Predicting Tennis Player Performance

From historical match data*.

October 9, 2020
Barrett Williams

*Men's singles matches and rankings.
Would like to run a similar analysis on women's data, though there is less of it.

Interesting questions to ask:

Which players are on the rise?

Who will unexpectedly perform super well next season?

What are the other, unexpected indicators of ranking, wins, or lifetime prize money?

How does performance in one year serve as a leading indicator of performance in the next year?

Which players had bad or good luck this season, given other stats?

Hypothetical presentation audience:

ESPN is looking to predict winners in advance of matches, to highlight anomalies, or identify underdogs with promise.

Measures of overall player performance:

Win-loss ratio:

- All time
- Per year

Prize money won

- All time
- Per year

Rank (integer values)

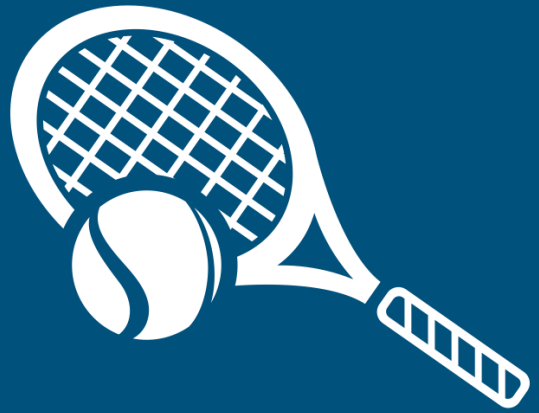
Demonstrated useful stats to use in our model:

More useful:

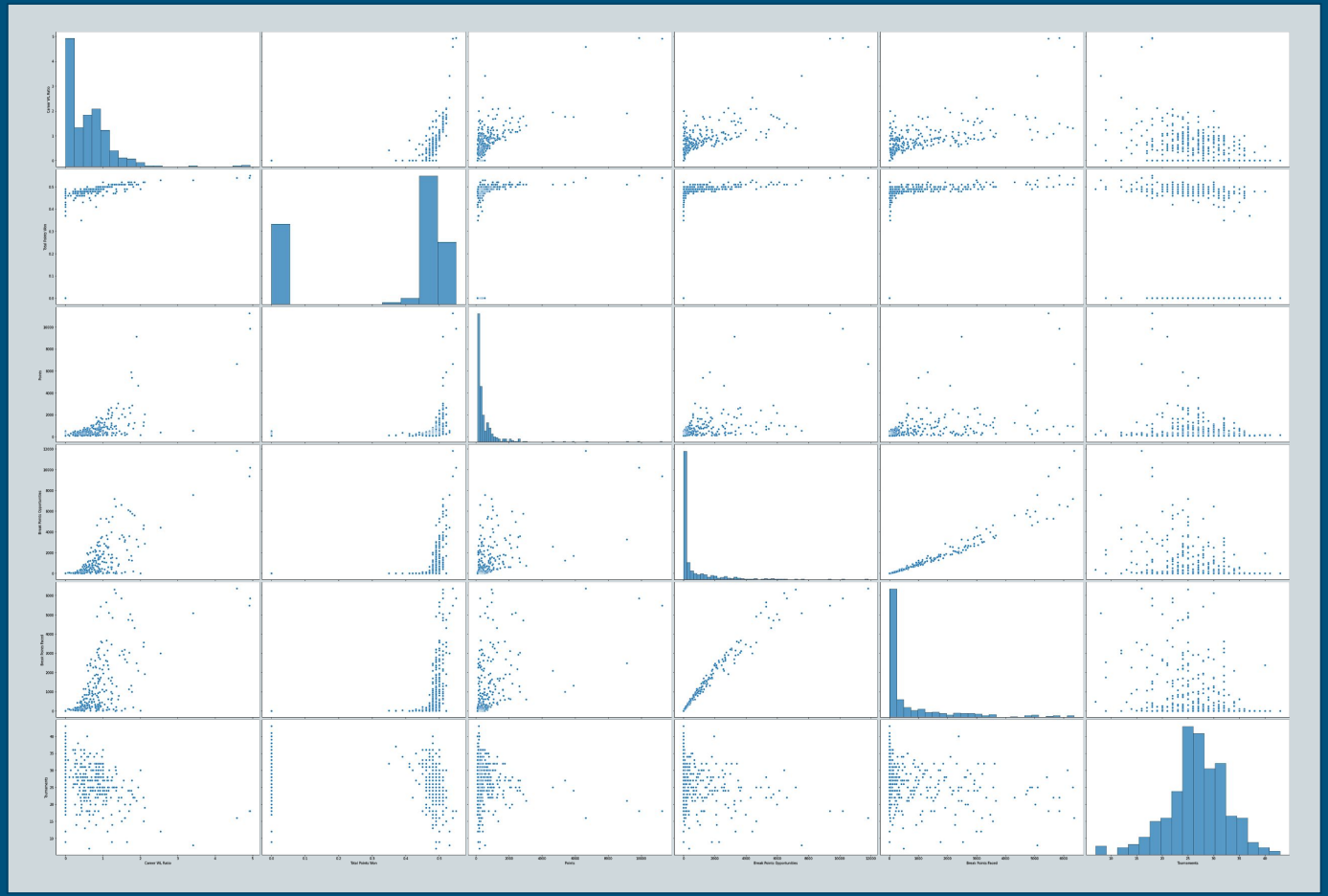
- Total points won
- Break points won

Less useful:

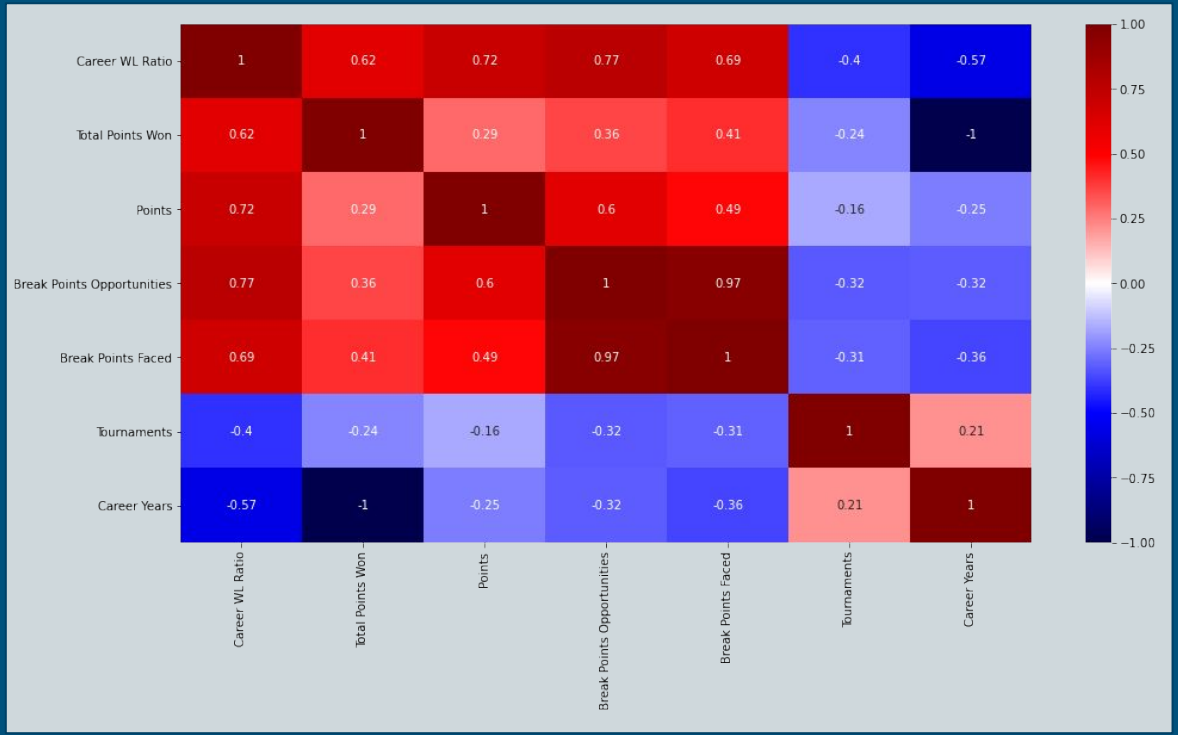
- Tournaments played (negative correlation!)
- Double faults



Pairplots



Correlation heatmap (post feature reduction)

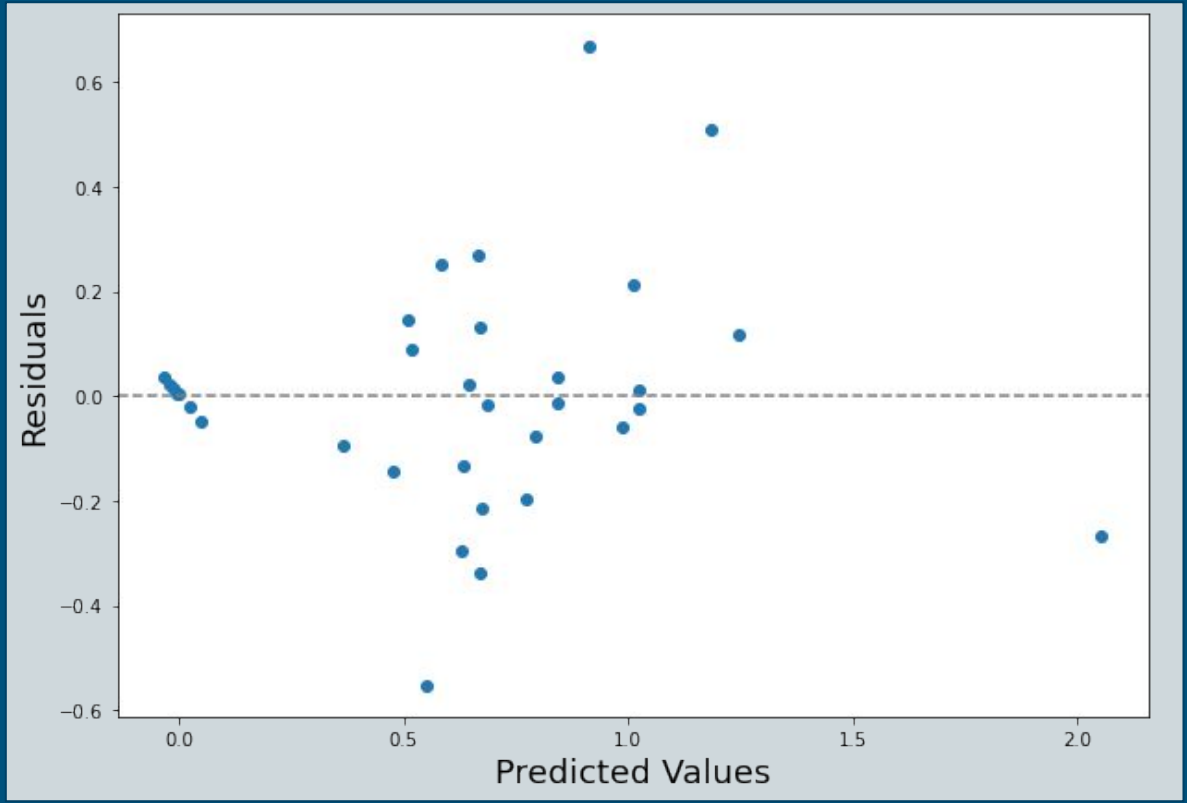


How well do our models work?

100 players			**262 players**	(4-way fold)	
Model	**R^2**	**RMSE**	**Model**	**R^2**	**RMSE**
Std reg	0.94	0.17	*Standard reg*	0.847 ± 0.081	0.214 ± 0.013
Lasso	0.94	0.26	*LassoCV*	0.810 ± 0.099	0.233 ± 0.012
Ridge	0.98	0.17	*RidgeCV*	0.842 ± 0.083	0.215 ± 0.014
			Feat-reduced standard reg	0.855 ± 0.048	0.212 ± 0.015

Note: polynomial transformations do not provide any measurable improvement.
For future investigation, a Box Cox transformation may prove useful.

Residual plot for final model



Interesting conclusions:

- Playing in more tournaments means a greater risk of losing,
  so tournament count is inversely correlated with win-loss ratio
- Age has little bearing on win-loss ratio, except perhaps, at the very top of the range
- Double faults had little impact: at the pro level they can be managed
- Break points have greater impact on win-loss ratio than aces

Methodology:

- Web scraping with Selenium and Beautiful Soup
- Linear regression
- Regularization
- Test-train split, with holdout for validation
- Lasso and Ridge regressions
  - Now with 4-fold cross-validation, reduced from 5 due to small dataset size
- Dimensionality reduction
- Model simplification

Appendix

Feature lists:

Full dataset:

- Rank, Age, Points, Tournaments, Aces
  Double Faults (int)
- 1st Serve, 1st Serve Points Won, 2nd Serve Points Won, (float)
- Break Points Faced, Break Points Opportunities (int)
- (float)
- Break Points Saved, Service Games Won, Total Service Points Won, 1st Serve Return Points Won, 2nd Serve Return Points Won (float)
- Service Games Played, Return Games Played (int)
- Break Points Converted, Return Games Won
  Return Points Won, Total Points Won (float
- Year Turned Pro (int)
- Career WL Ratio (float)
  Lifetime Prize (int)

Twice reduced:

- 'Total Points Won', (%)
- 'Points',
- 'Break Points Opportunities',
- 'Break Points Faced',
- 'Tournaments'

Challenges:

- CAPTCHA inhibited scraping at random intervals
  - HTML DOM for ATP Tour website was inconsistent
- Higher R^2 values on smaller sample size of 100 players
- Although 400 players scraped, only 262 had detailed stats
- For next time:
  - Make scraping process more robust
  - Break out error calculation and charting into separate helper functions
  - Predict rank, even though it is integer
  - Look at intra-year data to predict break-out players
  - Logarithmic transforms to account for high-performers skewing the data