Capstone Project 1 – ATP Tennis Match Statistics

Project Proposal

Sports statistics have been around for a long time. In recent years, it seems like advances in technology has increased the number of statistics that have been published in virtually every major professional sport. The problem then, isn't the availability of sports statistics, the problem is understanding what those statistics mean and whether sports participants can improve aspects of their performance, thereby increasing their chances of winning.

If you've ever watched a professional tennis match on television, you've no doubt seen various match statistics displayed on the screen and discussed by the commentators—before, during and after a match. They may comment on the differences between key statistics for each player and try to tie those statistics to why the players are winning or losing. Most often, however, there is no clear "winner" in each category, much less any consistency from one match to the next on which category(ies) matters most. In fact, it is not uncommon for a player to perform well in one category in one match and win, and then perform as well or better in that same category in the next match and lose.

After a match, players and their coaches review match statistics to determine what their next practice sessions need to focus on. Unfortunately, absence of any meaningful analysis on those match statistics, practice time could result in spending time trying to improve a particular skill that won't necessarily translate into more match wins. It is easy to say, "Your opponent had fewer unforced errors on her backhand side, so let's work on your backhand." It may be true that the opponent had fewer unforced errors, but it's difficult to say whether reducing the number of unforced backhand errors will result in a win the next time with any degree of certainty. Or a coach will say, "You missed all of your drop shots; let's work on that." Yet the player only missed three drop shots in a match with over 500 hundred total shots, so even if the player made all three drop shots, it's unlikely the match outcome would be any different.

So the challenge is to develop a machine learning model of match statistics that will result in meaningful information that will help a player her team spend time practicing what will be most likely to improve match outcomes.

Wikipedia describes sports analytics as a "collection of relevant, historical, statistics that when properly applied can provide a competitive advantage to a team or individual." That's what this project is all about.

Goals

1. Identify which feathers (match statistics) predict match winners.
2. Explore relationship between match statistics and winning matches.
3. Create machine learning model to predict match winners.

The Data

The data for the project came from the ATP World Tour Website as distributed on datahub.io. The data contains ATP tournaments, match scores, match stats, rankings and players overview. The data is scraped from the ATP's website using python scripts written by Kevin Lin.

The dataset has 53 CSV files broken into 5 different categories: tournaments, match scores, match stats, player rankings and player overviews. Tournament and match score files go back to 1877, while the rankings data goes back to 1973. The match stats, however, only go back to 1991. The focus of this project, therefore, will be to analyze all statistics from 1991 to 2015 since the match stats are such a critical component of the analysis.

There were 95,359 match scores in 2,054 tournaments included in the original data meeting the criteria above.

Data Wrangling

Summary Files. The results of the data scraping produced five separate files. For analysis purposes and faster processing, these files were merged into one comprehensive dataframe.

Performance Features for Exploratory Data Analysis. Due to the unique nature of tennis scoring, the variety of match lengths (some matches end relatively quickly with few statistics in all categories, while some can last five full sets), and the large number of statistics, a few additional performance metrics were added. This addition will allow for an easier EDA phase later on. The additional metrics are as follows:

- Rank differential – a ratio of the difference between opposing player rankings
- Percentage of aces
- Percentage of double faults
- Percentage of first serves in
- Percentage of break points won
- Percentage of return points won, and
- Percentage of total points won.

Duplicates, Unnecessary, Incomplete and/or Missing Values.

In merging the different datasets, duplicate and/or irrelevant columns were removed. Also, several matches included in the overall dataset were from matches that either were not played or were due to retirement. These matches were removed.

In today's tennis world, there is a much more robust set of match statistics that are being collected. It would be interesting to have access to this data for this project. Shot related data, such as cross court , down-the-line, drop shot, overheads, etc. broken down into winners, forced errors and unforced errors, length of rallies, serve and return placement, etc., would no doubt provide some additional insights not captured in this project.

Tournament and Match Dates

Unfortunately, the match statistics for each match did not include the specific data of the match. As such, all matches were assigned the start date of the tournament in which they were played. The dates

were then broken into year, month and day for data exploration purposes, as well as providing additional options for training and testing the machine learning model.

Two Final Datasets

The data was ultimately put into two main datasets for Exploratory Data Analysis and Machine Learning: 1) A match statistics dataset that allows for analysis and comparison between match statistics in general, and 2) a "results" dataset that separates winner players' statistics from loser players' statistics and then combines them into one dataset to allow for analysis on match statistics differences between match winners and losers.

Exploratory Data Analysis (EDA)

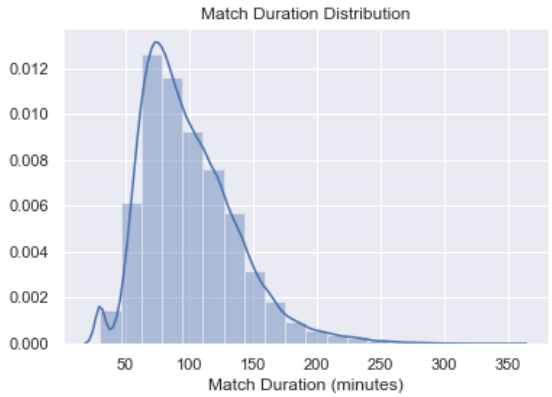
Exploring the data will be split into two stages. The first will include an analysis of statistics within matches. The second will be an analysis of winners' and losers' statistics.

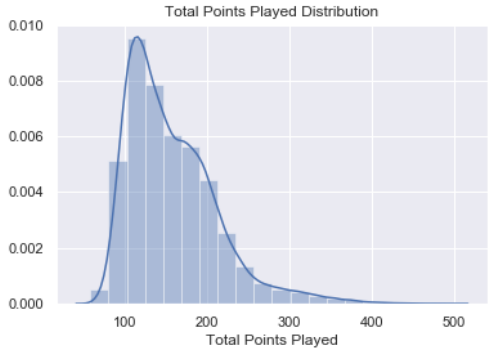
Match Statistics

After cleaning and wrangling the data, there are match statistics from 88,206 matches. Here are some interesting findings about the match statistics:

- The average match duration is 99.98 minutes.
- On average, the matches had a total of 157.04 points played.
- On average, matches had a total of 24.63 games played.

Here are plots of the distributions of match duration and total points played:





As expected, the distributions for match duration and total points played are very similar.

Even more interesting are the following:

Percent of Times the Match Winner Performed Better in Each Category

Category	% Winner Performed Better
Percentage of Total Points Won	94.3
Percentage of Service Aces	63.9
Percentage of Double Faults	58.8
Percentage of Service Points Won	92.7
Percentage of Return Points Won	92.6
Percentage of Break Points Won	70.7
Winner Ranked Higher	65.5

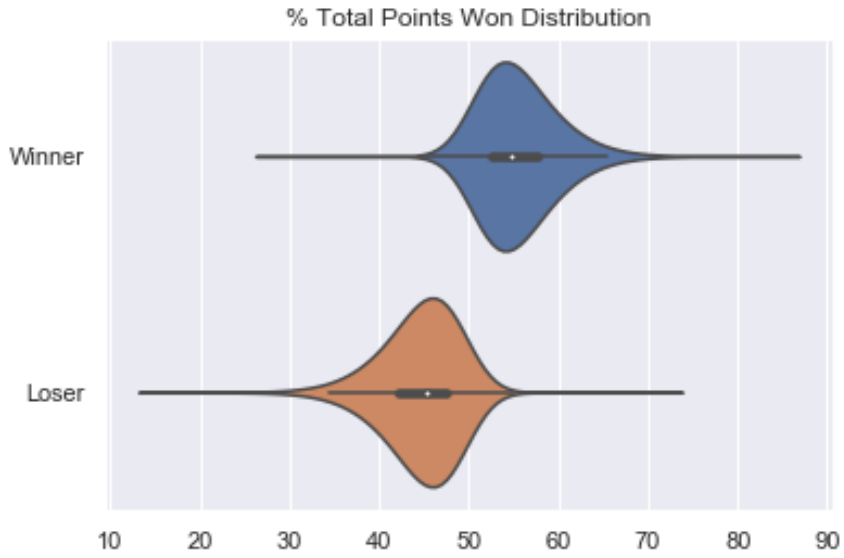
What is so interesting about these statistics is that although the winner had better statistics in all of the categories, primary Percentage of Total Points Won, Percentage of Service Points Won and Percentage of Return Points Won, the margin of victory was an average of 10.9 percent for Percentage of Total Points Won, or 14.3 points per match. With an average of 24.6 games played per match, that's only 0.58 points per game.

Match Winner and Loser Statistics

Here are the plots and the related statistics for each of the categories listed above, except for player rankings.

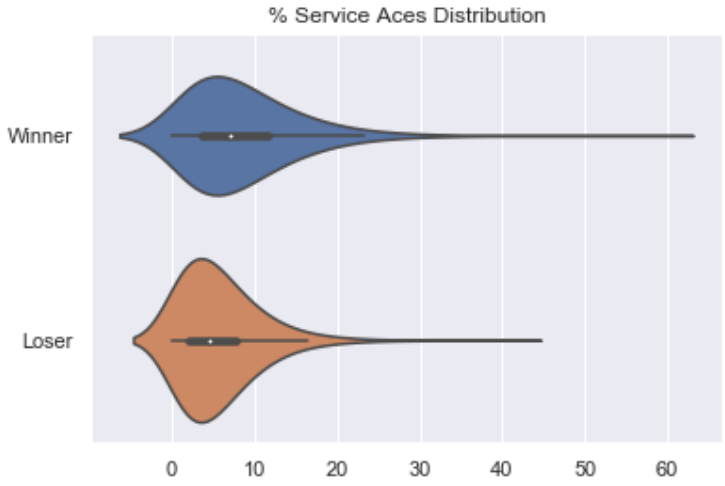
Percent Total Points Won

	count	mean	std	min	25%	50%	75%	max
Winner	88206.0	55.467523	4.115671	30.303030	52.525253	54.819277	57.731959	82.758621
Loser	88206.0	44.532477	4.115671	17.241379	42.268041	45.180723	47.474747	69.696970



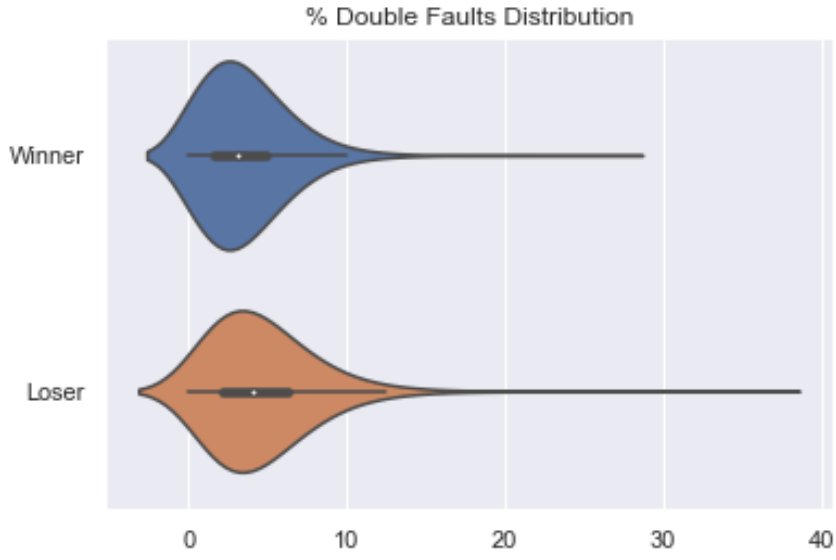
Percent Service Aces

	count	mean	std	min	25%	50%	75%	max
Winner	88206.0	8.340684	6.336654	0.0	3.676471	6.976744	11.538462	56.756757
Loser	88206.0	5.599402	4.661899	0.0	2.127660	4.587156	7.865169	40.000000



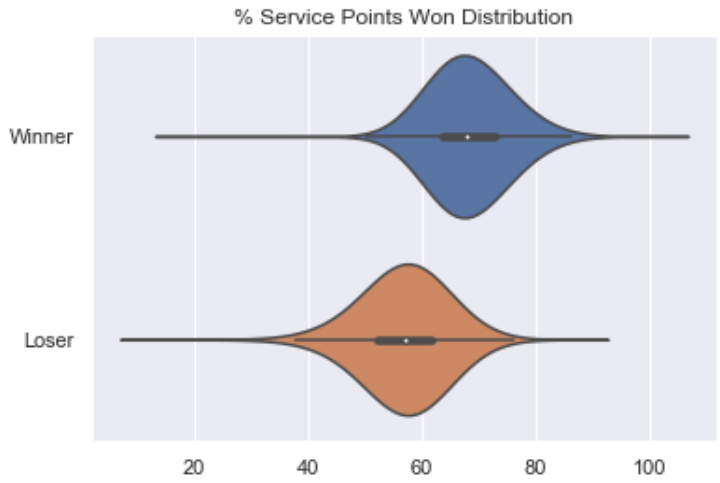
Percent Serve Double Faults

	count	mean	std	min	25%	50%	75%	max
Winner	88206.0	3.471327	2.647796	0.0	1.612903	3.076923	4.938272	26.027397
Loser	88206.0	4.511130	3.174744	0.0	2.150538	4.000000	6.250000	35.416667



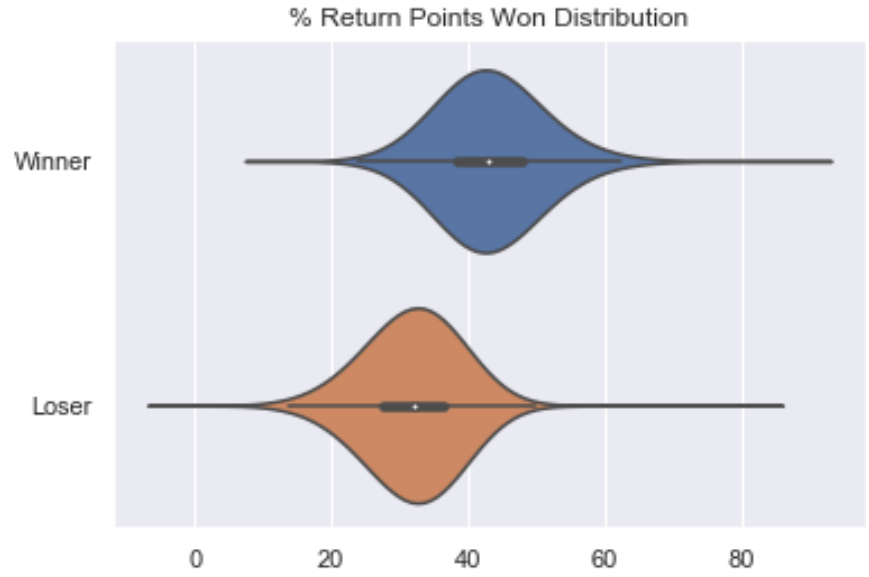
Percent Serve Points Won

	count	mean	std	min	25%	50%	75%	max
Winner	88206.0	68.222500	6.719134	19.696970	63.529412	67.857143	72.580645	100.000000
Loser	88206.0	56.694481	7.437214	14.285714	52.054795	57.031250	61.682243	85.185185



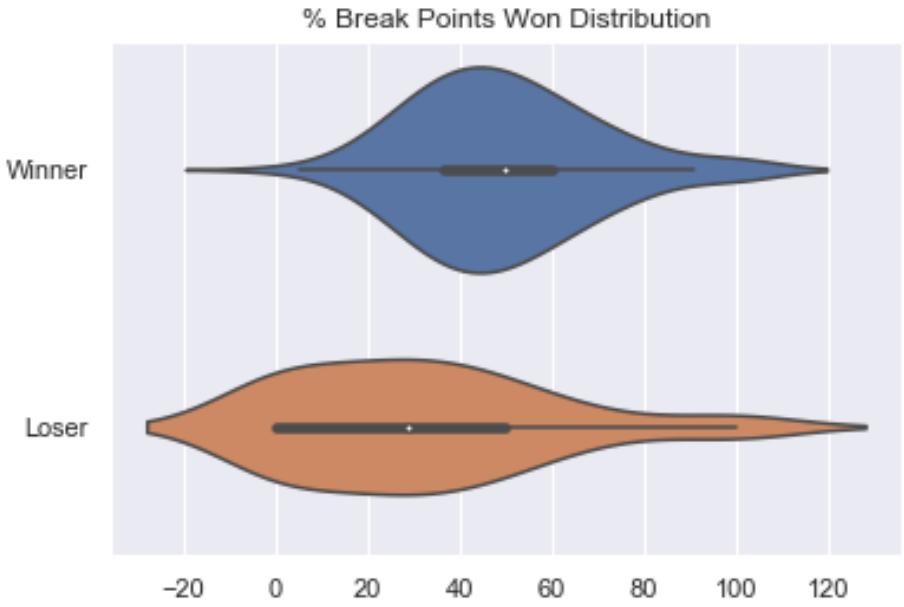
Percent Return Points Won

	count	mean	std	min	25%	50%	75%	max
Winner	88206.0	43.319388	7.449335	14.814815	38.317757	42.982456	47.959184	85.714286
Loser	88206.0	31.802591	6.894812	0.000000	27.419355	32.142857	36.486486	79.166667



Percent Break Points Won

	count	mean	std	min	25%	50%	75%	max
Winner	88206.0	49.534994	19.620488	0.0	36.363636	50.000000	60.0	100.0
Loser	88206.0	31.585153	28.057964	0.0	0.000000	28.571429	50.0	100.0



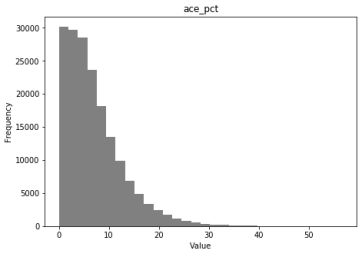
Inferential Statistical Analysis

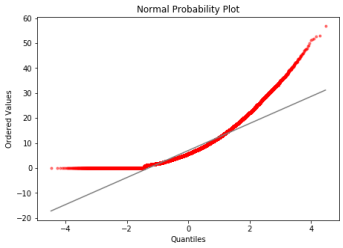
This section loos at statistical significance on observations made and thoughts about the findings during the EDA section. This is an essential step to understanding whether the differences between Match Winners and Losers are unlikely to have happened by chance alone.

The focus lies primarily on three categories: 1) Distribution between Match Winners and Match Losers, 2) Correlation with the Target Variable (i.e., Win or Non-Win), and 3) Collinearity between Features.

Challenges

A challenge for this project was the lack of normally distributed data (see Q-Q plots below). Normally distributed data are often a requirement for classical statistical tests. Fortunately, the Central Limit Theorem allows us to use the Z-test to compare sample distribution differences. (Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This is especially true for sample sizes over 30. In other words, as you take more samples, especially large ones, the graph of the sample means will look more like a normal distribution.)


ace_pct
Frequency
Value
0
5000
10000
15000
20000
25000
30000
0
10
20
30
40
50


Normal Probability Plot
Ordered Values
Quantiles
60
50
40
30
20
10
0
−10
−20
−4
−2
0
2
4