Tennis Prediction Model

Progress Report

Shayla Grymaloski

Ricky Huang

Amaan Makhani

John Schriemer

Bryan Travers

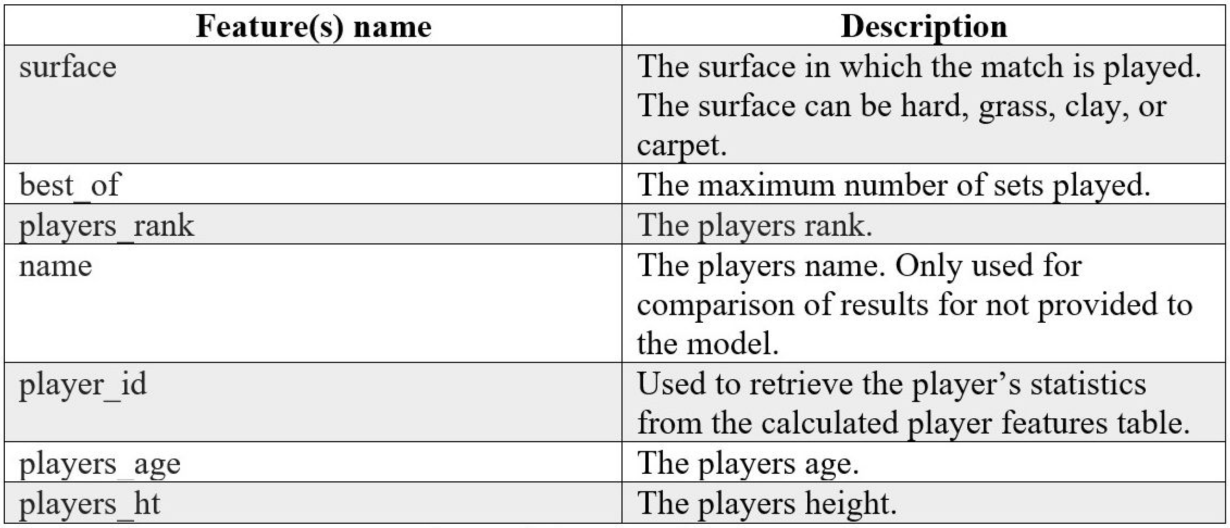
1. Problem Overview

The global sports market reached a value of nearly $488.5 Billion in 2018 [1] and has been growing steadily. Some of this growth is attributed to the increase in sports analytics utilizing statistics and, more recently, machine learning. Using machine learning to analyze and predict sports results has proved both helpful and lucrative for data scientists, coaches, and bookkeepers alike. In particular, prediction models using machine learning have been utilized by those in the growing sports betting market, which has an estimated market value of USD 155.49 billion by 2024 [2]. The demand for accurate predictive models in these markets has prompted a new problem space for machine learning, with many challenges yet to be solved

Tennis poses some unique challenges concerning machine learning. With three different playing surfaces, varying climates and conditions, and diverse playing styles and strategies, the potential for noisy data is nearly guaranteed. Also, it has been difficult to predict a match outcome when the momentum of any tennis match can swing in the matter of a few successfully placed strokes. David Foster Wallace has well-described tennis' unpredictability as "Just one single shot in one exchange in one point of a high-level match is a nightmare of mechanical variables". The large number of variables that define a match has made the problem of using machine learning to predict match outcomes intriguing and worthwhile for our team.

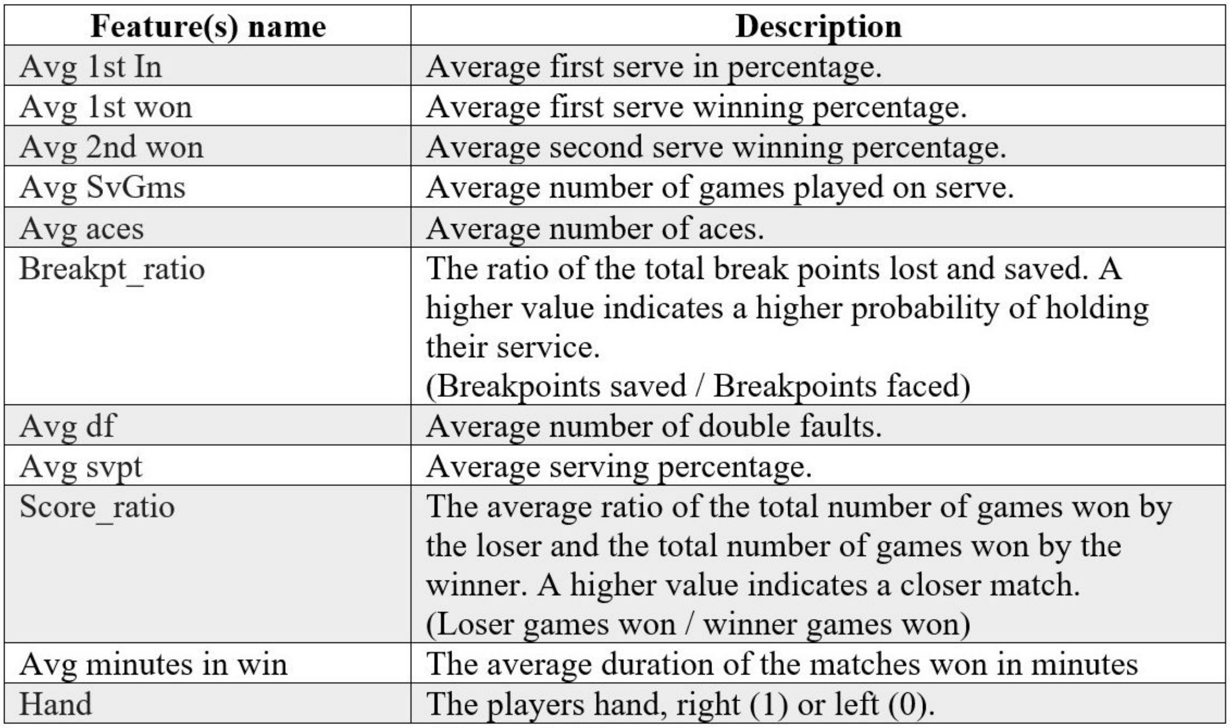
Our team will be using a well-established ATP tour dataset from Kaggle [3]. We will then train the models detailed in the task breakdown and predict the outcome of the 2017 tennis season matches. We will then compare the results with the actual outcome of the 2017 season to obtain our accuracy. The features we chose to use are detailed in ***Figure 1 and 2*** below. The features of each match are detailed in ***Figure 1***. The whole dataset was then used to create player specific features that used the columns from the original dataset and produced the player's averages over the timeframe. These calculated features are described in ***Figure 2***. The columns that were removed from the initial dataset as well as the cleaning process are described in **Section 3.1**.

We chose this dataset because there was a large degree of data spanning various years and many matches. This dataset has also existed for multiple years and is stable. The hope for using this dataset is to discover factors that continuously impact tennis matches and to train our models to pick up on these trends.



Feature(s) name	Description
surface	The surface in which the match is played. The surface can be hard, grass, clay, or carpet.
best_of	The maximum number of sets played.
players_rank	The players rank.
name	The players name. Only used for comparison of results for not provided to the model.
player_id	Used to retrieve the player's statistics from the calculated player features table.
players_age	The players age.
players_ht	The players height.

Figure 1: Pre-match features.



Feature(s) name	Description
Avg 1st In	Average first serve in percentage.
Avg 1st won	Average first serve winning percentage.
Avg 2nd won	Average second serve winning percentage.
Avg SvGms	Average number of games played on serve.
Avg aces	Average number of aces.
Breakpt_ratio	The ratio of the total break points lost and saved. A higher value indicates a higher probability of holding their service. (Breakpoints saved / Breakpoints faced)
Avg df	Average number of double faults.
Avg svpt	Average serving percentage.
Score_ratio	The average ratio of the total number of games won by the loser and the total number of games won by the winner. A higher value indicates a closer match. (Loser games won / winner games won)
Avg minutes in win	The average duration of the matches won in minutes
Hand	The players hand, right (1) or left (0).

Figure 2: Player Features.

Since our proposal, we have decided to focus on the variation of models rather than varying platforms. This means we will use Sklearn with a variety of models as described in this report to generate the final report and presentation. We chose to move in this direction because we felt it would be more effective to spend time on the creation, training, and evaluation of our models rather than spending time to gain proficiency in TensorFlow. As previously mentioned, we also modified our plan for the dataset by using the dataset to gather player statistics then dropping those columns. The original dataset had combined the winners and losers of each match which we split. We will then provide the model with one player’s data and the matches outcome. This is because the model could not process rows with multiple player’s data within it.

2. Goals

Initially, we had the goal of evaluating our data, examining two predictions, correct match predictions and accurate tournament predictions. As well, we also had the original intent of having two platforms to run machine learning experiments. Early on into the implementation of our various models, it became evident that our dataset and goals required further refinement. To combat these new challenges, we looked into several resources, outlined in section 3, to see how practitioners in the field are overcoming similar obstacles. For one, the chosen dataset included several features that did not provide much insight regarding match outcomes.

Additionally, the proposed goal of predicting match and tournament outcomes initially relied on models accurately predicting the winner's ID, but in terms of model performance, predicting player IDs created several issues involving standardization and training of the dataset. As a team, we decided it would be best to focus on predicting match outcomes and not include predictions for tournament outcomes. In addition, we have changed our goal of looking at different machine learning platforms to five different machine learning models. Our goal is for our models to perform with similar accuracies or more effectively than similar reports and that each model we test will be in the range of accuracy between 65% to 70%. This value is based on another report comparing tennis match outcomes where the accuracy came out to be 69.4% [4]. We hope to have an average within the above-mentioned range and will attempt to outperform the accuracy of this report. We are also interested in comparing our trained models' accuracy to compare the accuracy of predicting historic tennis matches to more recent matches. In our final report, we are hoping to compare the five models used to predict our outcomes with k-fold cross-validation. This way, we can have a useful performance metric for our trained models.

Our original goals were formulated at a point in the course where we did not yet understand what was achievable. Since then, we have had more hands-on experience with machine learning tools and a better understanding of what is expected in terms of content depth.

3. Plan and Progress-to-date

Originally the goal for the group, to be achieved by the progress report’s due date, was for each member to experiment independently with their chosen model in order to establish a baseline of accuracy. This involved importing the dataset and creating a model without any parameter tuning, or preprocessing beyond the automatic methods in Scikit Learn’s implementations. Initial results between models were to be compared to determine which of the chosen implementations showed the greatest promise in predicting match and tournament outcomes. However, upon closer examination, it became evident that both the dataset and the utilization of the dataset required further refinement as outlined in **section 1**. Given the time constraints, we adjusted our goal from having baseline model performance metrics, to having a dataset and goal formulation that would work for the purpose of our project.

With this change of focus in mind, our progress to date is promising. We have refined our original objectives to be more attainable and spent considerable time working with the data, with the end goal of creating more accurate models.

With the data cleaning and goal formulation largely behind us, we can now spend the coming weeks tuning our models and pooling our findings. From there, we intend to analyze the results and further refine our models using optimization tools such as grid search and randomized parameter optimization.

3.1. Data Cleaning

The ATP Matches dataset contains details about ATP matches played since 1968, however, detailed match statistics are only available for matches after 1990. Since our group plans on exploring what statistics dictate tennis match results in recent years, we decided to only focus on the matches played from 2013 to 2016. Any columns of missing data were either filled with the appropriate values (0, median, or mode) or the row dropped. One of the challenges our group faced was dealing with the match scores. Tennis has a unique scoring system, playing sets up to 6 or 7, and playing best out of 3 or 5 sets. To combat these inconsistent scores, we created a new column, named score_ratio, that takes the loser's number of games and divides them by the winner's games. The higher the ratio, the closer the match. The next challenge to face is dealing with both the loser and winner data in each row. Our team has been brainstorming some ideas as to how to deal with this and has decided to provide only the information of one player and the corresponding match outcome to the model. We have also used the original dataset to produce average statistics about players over this time frame. These stats were created through excel and using columns from the original dataset. The new calculated player features are described in **section 1.**

4. Task Breakdown

In our previous formal proposal, each team member was responsible for cleaning their assigned dataset. However, since the original proposal, we have decided to stick to one dataset for our project. Therefore, in order for each member to have a unique role, each team member will be responsible for experimenting with at least one model.

In response to the feedback from our proposal submission, our group has split the remaining tasks up explicitly. Amaan will be in charge of experimenting with an ANN model using stochastic gradient descent and the creation of the player dataset. Similarly, Shayla will examine ANN performance under the perceptron model. In addition to cleaning the dataset, John will implement a random forest implementation. Ricky is responsible for investigating the accuracy of a Naive Bayes predictive model, and Bryan will attempt to use logistic regression and SVMs to predict match outcomes. Upon completion of the necessary experiments, report planning will begin and the task will be split evenly among members.

Date	Goals
Week 1 **July 13-19**	Finishing up data parsing and generation. **Mid-week:** Player data table is complete and a clear way of using this dataset is agreed upon. **End of the week:** Initial model implementation underway.
Week 2 **July 20-26**	Each individual team member's model is optimized and produces some visuals and values. **Mid-week:** Meet to discuss the initial optimization process and create a consistent visual representation of data.
Week 3 **July 27- August 3**	Writing and finishing up the report. The report is split into sections where each team member is responsible for one. **Mid-week:** Team members begin writing sections of the report. **End of the week:** Report is ready to be edited and will be finalized prior to submission.

5. Initial Results

Section 3 explains the initial challenges of working with our chosen data set, however, our team was still able to create some initial visualizations to gain insights into some of the ATP data set's features. **Figure 4** displays a correlation heat map for all columns of the initial dataset. The four pockets of higher correlations are serving, returning, and match statistics. This can be attributed to the fact that a higher first-serve percentage and the ace count will correlate with higher amounts of total won serving games. On the other hand, areas with low correlation are columns like player name, player country, and player height. We plan on creating another visualization like this once we finalize our dataset, and decide how to structure this data to avoid having both the winner and loser data in the same row. The winner and loser data within the same rows make it difficult in projecting how the models will use the data given and possibly correlate the winner to the order of player data given.

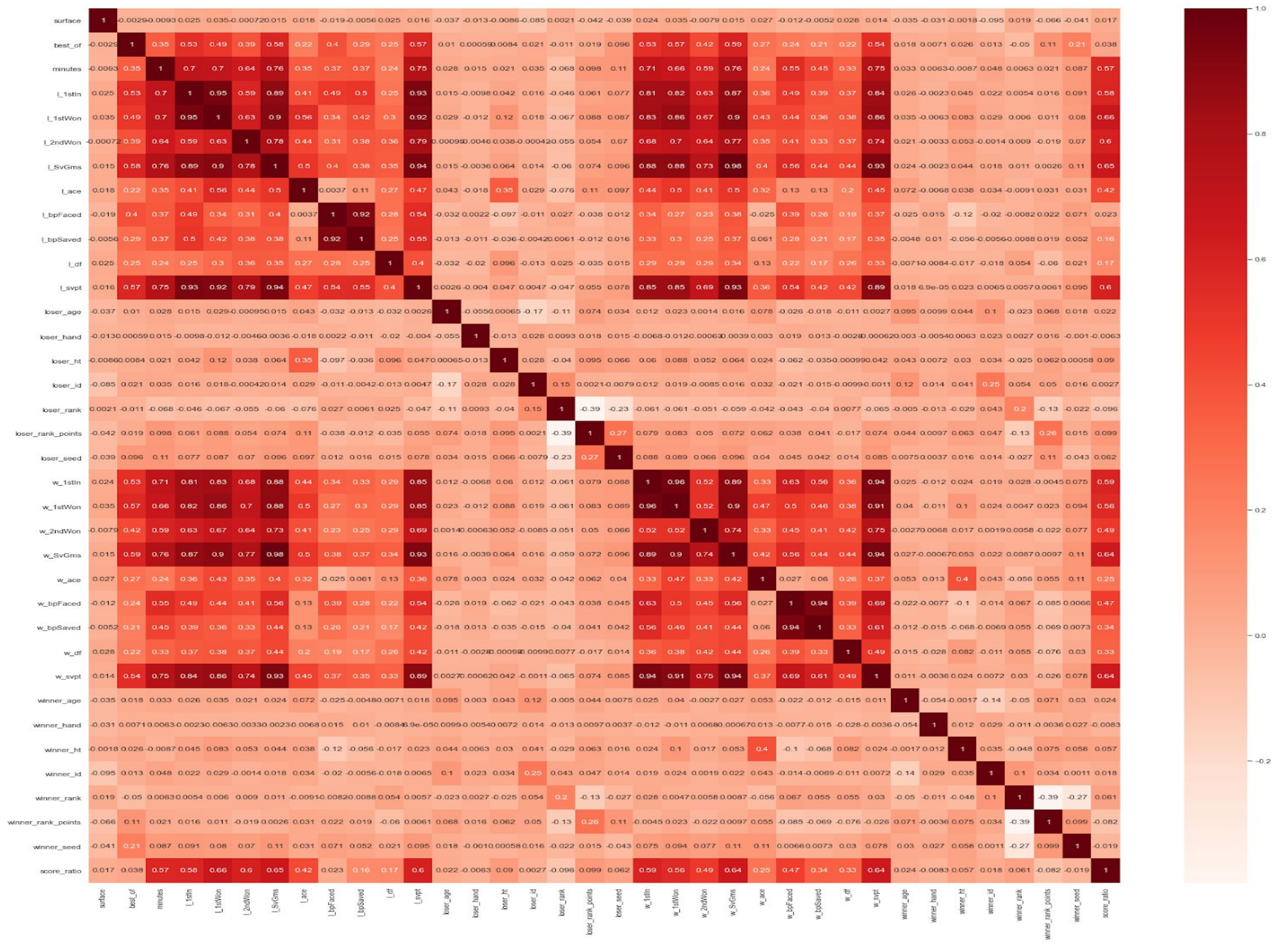
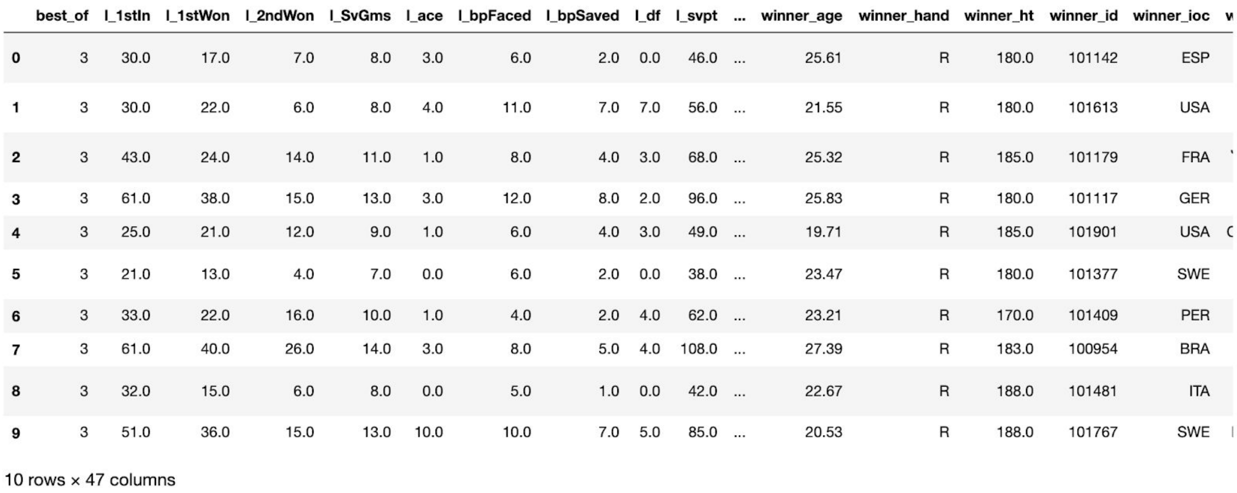


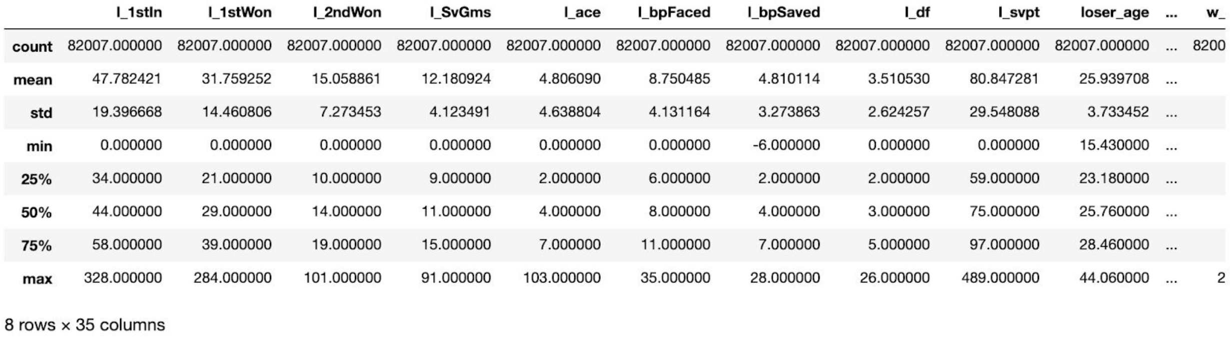
Figure 4: Correlation Heatmap for Initial ATP Dataset



	best_of	l_1stIn	l_1stWon	l_2ndWon	l_SvGms	l_ace	l_bpFaced	l_bpSaved	l_df	l_svpt	...	winner_age	winner_hand	winner_ht	winner_id	winner_ioc	w
0	3	30.0	17.0	7.0	8.0	3.0	6.0	2.0	0.0	46.0	...	25.61	R	180.0	101142	ESP	
1	3	30.0	22.0	6.0	8.0	4.0	11.0	7.0	7.0	56.0	...	21.55	R	180.0	101613	USA	
2	3	43.0	24.0	14.0	11.0	1.0	8.0	4.0	3.0	68.0	...	25.32	R	185.0	101179	FRA	
3	3	61.0	38.0	15.0	13.0	3.0	12.0	8.0	2.0	96.0	...	25.83	R	180.0	101117	GER	
4	3	25.0	21.0	12.0	9.0	1.0	6.0	4.0	3.0	49.0	...	19.71	R	185.0	101901	USA	C
5	3	21.0	13.0	4.0	7.0	0.0	6.0	2.0	0.0	38.0	...	23.47	R	180.0	101377	SWE	
6	3	33.0	22.0	16.0	10.0	1.0	4.0	2.0	4.0	62.0	...	23.21	R	170.0	101409	PER	
7	3	61.0	40.0	26.0	14.0	3.0	8.0	5.0	4.0	108.0	...	27.39	R	183.0	100954	BRA	
8	3	32.0	15.0	6.0	8.0	0.0	5.0	1.0	0.0	42.0	...	22.67	R	188.0	101481	ITA	
9	3	51.0	36.0	15.0	13.0	10.0	10.0	7.0	5.0	85.0	...	20.53	R	188.0	101767	SWE	

10 rows × 47 columns

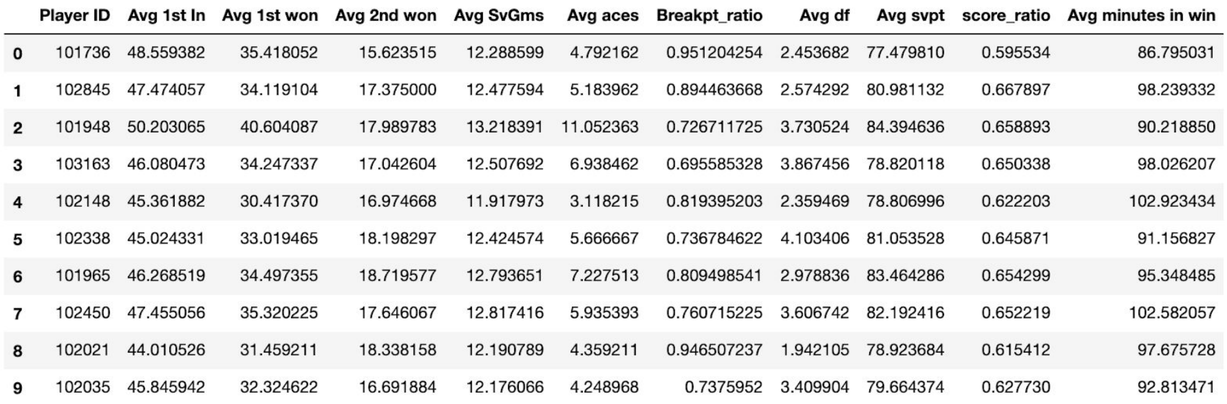
Figure 5: Cleaned data set overview.



	l_1stIn	l_1stWon	l_2ndWon	l_SvGms	l_ace	l_bpFaced	l_bpSaved	l_df	l_svpt	loser_age	...	w_
count	82007.000000	82007.000000	82007.000000	82007.000000	82007.000000	82007.000000	82007.000000	82007.000000	82007.000000	82007.000000	...	8200
mean	47.782421	31.759252	15.058861	12.180924	4.806090	8.750485	4.810114	3.510530	80.847281	25.939708	...	
std	19.396668	14.460806	7.273453	4.123491	4.638804	4.131164	3.273863	2.624257	29.548088	3.733452	...	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-6.000000	0.000000	0.000000	15.430000	...	
25%	34.000000	21.000000	10.000000	9.000000	2.000000	6.000000	2.000000	2.000000	59.000000	23.180000	...	
50%	44.000000	29.000000	14.000000	11.000000	4.000000	8.000000	4.000000	3.000000	75.000000	25.760000	...	
75%	58.000000	39.000000	19.000000	15.000000	7.000000	11.000000	7.000000	5.000000	97.000000	28.460000	...	
max	328.000000	284.000000	101.000000	91.000000	103.000000	35.000000	28.000000	26.000000	489.000000	44.060000	...	2

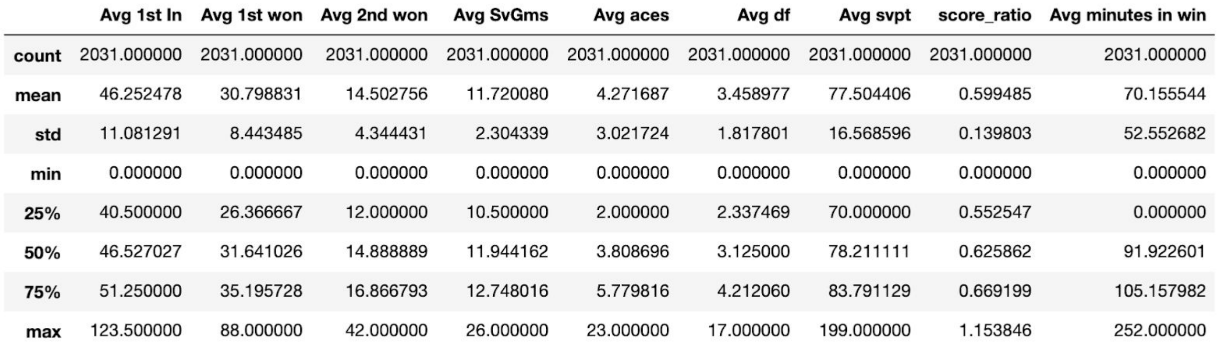
8 rows × 35 columns

Figure 6: Stats on our cleaned data set.



	Player ID	Avg 1st In	Avg 1st won	Avg 2nd won	Avg SvGms	Avg aces	Breakpt_ratio	Avg df	Avg svpt	score_ratio	Avg minutes in win
0	101736	48.559382	35.418052	15.623515	12.288599	4.792162	0.951204254	2.453682	77.479810	0.595534	86.795031
1	102845	47.474057	34.119104	17.375000	12.477594	5.183962	0.894463668	2.574292	80.981132	0.667897	98.239332
2	101948	50.203065	40.604087	17.989783	13.218391	11.052363	0.726711725	3.730524	84.394636	0.658893	90.218850
3	103163	46.080473	34.247337	17.042604	12.507692	6.938462	0.695585328	3.867456	78.820118	0.650338	98.026207
4	102148	45.361882	30.417370	16.974668	11.917973	3.118215	0.819395203	2.359469	78.806996	0.622203	102.923434
5	102338	45.024331	33.019465	18.198297	12.424574	5.666667	0.736784622	4.103406	81.053528	0.645871	91.156827
6	101965	46.268519	34.497355	18.719577	12.793651	7.227513	0.809498541	2.978836	83.464286	0.654299	95.348485
7	102450	47.455056	35.320225	17.646067	12.817416	5.935393	0.760715225	3.606742	82.192416	0.652219	102.582057
8	102021	44.010526	31.459211	18.338158	12.190789	4.359211	0.946507237	1.942105	78.923684	0.615412	97.675728
9	102035	45.845942	32.324622	16.691884	12.176066	4.248968	0.7375952	3.409904	79.664374	0.627730	92.813471

Figure 7: The player dataset consisting of calculations we made.



	Avg 1st In	Avg 1st won	Avg 2nd won	Avg SvGms	Avg aces	Avg df	Avg svpt	score_ratio	Avg minutes in win
count	2031.000000	2031.000000	2031.000000	2031.000000	2031.000000	2031.000000	2031.000000	2031.000000	2031.000000
mean	46.252478	30.798831	14.502756	11.720080	4.271687	3.458977	77.504406	0.599485	70.155544
std	11.081291	8.443485	4.344431	2.304339	3.021724	1.817801	16.568596	0.139803	52.552682
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	40.500000	26.366667	12.000000	10.500000	2.000000	2.337469	70.000000	0.552547	0.000000
50%	46.527027	31.641026	14.888889	11.944162	3.808696	3.125000	78.211111	0.625862	91.922601
75%	51.250000	35.195728	16.866793	12.748016	5.779816	4.212060	83.791129	0.669199	105.157982
max	123.500000	88.000000	42.000000	26.000000	23.000000	17.000000	199.000000	1.153846	252.000000

Figure 8: Stats on our calculated data set.

6. References

1) ltd, R., 2020. Sports Global Market Opportunities And Strategies To 2022. [online] Researchandmarkets.com. Available at: <https://www.researchandmarkets.com/reports/4770417/sports-global-market-opportunities-and-strategies?utm_source=BW&utm_medium=PressRelease&utm_code=ctvc8g&utm_campaign=1244426+-+Sports+-+%24614+Billion+Global+Market+Opportunities+%26+Strategies+to+2022&utm_exec=joca220prd> [Accessed 20 June 2020].

2) R. ltd, "Sports Betting Market by Platform, by Type, and by Sports Type: Global Industry Perspective, Comprehensive Analysis, and Forecast, 2017-2024", Researchandmarkets.com, 2020. [Online]. Available: https://www.researchandmarkets.com/reports/4853933/sports-betting-market-by-platform-by-type-and. [Accessed: 20- Jun- 2020].

3) S. VM, "ATP matches, details of the ATP matches since 1968", *Kaggle.com*, 2020. [Online]. Available: https://www.kaggle.com/sijovm/atpdata. [Accessed: 08- Jul- 2020].

4) A. Cornman, G. Spellman and D. Wright, "Machine Learning for Professional Tennis Match Prediction and Betting", Cs229.stanford.edu, 2020. [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5242116.pdf. [Accessed: 20- Jun- 2020].

5) A. Wagner and D. Narayanan, "Using Machine Learning to predict tennis match outcomes", Deepakn94.github.io, 2020. [Online]. Available: http://deepakn94.github.io/assets/papers/6.867.pdf. [Accessed: 20- Jun- 2020].