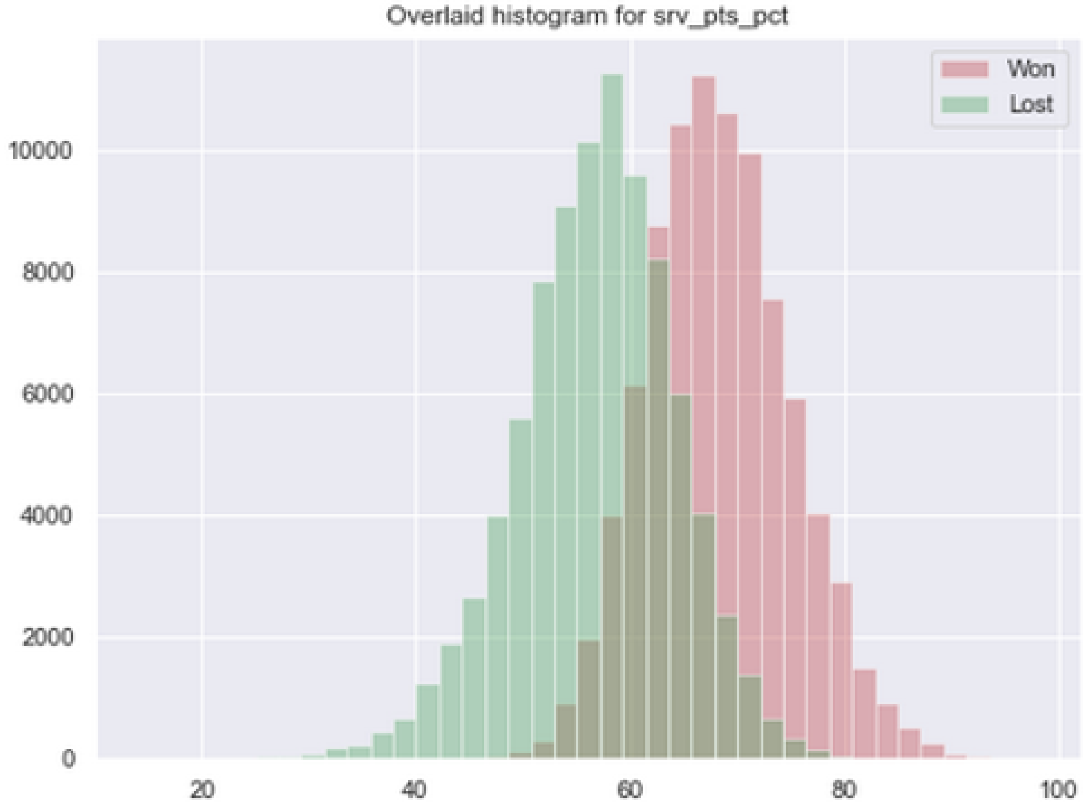
```
Wins Mean:
76.23225218749391
No-Wins Mean:
65.35481512219901
Mean Diff:
10.877437065294899
H0 Diff: 0

p: 0.0
CI: [-0.09521021
0.09905118]
ME:
0.09875387007053657
```

Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS



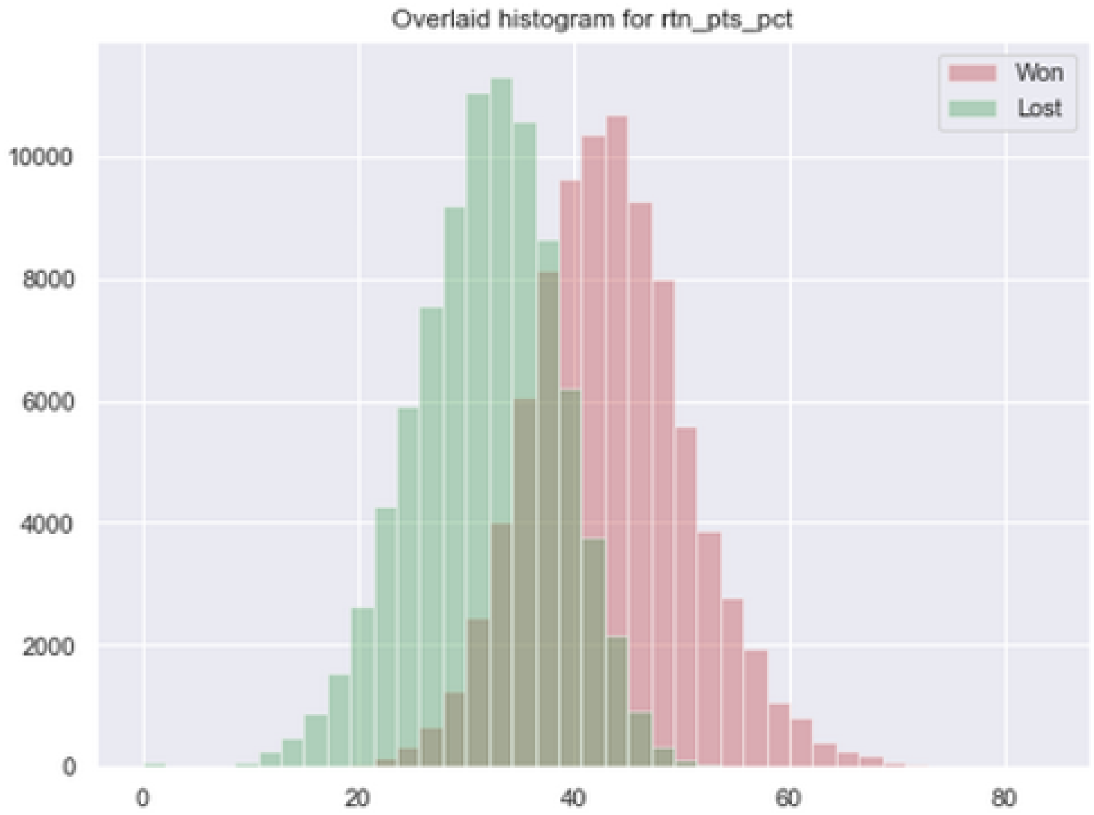
```
Wins Mean:
68.22250043301648
No-Wins Mean:
56.694481221203695
Mean Diff:
11.528019211812783
H0 Diff: 0

p: 0.0
CI: [-0.08612867
0.08654083]
ME:
0.0869216829786314
```

Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS



```
Wins Mean:
43.3193882825373
No-Wins Mean:
31.80259127897296
Mean Diff:
11.516797003564335
H0 Diff: 0

p: 0.0
CI: [-0.08571761
0.08432219]
ME:
0.08365446555536553
```

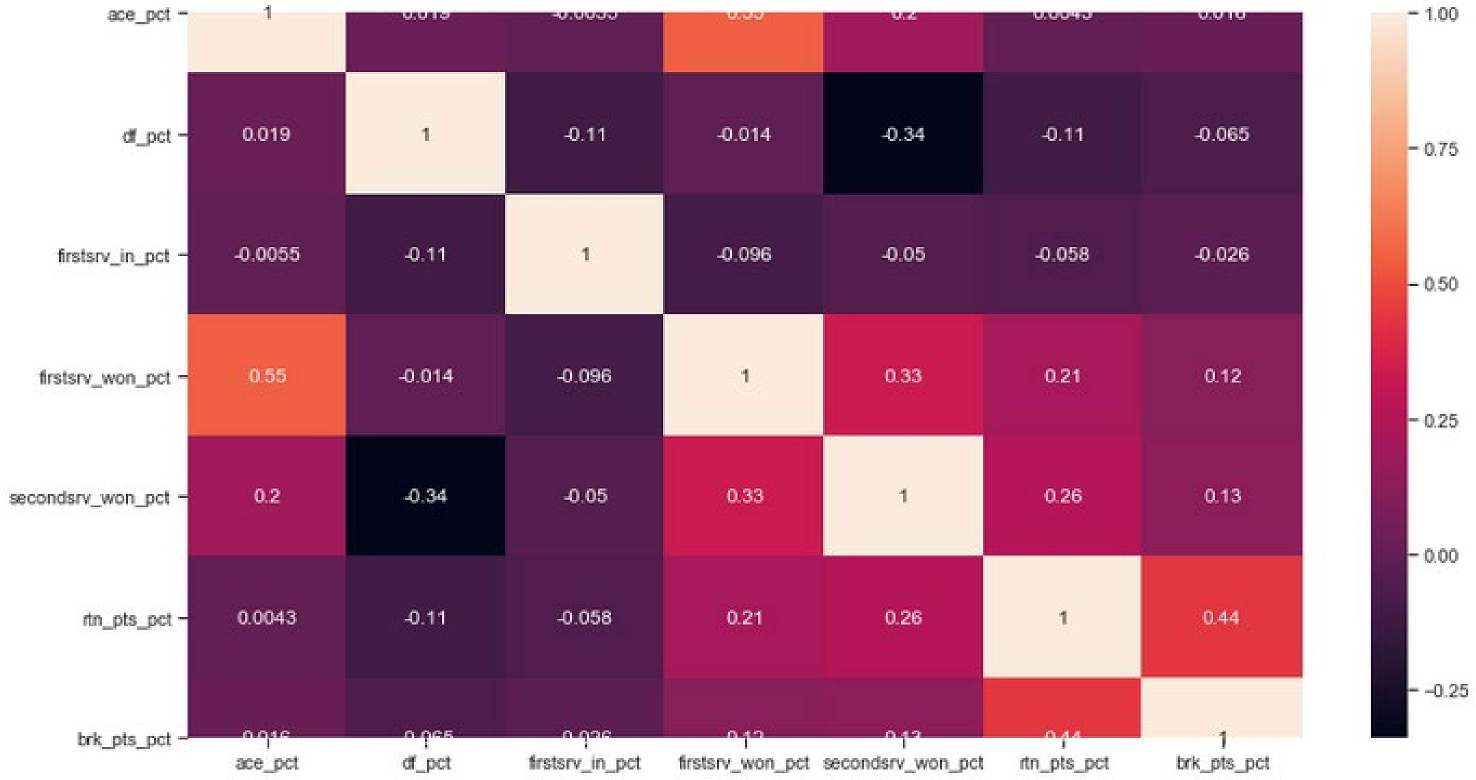
Baseline Features

DISTRIBUTION PLOTS:

WINNERS VS LOSERS

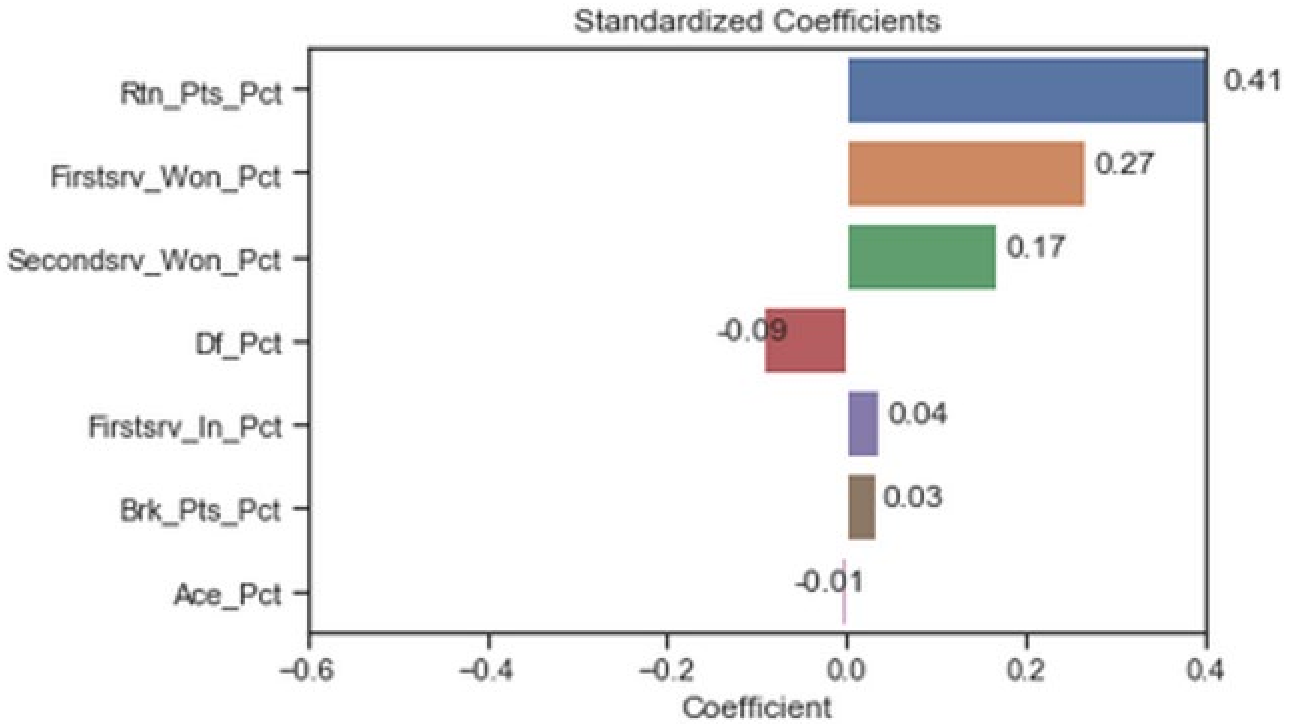
CORRELATION MATRIX

TUNING VARIABLES HEATMAP



Results

LOGISTIC REGRESSION FEATURE IMPORTANCE



Summary

FINDINGS & NEXT STEPS

What Matters

- Win Return Points
- Win Service Points on either serve
- More important to get the serve in than to serve an ace.

Summary

FINDINGS & NEXT STEPS

What Next

- Need more data on specific types of shots and shot placement
- Need more data on specific game scores