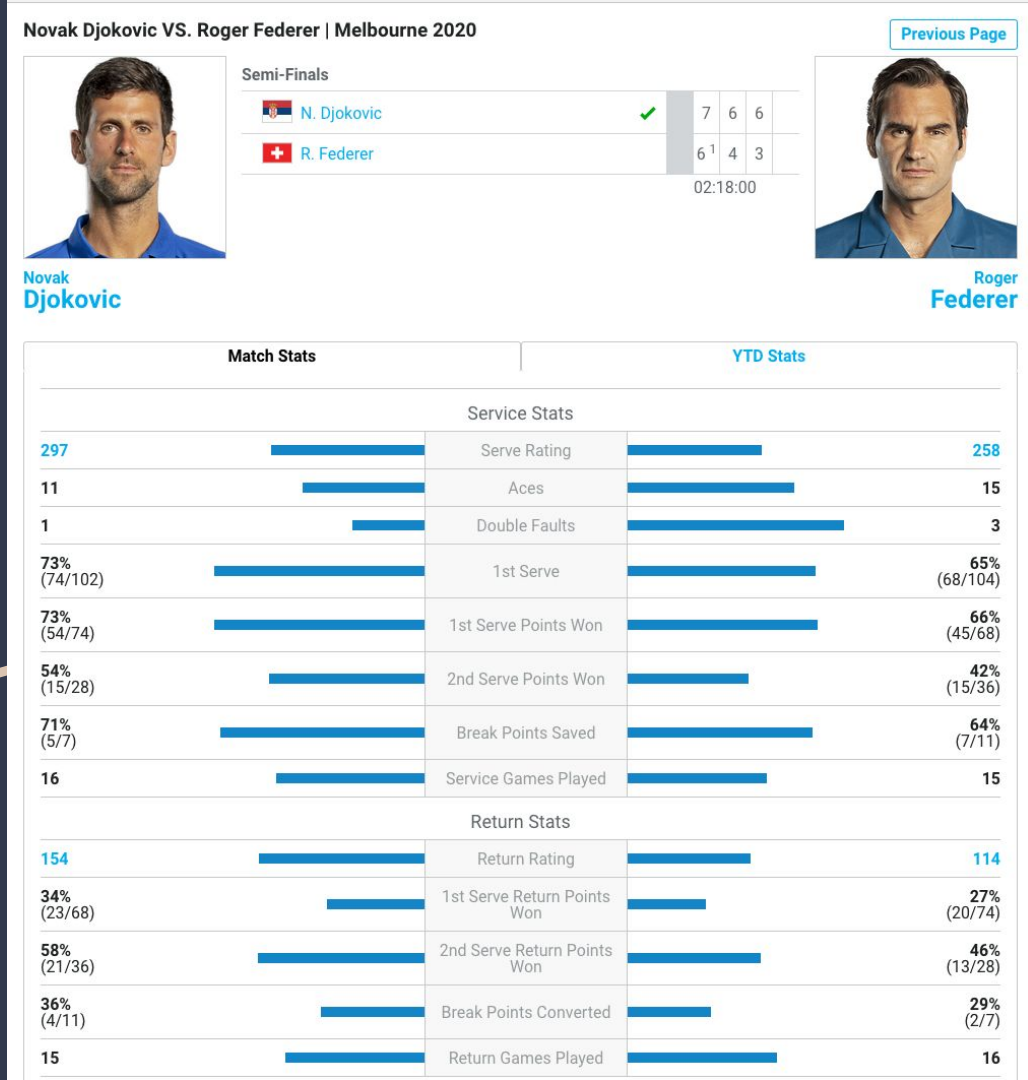
Capstone Project 1: Tennis

Scott Penn

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)



Data Wrangling Part I

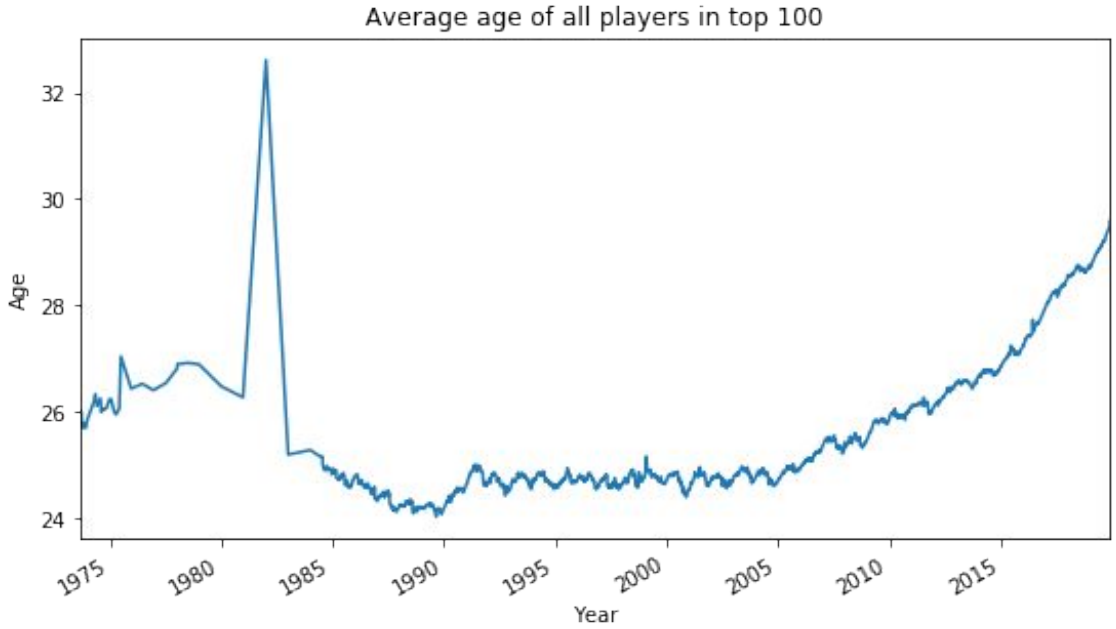
1. Reducing the player set to Top 100 players (from 54405 to 1094)

2. Splitting match data for the winner and loser into separate rows.

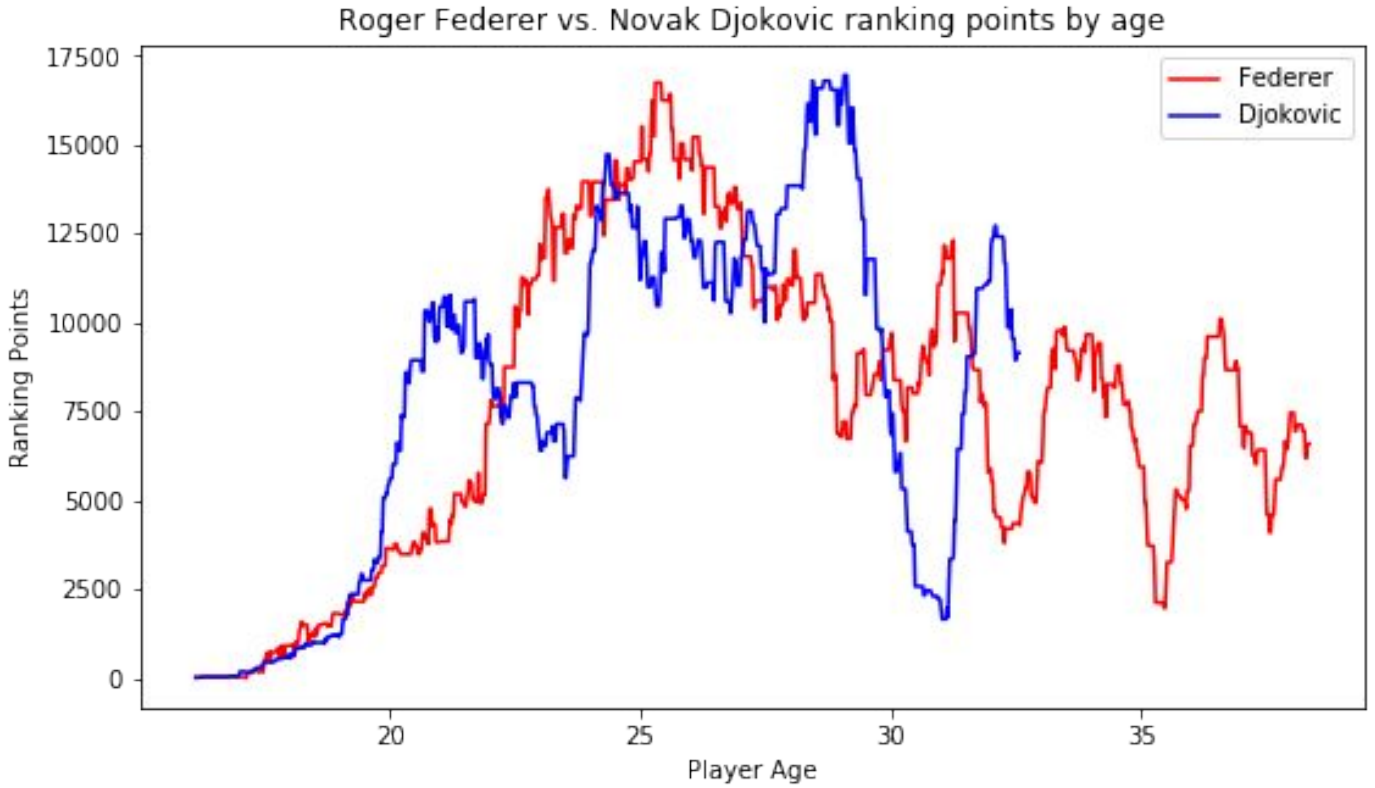
3. Computing additional match statistics (e.g. 1st Serve Percentage)

Questions

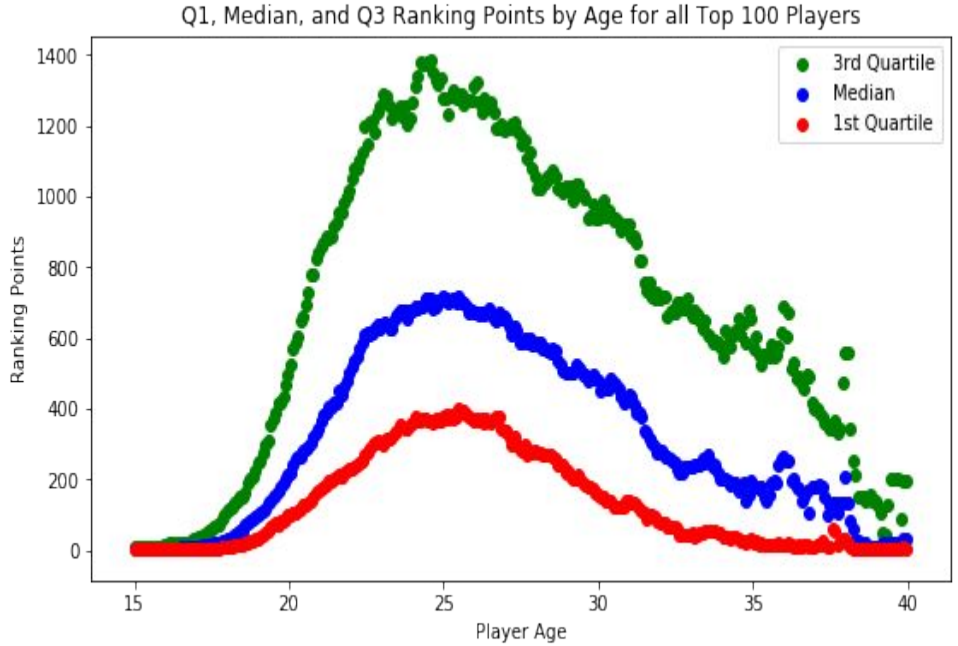
- What defines a Top 100 player?
- Are the official statistics sufficient for predicting a player's performance?

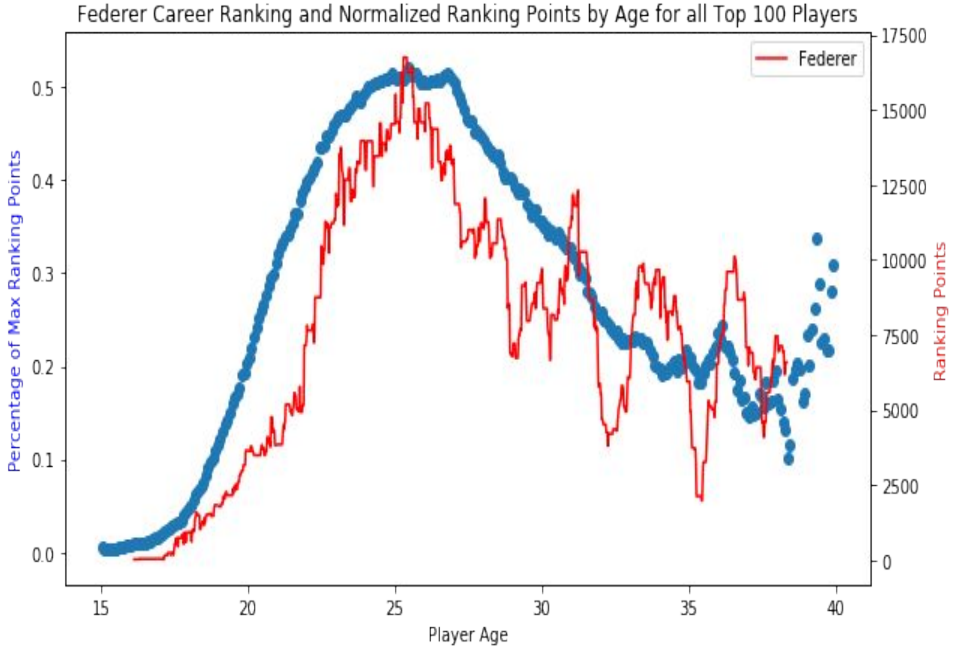


Visualizing Time Series Data



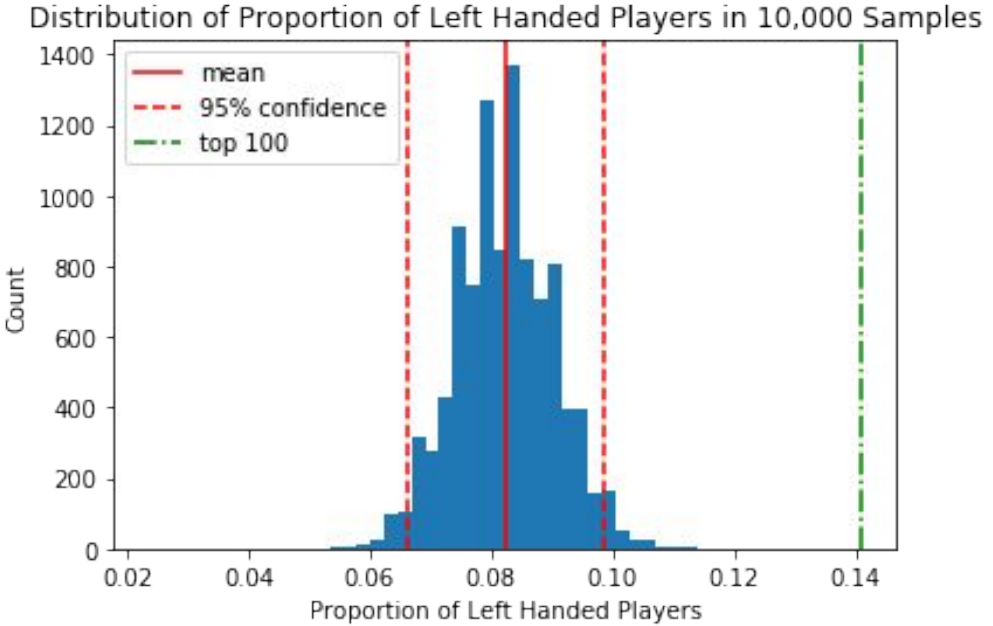
Visualizing Average Career Ranking Data





Statistical Analysis

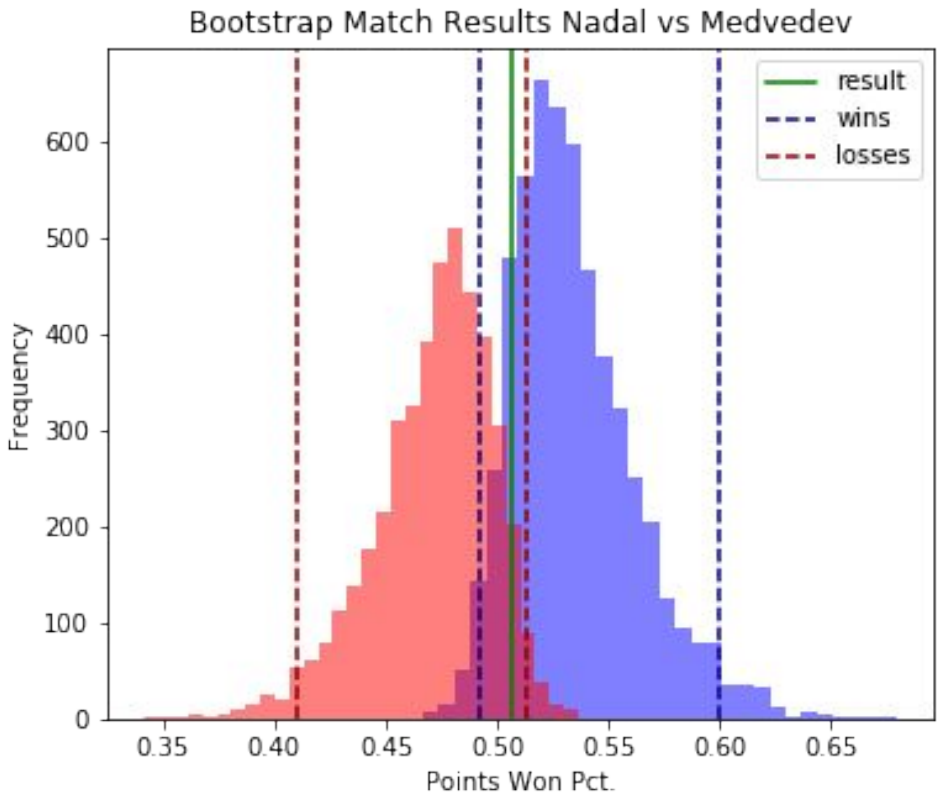
Top 100 players are significantly more likely to be left handed.



Bootstrap Analysis

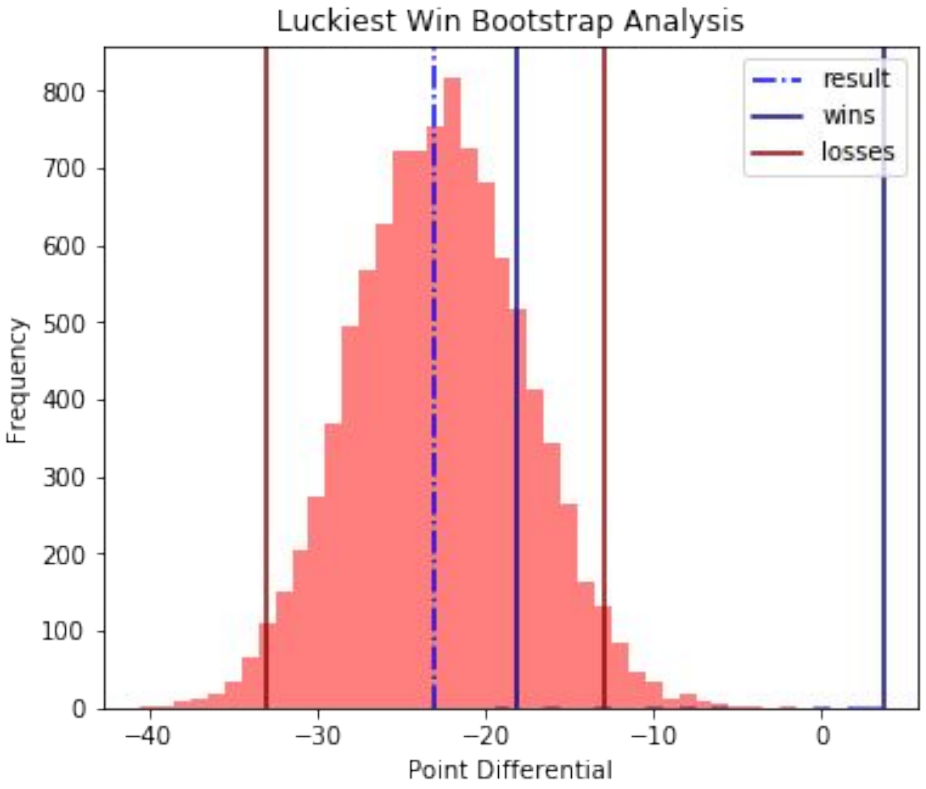
How likely is a result given the match statistics? There is only one match played, so generating bootstrap samples can provide an answer.

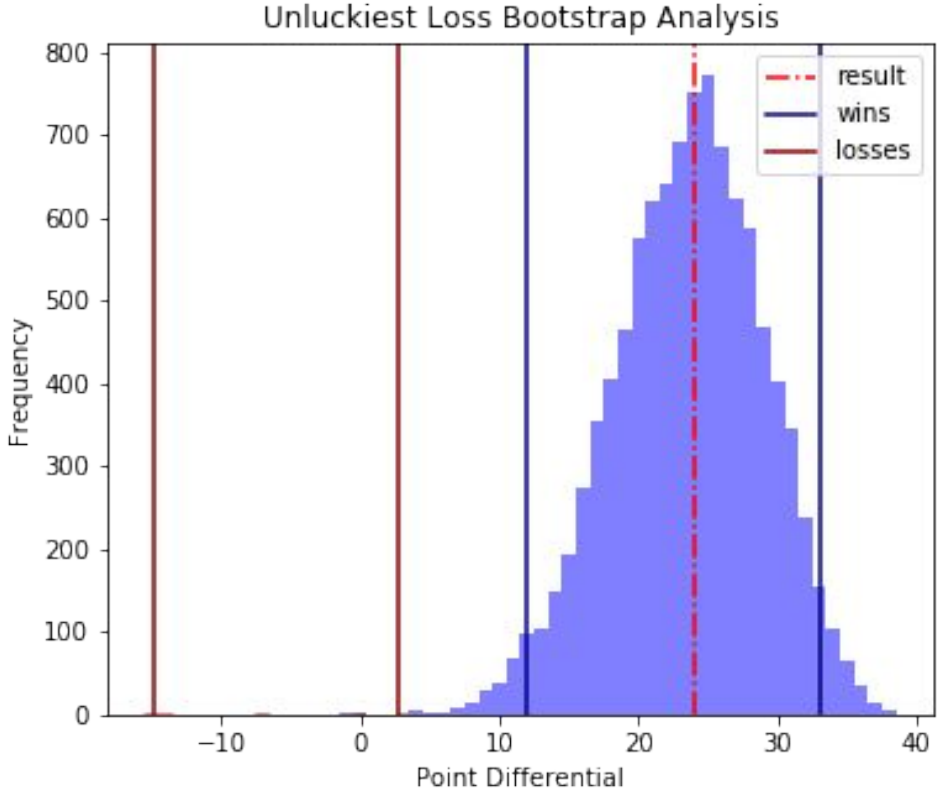
Some matches are lost despite winning more points than the opponent. This overlap can provide uncertainty when predicting the result of a match using regression techniques.



The Luckiest Win and the Unluckiest Loss

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





Data Wrangling Part II

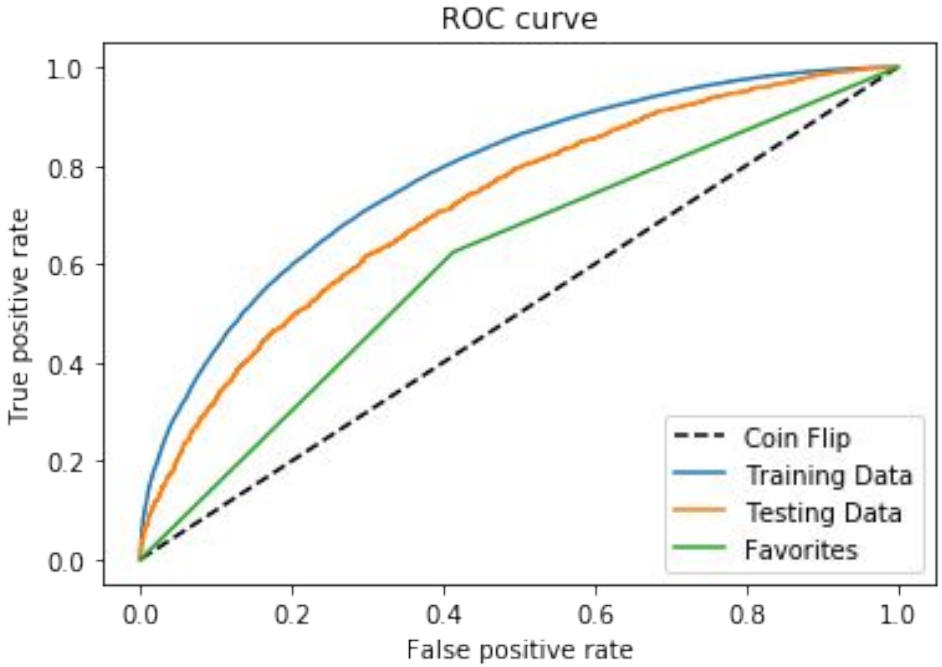
1. Determining career statistics using an expanding mean.

2. Determining recent statistics using a rolling mean.

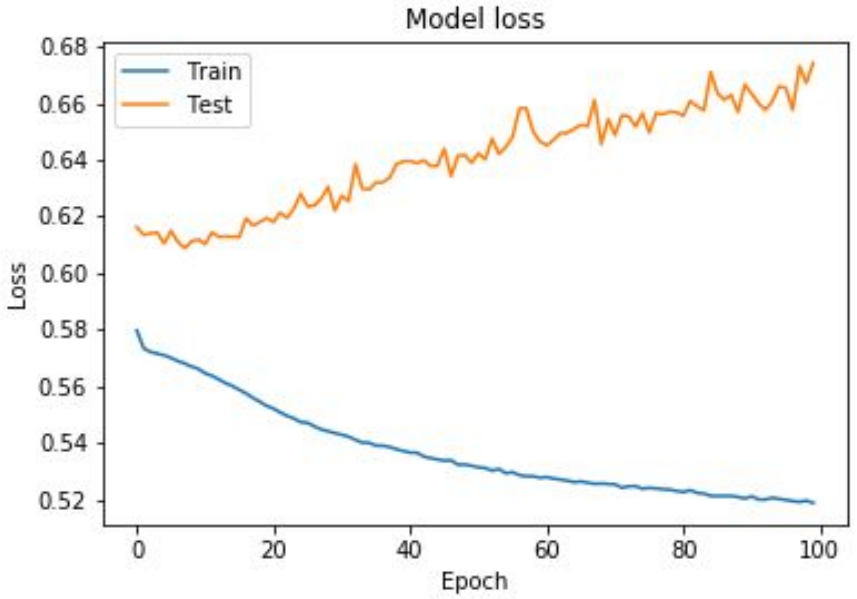
3. Computing additional features such as player head-to-head.

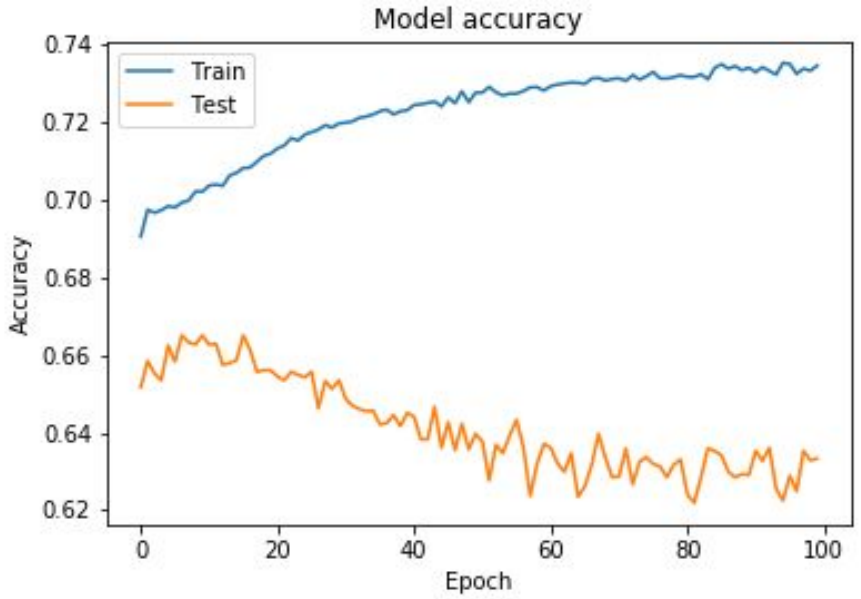
Match Result Predictor

- 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.



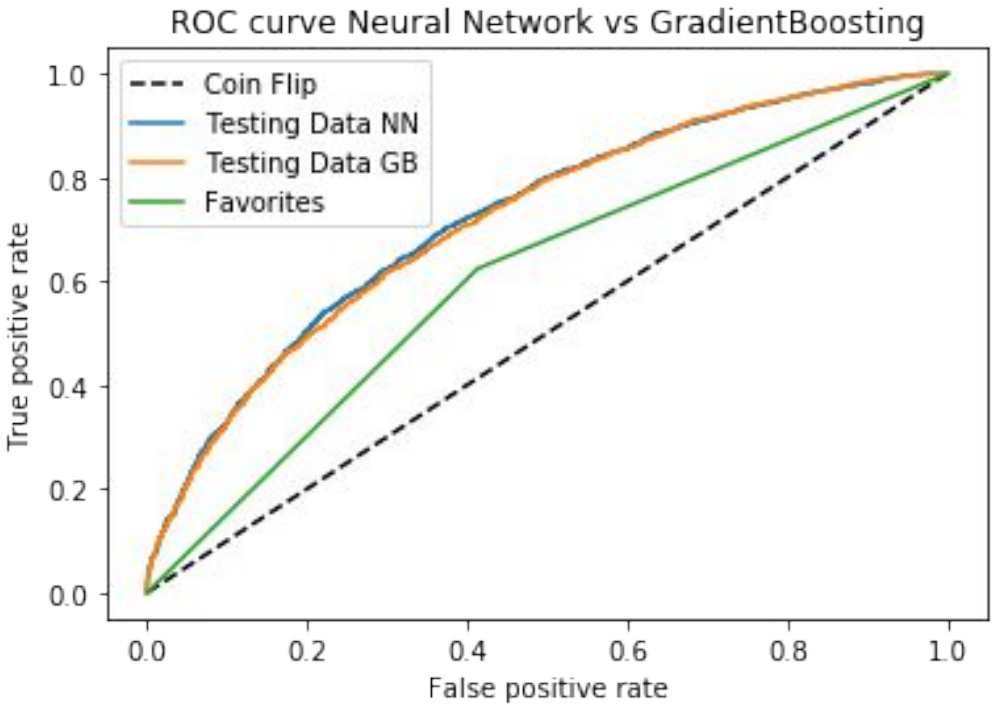
Bonus: Deep Learning





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.