

Capstone Project 1: Tennis

Scott Penn

A dark blue diagonal gradient bar that starts from the bottom left corner and extends towards the top right corner, covering the lower half of the slide.

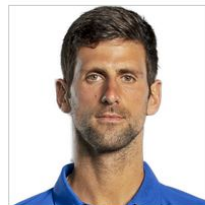
Dataset:

https://github.com/JeffSackmann/tennis_atp

- Contains match data, ranking data, and player data from 1969–2019
- Official statistics are provided by the ATP, but include some gaps and errors.
- Many popular statistics are not present, and must be computed first. (e.g. H2H)

Novak Djokovic VS. Roger Federer | Melbourne 2020

[Previous Page](#)



Novak
Djokovic

Semi-Finals

N. Djokovic



7 6 6

R. Federer

6¹ 4 3

02:18:00



Roger
Federer

Match Stats

YTD Stats

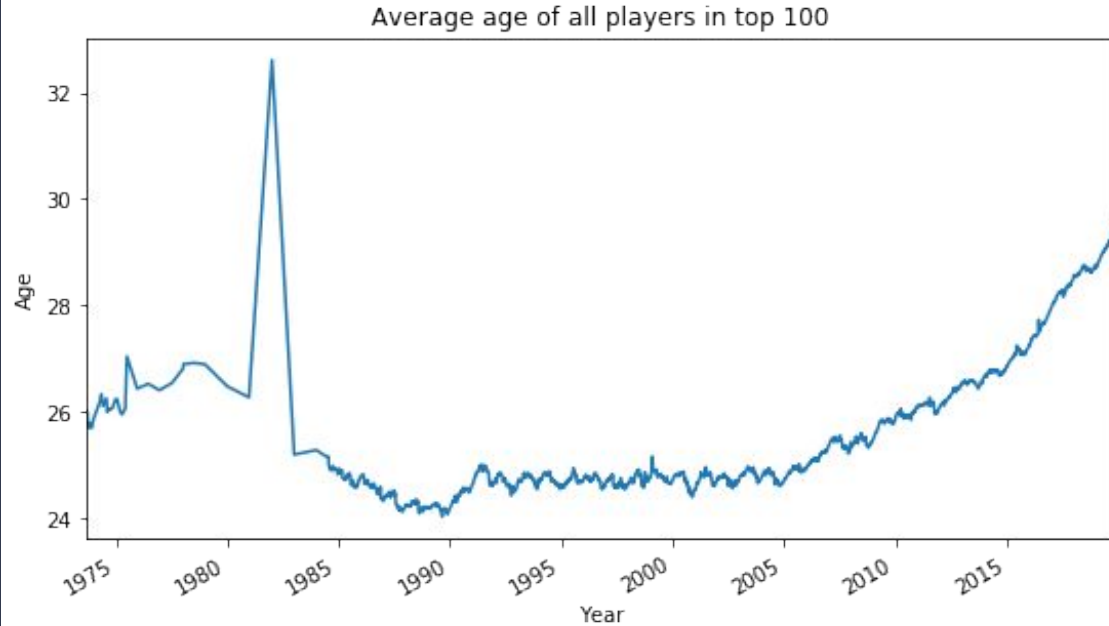
Service Stats	
297	258
11	15
1	3
73% (74/102)	65% (68/104)
73% (54/74)	66% (45/68)
54% (15/28)	42% (15/36)
71% (5/7)	64% (7/11)
16	15
Return Stats	
154	114
34% (23/68)	27% (20/74)
58% (21/36)	46% (13/28)
36% (4/11)	29% (2/7)
15	16

Data Wrangling Part I

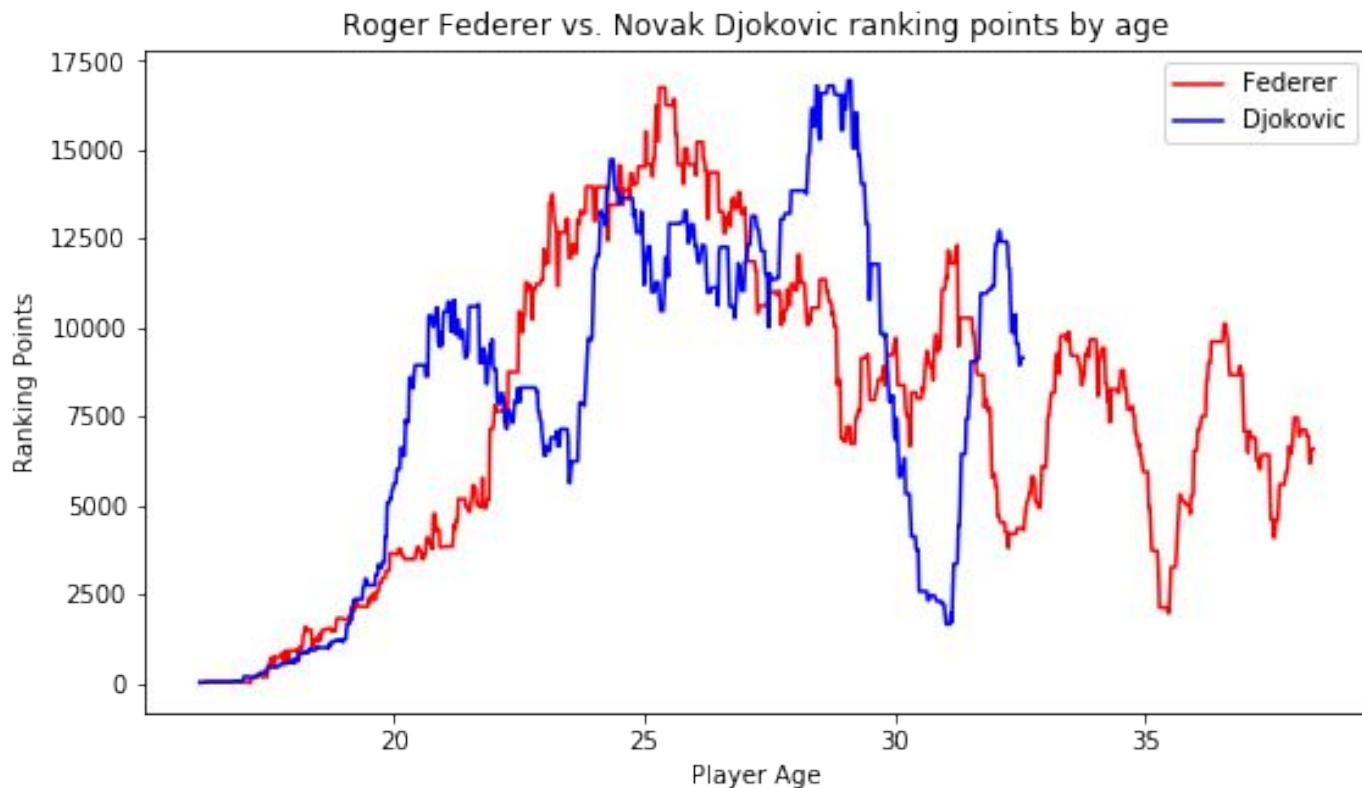
1. Reducing the player set to Top 100 players (from 54405 to 1094)
2. Splitting match data for the winner and loser into separate rows.
3. Computing additional match statistics (e.g. 1st Serve Percentage)

Questions

- What defines a Top 100 player?
- Are the official statistics sufficient for predicting a player's performance?

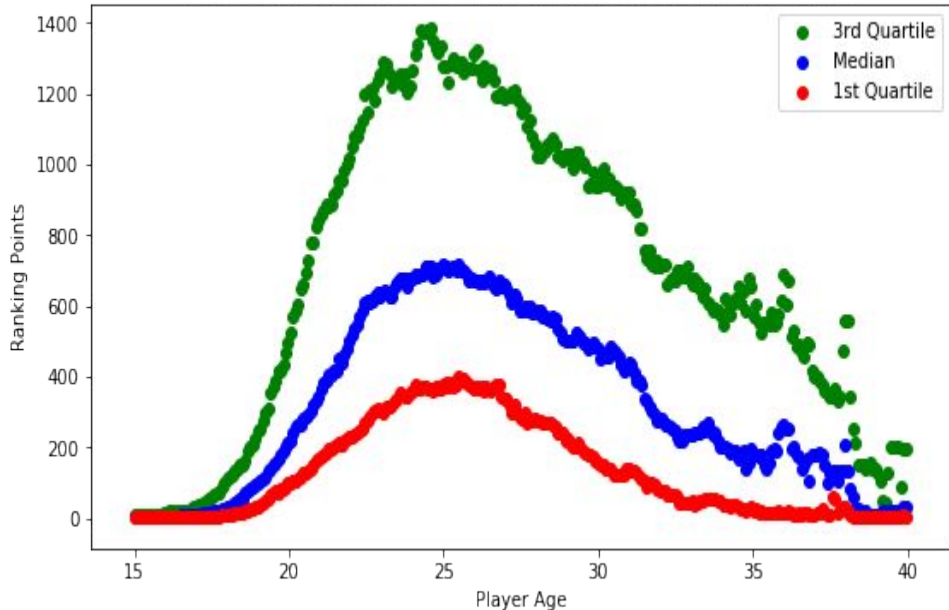


Visualizing Time Series Data

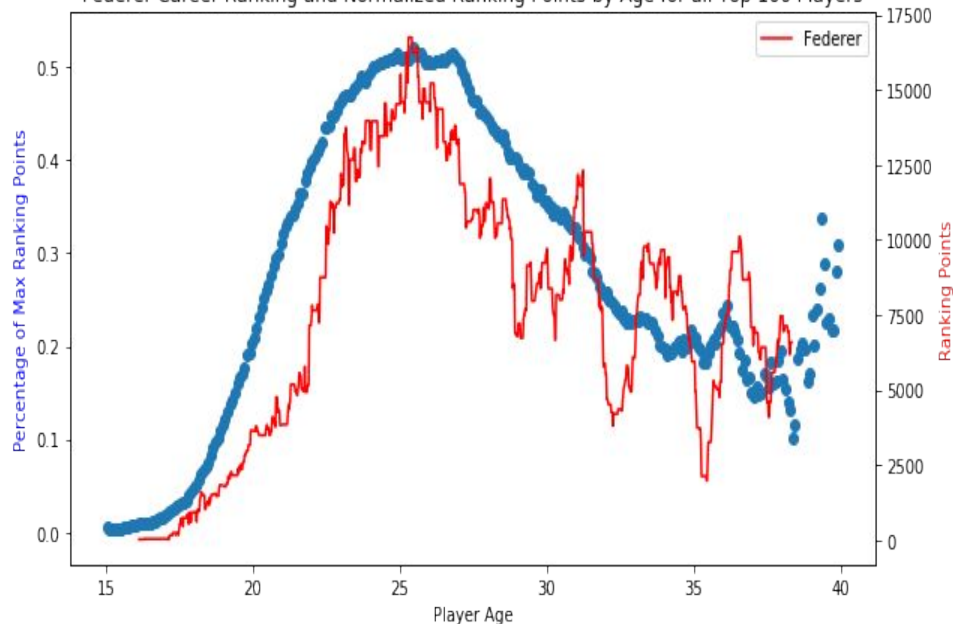


Visualizing Average Career Ranking Data

Q1, Median, and Q3 Ranking Points by Age for all Top 100 Players

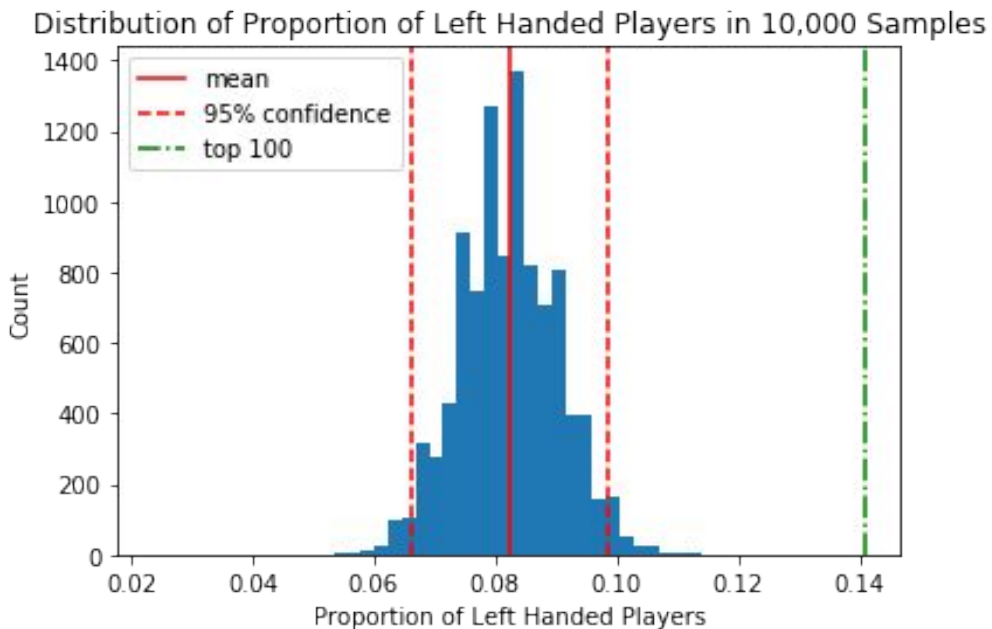


Federer Career Ranking and Normalized Ranking Points by Age for all Top 100 Players



Statistical Analysis

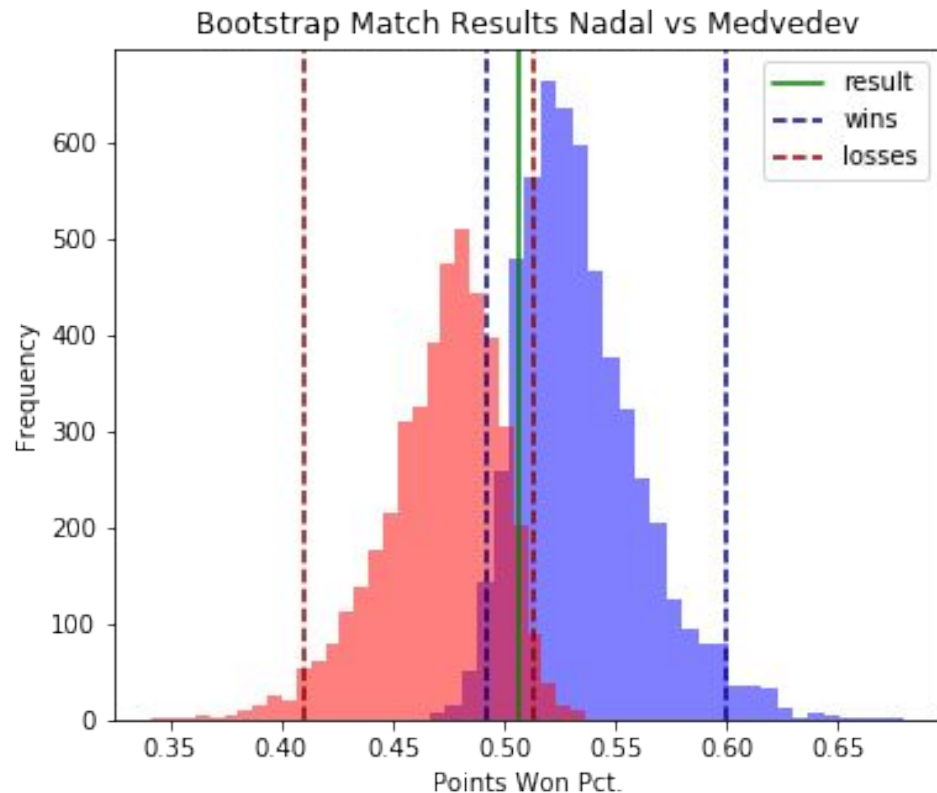
Top 100 players are significantly more likely to be left handed.



Bootstrap Analysis

How likely is a result given the match statistics? There is only one match played, so generating bootstrap samples can provide an answer.

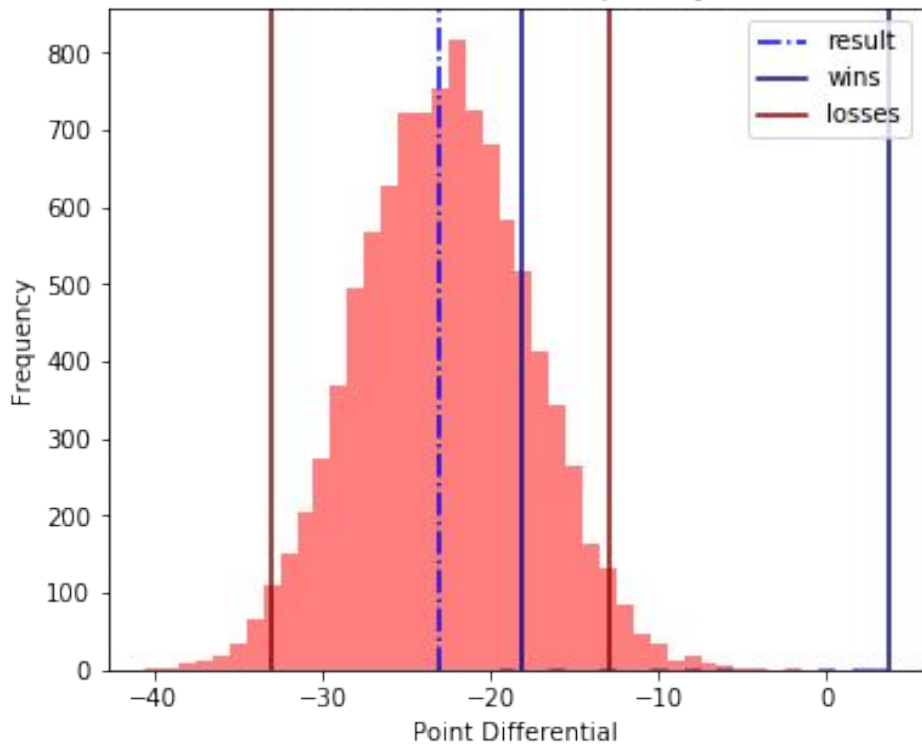
Some matches are lost despite winning more points than the opponent. This overlap can provide uncertainty when predicting the result of a match using regression techniques.



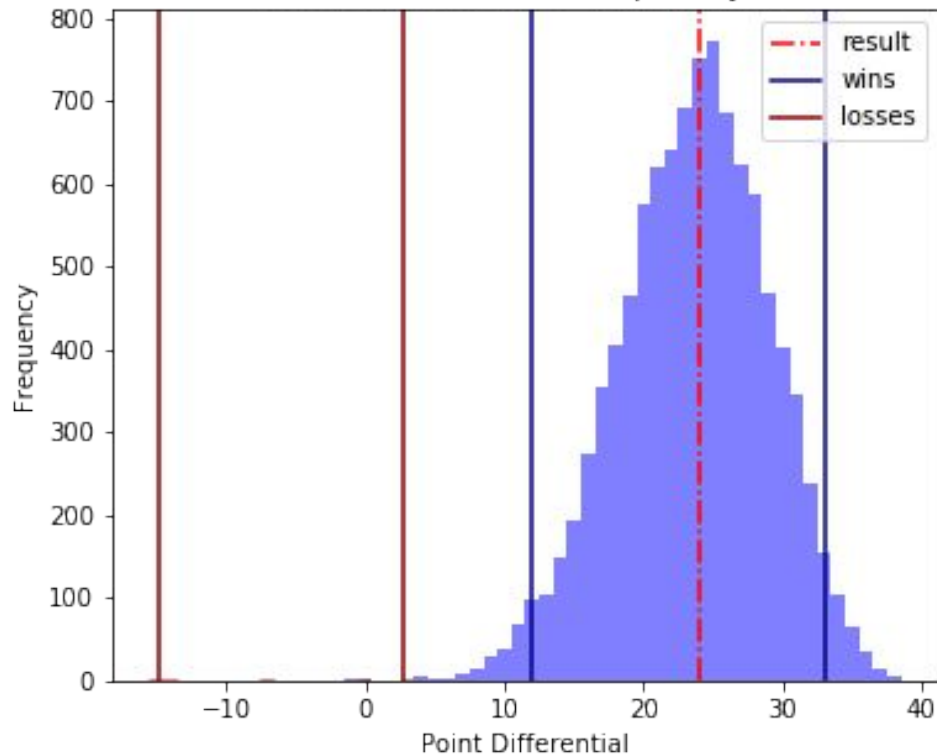
The Luckiest Win and the Unluckiest Loss

Both results fall well outside the 95% confidence interval for a win or loss given the statistics.

Luckiest Win Bootstrap Analysis



Unluckiest Loss Bootstrap Analysis

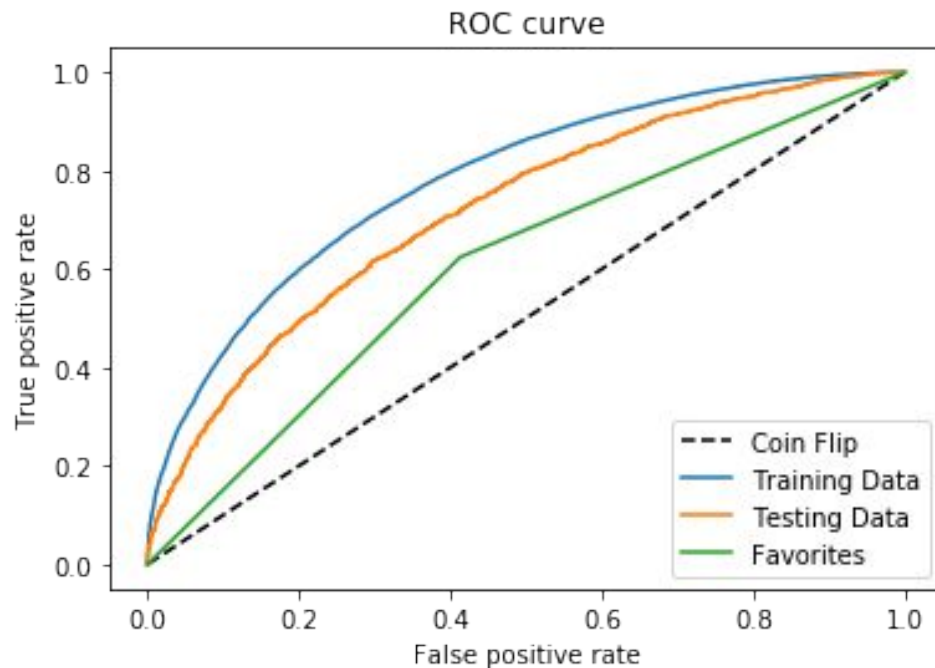


Data Wrangling Part II

1. Determining career statistics using an expanding mean.
2. Determining recent statistics using a rolling mean.
3. Computing additional features such as player head-to-head.

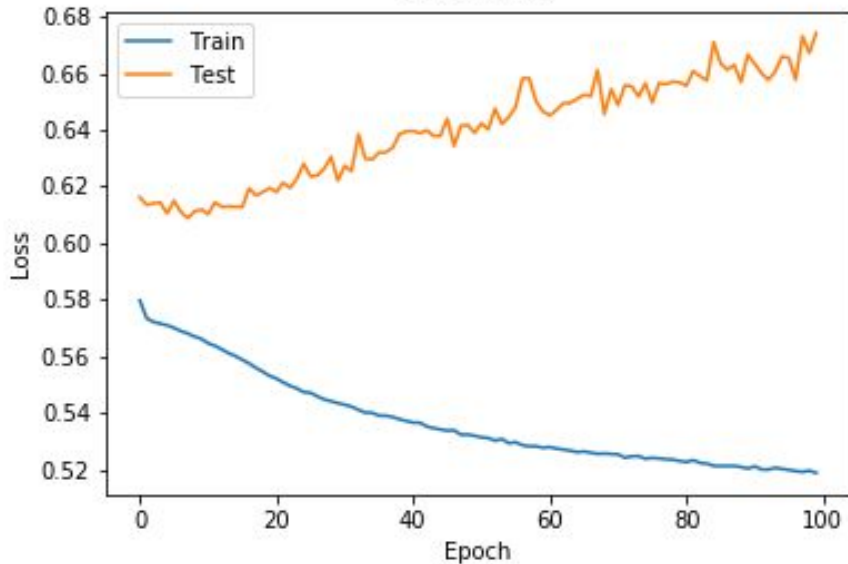
Match Result Predictor

- Using Gradient Boosting to classify a match result.
- Career statistics had the most predictive power.
- ~65% accuracy, 5 percent higher than just picking the favorite.
- Improvement would require more features that more accurately represent player performance.

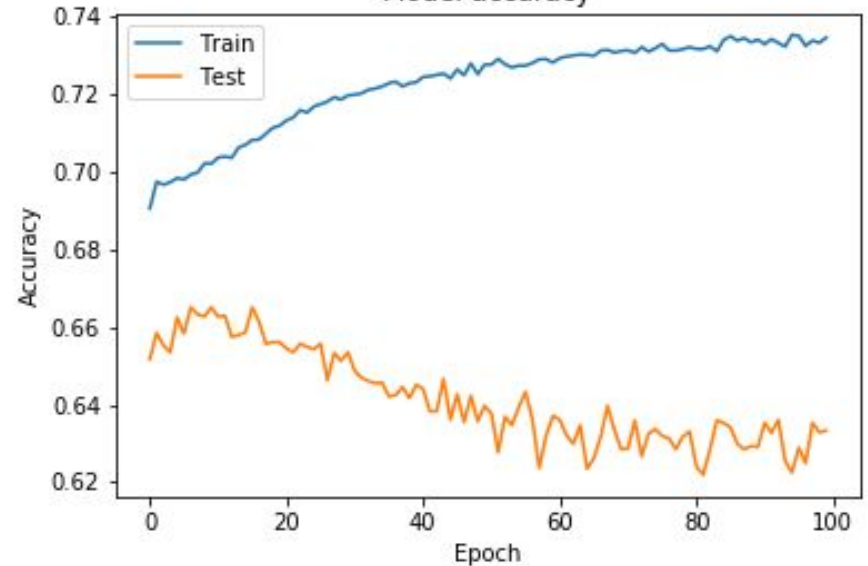


Bonus: Deep Learning

Model loss

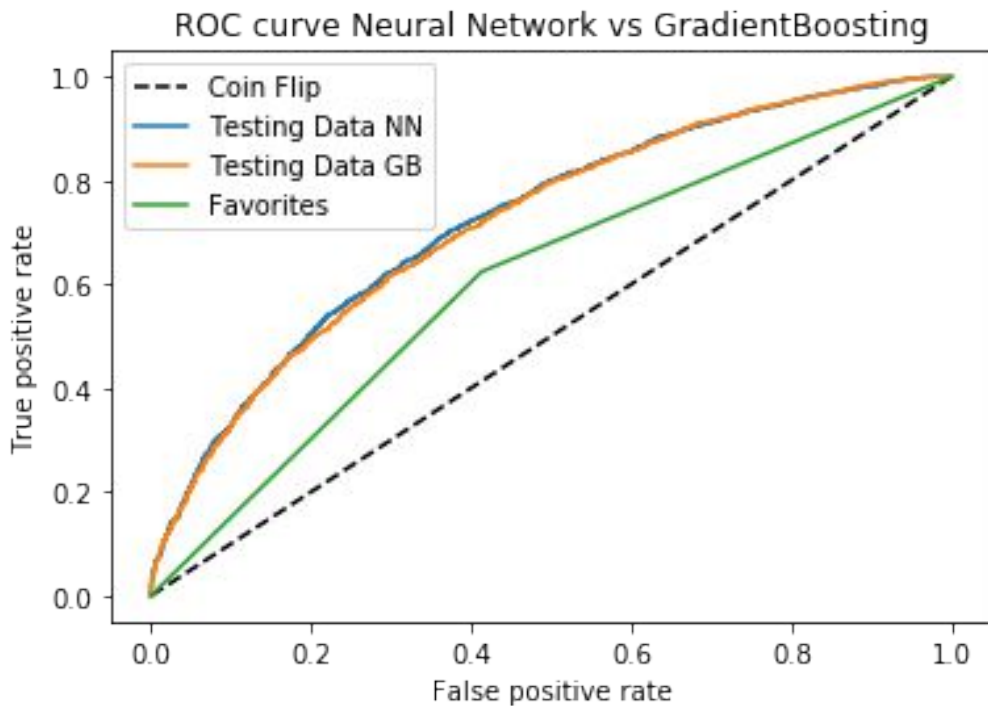


Model accuracy



Results

- 66.5% accuracy on test data.
- Best performance with only one hidden layer containing 100 nodes.
- Model overfits quickly, so Early Stopping is necessary to reduce training time.
- Upsets make it difficult to generalize.



Conclusions

- Many players follow a similar career path.
- Current pro tennis players are remaining competitive for longer.
- Official statistics have a limited use for predicting player success.
- Better performance based statistics could help players improve.
- Tennis lags behind other sports such as baseball in both use and availability of advanced analytics.