# Project Tennis ML

Report

Marvin Ford

Vano Mahi

## Introduction

Tennis is a multi-billion dollar sport that relies extensively on data and up-to-the minute analysis to provide insights so that players can improve their games; for coaches to be more intentional with training regiments and match scouting; for television networks to provide meaningful coverage; for engagement of fans in the tournaments through fantasy leagues, and looking at head-to-head win loss rates between players; for gaming companies to provide odds-making analysis; for commentators to have background informations on players to speak knowingly and intelligently , etc.

The main goal of this project was to leverage the ATP and WTA results to predict who will win the 2021 French Open men's and women's championships. The minor goal is to apply what we have learned from our models to predict the winners of other tournaments. Ultimately, we want to continue this project and provide meaningful insights in the form of an app to engage and encourage tennis fans across the world to build their own fantasy leagues as per our predictive scores. And a pipe dream is to engage players, commentators, networks, and betting companies to leverage our analysis for their unique needs.

Datasets of tennis matches were acquired and explored. The exploration provided insights on what features needed to be cleansed, typed-converted, binary encoded, et cetera. The data was subsequently cleaned and pickled for PCA Analysis, scaling, dimensionality reduction, machine learning analysis and evaluation.

Several notebooks were created and the project was  managed through our github repository.

Again, our goal was to leverage machine learning to predict who will win the 2021 French Open Championship.  We trained and tested our models on all the data except for the 2020 French Open Results. These hold-out sets were used to validate if the best model produced the same results with optimal accuracy.

## Background

The French Open is the second grand slam tournament of each calendar year. It has been held in Paris at the Roland Garros stadium every year since 1891. A grand slam tournament is what every professional tennis player dreams of winning. Ultimately, our goal is to predict winners and losers.

The problem is that we have tennis data and we want to preprocess it and leverage machine learning to predict winners and losers. The data contains many important dimensions and metrics that are standard in the tennis arena.

## Data

We acquired the data from Jeff Sackmann's github repository. Jeff Sackmann is an author and software developer who has worked in the fields of sports statistics and test preparation. He curates tennis datasets from the 1940 of all WTA and ATP matches. We focused on matches played between the year 2000 to 2021.
We developed a suite of helper functions to preprocess the date— filled in missing values, converted string formatted columns to numerical values and date to date time format, dropped rows and irrelevant features after extracting insights from the RandomForrest model.

Although some of the features are common on both datasets, ATP and WTA, we decided to work on each separately. It was a good strategy to take this route because each dataset was computationally easier to handle. It is important to note that the helper functions were used for both sets.

After removing the tournament which had over 1600 unique names, we ended up with 55 features and 119,438 rows for the WTA data, and 46 features and 128,070 rows for the ATP data. These numbers included the target variables, and an extra feature to indicate the rows that are aligned to the 2020 french Open Tournaments to easily filter for our experiment later.

It's important to note that we had to expand each dataset, because winners and losers were on the same row. We had to separate the winner from the losers and unionized both datasets. We also created the target variable.

Helper functions were also developed for the PCA analysis—to generate PCA results and visualization seamlessly.

## Algorithm

We focused our attention on the sklearn machine learning library and leveraged algorithms such as PCA,  Logistic Regression, and RandomForest, Bagging Classifiers, Standard Scaler, etc.

## Results

### PCA

We fed the preprocessed data into the PCA algorithm. At first,  finding patterns in the data was difficult because we could not identify the variability in the first 20 principal components.  The preprocessed data for the ATP had over 1700 features.  We tried up to 100 principal components and still couldn't identify any patterns.

We used the RandomForrest model to find the important features. Based on the result, we realized that the tournament name had no predictive importance and it created many extra columns that increased the processing time, so we dropped it. We reran the results and got better results. Now we can explain 50% of the variance in the first five components of the ATP data, and approximately 40% on the WTA data, which is way better.

### Dimensionality reduction

We wanted to find the most important features to optimize the predictability of the model and most importantly to reduce the number of features. So we found the important features using the RandomForest Algorithms.  Some of the notable features were break-point faced, age of the player, number or ranking points, the ranking of the player, first serve won amongst others not mentioned here. These results were consistent across both datasets. Our analysis during this phase of the project gave us the insight to drop the tournament name from the process, which had no impact on the results of the model.

Machine Learning

For our experiment, we removed the French Open 2020 data from both ATP and WTA datasets. Our goal here was to apply the best model to this data to see if it makes the correct predictions. The data was standardized and separated into training and testing sets with a test size of 30%. The models were applied accordingly.

### The ATP Analysis

The logistic regression model produced an accuracy scores of on training set of 0.79 and 0.7852 testing set— the training set slightly performed better than the test set (overfitting). The f1 scores were 0.7920 on the training set and 0.7867 on the testing set. The precision scores on the training set were 0.7845 and 0.7812 on the testing set. The recall scores were 0.7996 on the training set and 0.7924 on the testing set.

Next, we tried Random Forest which generated accuracy scores on the training set of 0.9898 and 0.7415 on the test set. This model suffers from high variance, it performed better on the training set and significantly worse on the test set. The f1 score on the training set was 0.9898 and 0.7270 on the testing set. The precision scores were 0.9961 on the training sets and 0.7702 on the test set. And the recall scores were 0.9835 on the training set and 0.6883 on the test set.

We pretty much got the same scores on the Bagging Classifier compared to the Random Forest classifier— the accuracy scores were 0.9898 on the training set and 0.7415 on the test sets. It produced f1 scores of 0.9898 on the training set and 0.7270 on the testing set. The precision scores were 0.9961 on the training set and 0.7702 on the testing set. The recall scores were 0.9835 on the training set and 0.6883 on the testing set. This model overfitted significantly.

Finally, we tuned the hyperparameters of the Logistic Regression and cross-validated using GridSearchCV. It produced a best parameter of {'C': 1.0}, which resulted in a best score of 0.7899 and an accuracy score of 0.7900.

**The WTA Analysis**

The Logistic Regression produced accuracy scores of 0.7315 on the training set and 0.7308 on the testing set. The f1 scores were 0.7247 on the training set of 0.7247 and 0.7222 on the testing set. Additionally, the precision scores were on the training sets and 0.7460 on the testing set. The recall scores were 0.7068 on the training set and 0.6999 on the testing set.

The Random Forest produced accuracy scores on the training set and 0. 0.6956 on the test set. The f1 scores were 0.9860 on the training set and 0.6781 on the testing set. The precision scores were 0.9942 on the training set and 0.7193 on the test set. The recall scores were 0.9780 on the training set and 0.6414 on the testing set.

The Bagging classifier produced accuracy scores of 0.9861 on the training set and 0.6956 on the test set, the f1 scores were 0.9860 on the training set and 0.6781 on the testing set. The precision scores were 0.9942 on the training set and 0.7193 and on the testing set. The recall scores were 0.9780 on the training and 0.6414 on the test set.

The hyperparameter tuning of the Logistic Regression with GridSearchCV produced a best parameter of {'C': 100.0}. The best score was 0.7350, and an accuracy score was 0.7355.

### The Experiment

The analysis showed that random forest is the best model, therefore we decided to apply it on only the French 2020 tournament data for both ATP and WTA. We got an accuracy of 0.50 for both datasets. The model predicted that all of the players were winners. It resulted in too many false positives. The model is not good on new data.

## Conclusion

The accuracy of our models were fair. However, all of them that we tried overfitted—they performed better on the training sets but not on the test sets. The experiment of predicting on the French Open 2020 holdsets produced significant false positives at a rate of 50 percent for both ATP and WTA. The model didn't accurately predict the winners. More experimentation and exploration with machine learning is needed. This problem is completely solvable, but we have to do effective model selection and regularization. The Random Forest Classifier worked well to identify the important features, which helped to significantly reduce the dimensionality of each dataset. Our PCA results improved significantly, we could identify the variability in the dataset, at least 40 percent in the first five components.

## Key Code

The key codes are linked below:

- ❑ Data Cleansing
- ❑ Feature Importance
- ❑ PCA Analysis
- ❑ Machine Learning

## Map of the repository

1. The Folder called data contains the ATP and WTA datasets from 2000 to 2021 and their dictionary)

- atp_matches (datasets of the ATP matches from 2000 to 2021)
- wta_matches ( contains the datasets of the WTA matches from 2000 to 2021)
- ATP_WTA_matches_data_dictionary (Is the dictionary of both datasets)

2. The Folder called data_cleaning contains all the notebooks and functions we used to cleaning the data

- ATP_cleaning.ipynb
- data_cleaning.ipynb
- help_functions_cleaned_atp-Copy1.ipynb
- help_functions_cleaned_atp.ipynb
- helper_functions-wta_data_cleaning.ipynb

3. The Folder called feature_importance contains:

- ATP_Feature_Importance.ipynb (the features importance and selection of the ATP dataset)
- WTA_Feature_Importance.ipynb (the features importance and selection of the WTA dataset)

4. The Folder called machine_learning contains all the Machine Learning models we ran and compared.

- ATP_ML.ipynb (Machine learning model on the ATP dataset)
- WTA_ML.ipynb (Machine learning model on the WTA dataset)
- predicting_tennis_results.ipynb (our first draft of the first dataset)

5. The Folder called other contains:

- wtaFrenchOpen2020.ipynb (The French Open 2020 dataset drawn from the original dataset. We isolated the French Open of the year 2020 so that we could apply our best model on it and see how well our model would predict the winner of the French Open 2020.)

6. The Folder called pca_analysis contains the PCA for both ATP and WTA
    - PCAAnalysisATP.ipynb (PCA for ATP)
    - PCAAnalysisWTA.ipynb (PCA for WTA)

7. The file .ipynb is our first draft when trying to aggregate all the dataset into one
8. DSE2100_ML_project_dataset.csv is the aggregate dataset as a csv file
9. DSE2100_ML_project_dataset.pkl is the aggregate dataset created by pickle, a Python module that enables objects to be serialized to files on disk and deserialized back into the program at runtime. It contains a byte stream that represents the objects.

## Project Organization

The project was managed through our github repository repository. We cleaned, converted string types to numeric, dropped the features, imputed missing values, and binary encoded categorical features. That cleansed data was pickled so that we could both work independently on various bits. We opened issues that we faced on github and resolved them when we had a resolution.

Van and I committed our work to the repository. Before pushing to the repo we did a pull request (PR) to avoid merge conflicts when pushing to repo.

Two datasets were leveraged, so we have notebooks for both the WTA and the ATP. All the notebooks are arranged in their respective folders.

## Contribution

| Marvin Ford | Vano Mahi |
|---|---|
| Data Exploration<br>Data Cleaning— Built helper functions<br>PCA Analysis — Built helper functions<br>Dimensionality Reductions/MachineLearning<br>Report Writing | Data Acquisition<br>Data Exploration<br>Exploratorator Data Analysis<br>Machine Learning<br>Machine Learning Analysis [Report]<br>Map of the Repository |

## Bibliography

Works Cited

ATP Tour. “Coria: To Play Against Djokovic In His Home City Is 'A Dream'.” *ATP Tour*, www.atptour.com/.

“JeffSackmann - Overview.” *GitHub*, github.com/JeffSackmann.

“The Official Home of the Women's Tennis Association: WTA Tennis.” *Women's Tennis Association*, www.wtatennis.com/.

“Tennis Betting: Tennis Results: Tennis Odds.” *Tennis Betting | Tennis Results | Tennis Odds*, www.tennis-data.co.uk/alldata.php.

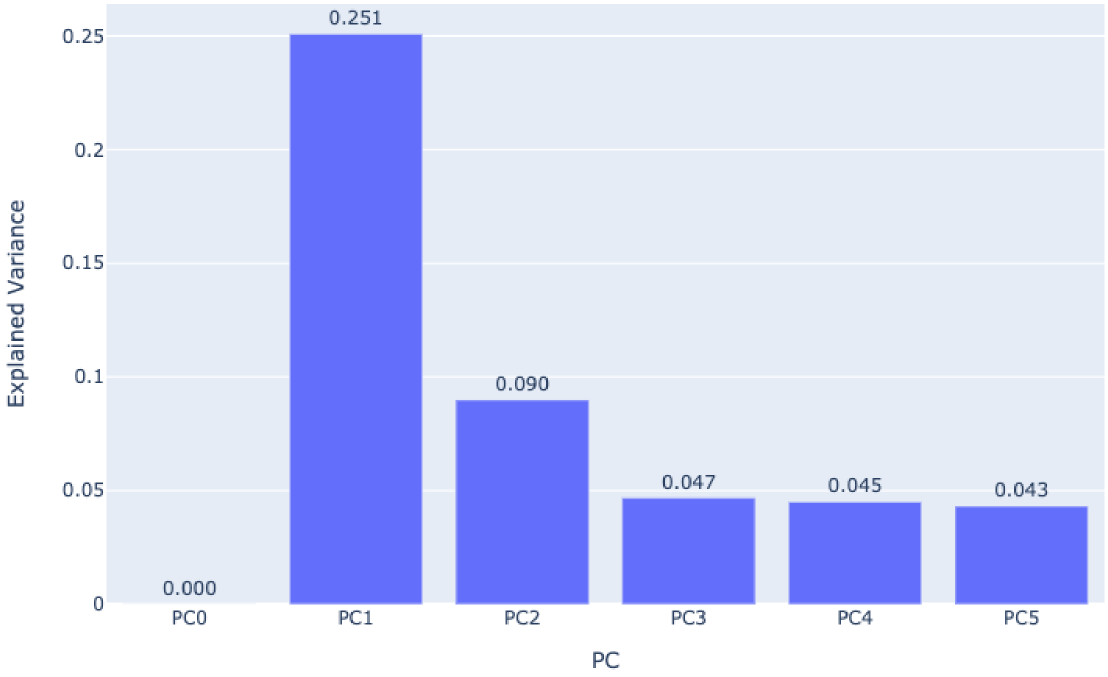
“Where the World Builds Software.” *GitHub*, www.github.com/.

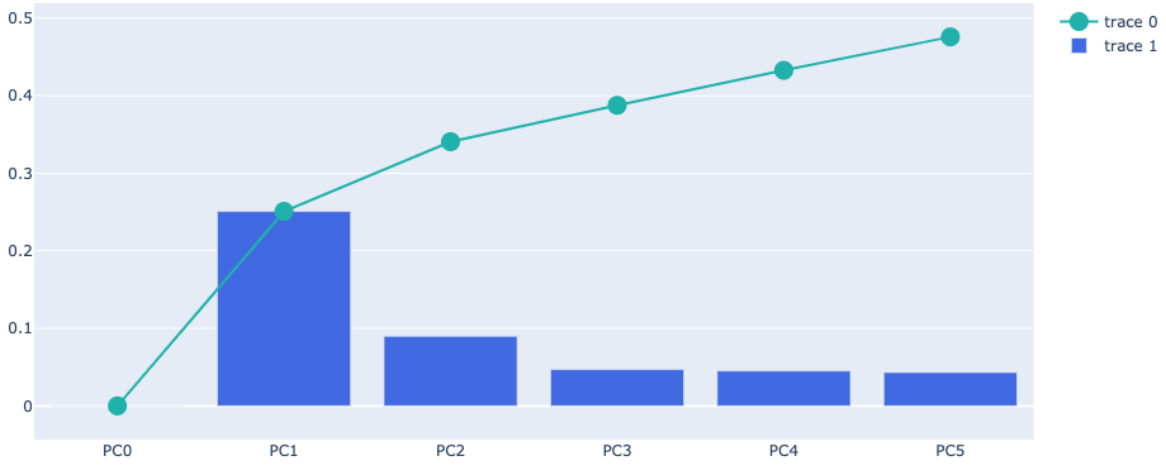
Chavda, Jyoti & Patel, Namrata & Vishwakarma, Prasant. (2019). Predicting tennis match-winner and comparing bookmakers odds using machine learning techniques.. 10.13140/RG.2.2.10920.32009.

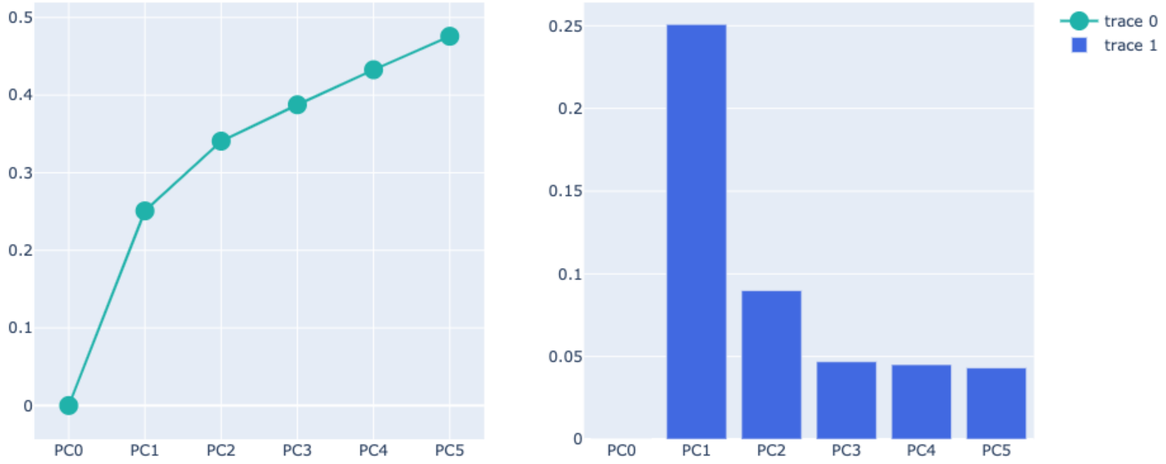
Gu, Wei & Saaty, Thomas. (2019). Predicting the Outcome of a Tennis Tournament: Based on Both Data and Judgments. Journal of Systems Science and Systems Engineering. 28. 317-343. 10.1007/s11518-018-5395-3.

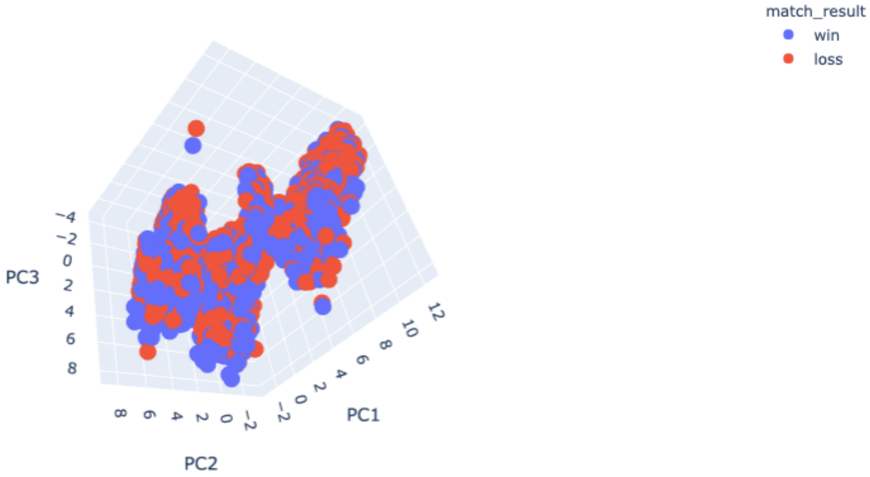
## Appendix

### PCA Visualizations for ATP

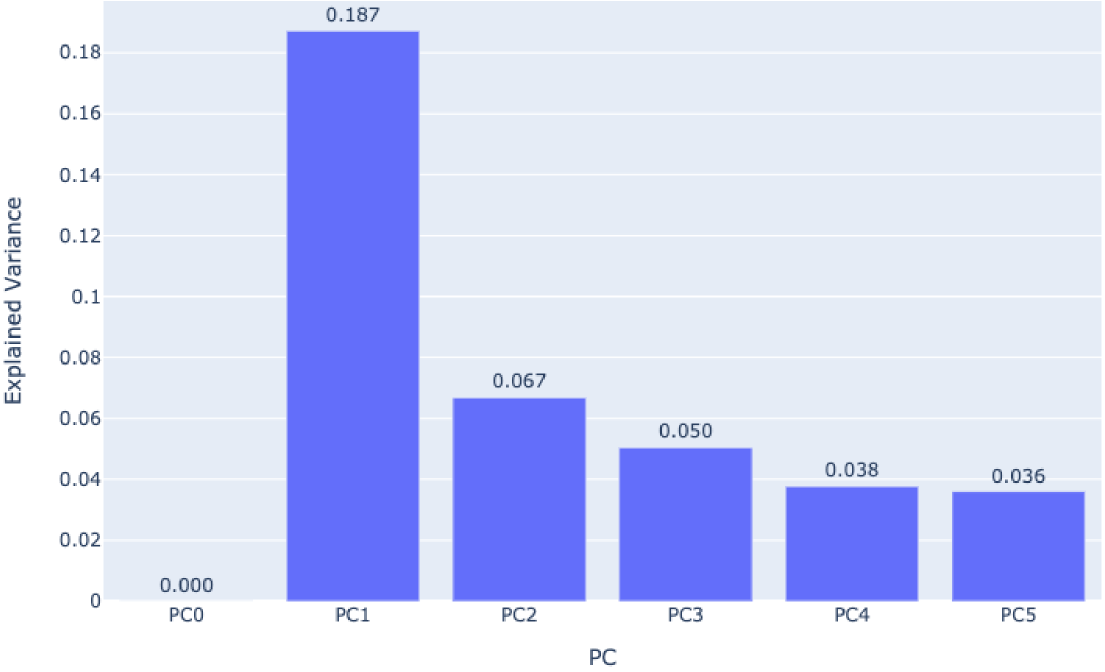


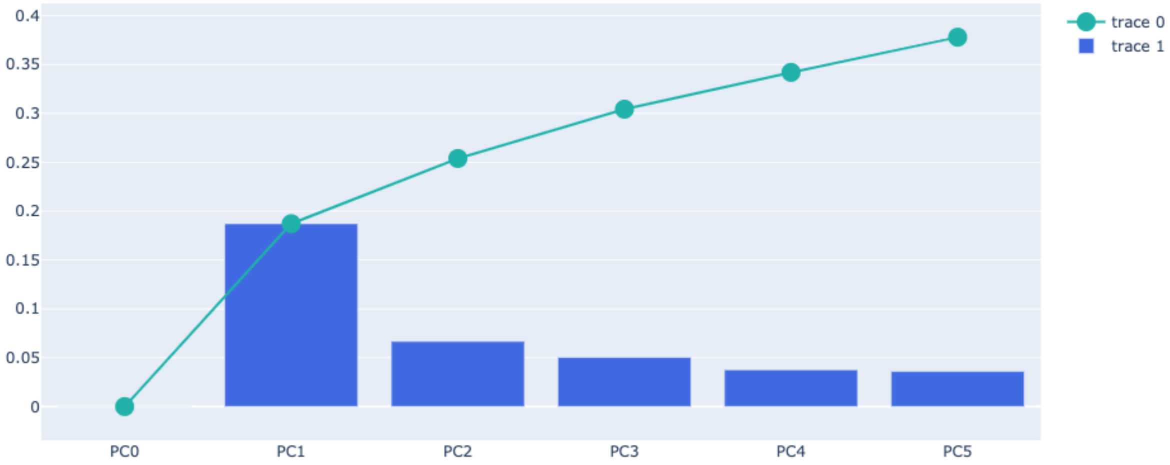

0.5
0.4
0.3
0.2
0.1
0
PC0
PC1
PC2
PC3
PC4
PC5
trace 0
trace 1

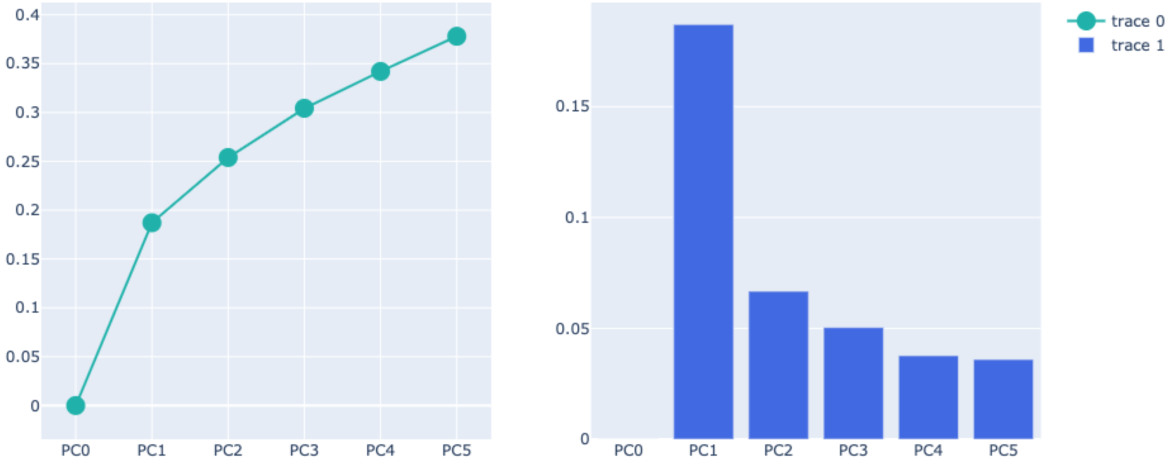

0.5
0.4
0.3
0.2
0.1
0
PC0
PC1
PC2
PC3
PC4
PC5
0.25
0.2
0.15
0.1
0.05
0
PC0
PC1
PC2
PC3
PC4
PC5
trace 0
trace 1

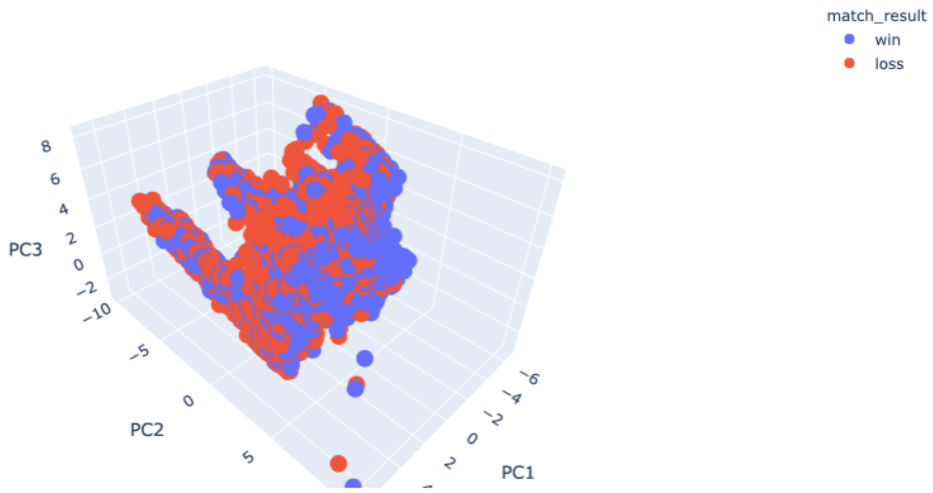


## PCA Visualization for WTA

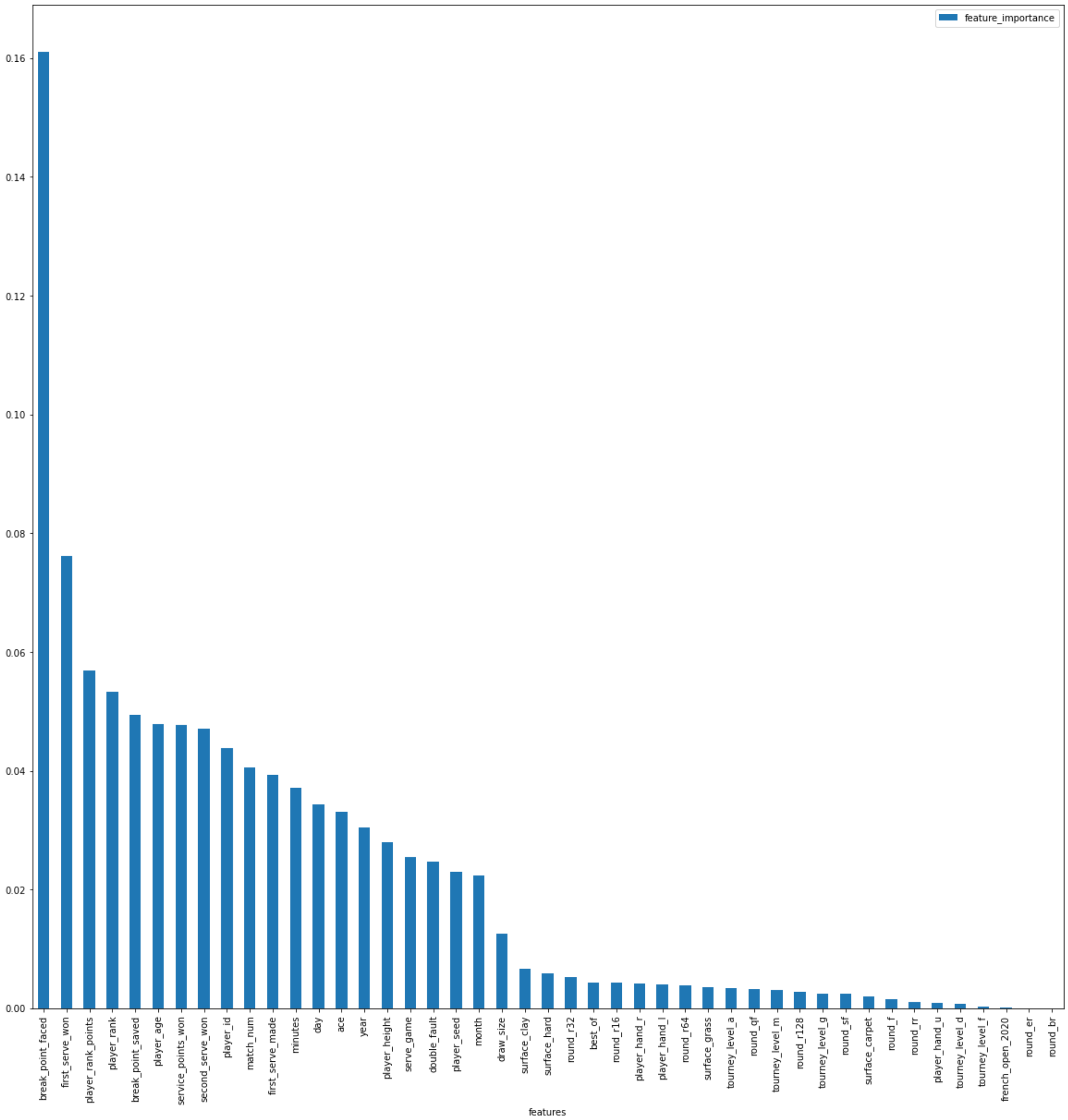



0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
PC0
PC1
PC2
PC3
PC4
PC5
trace 0
trace 1

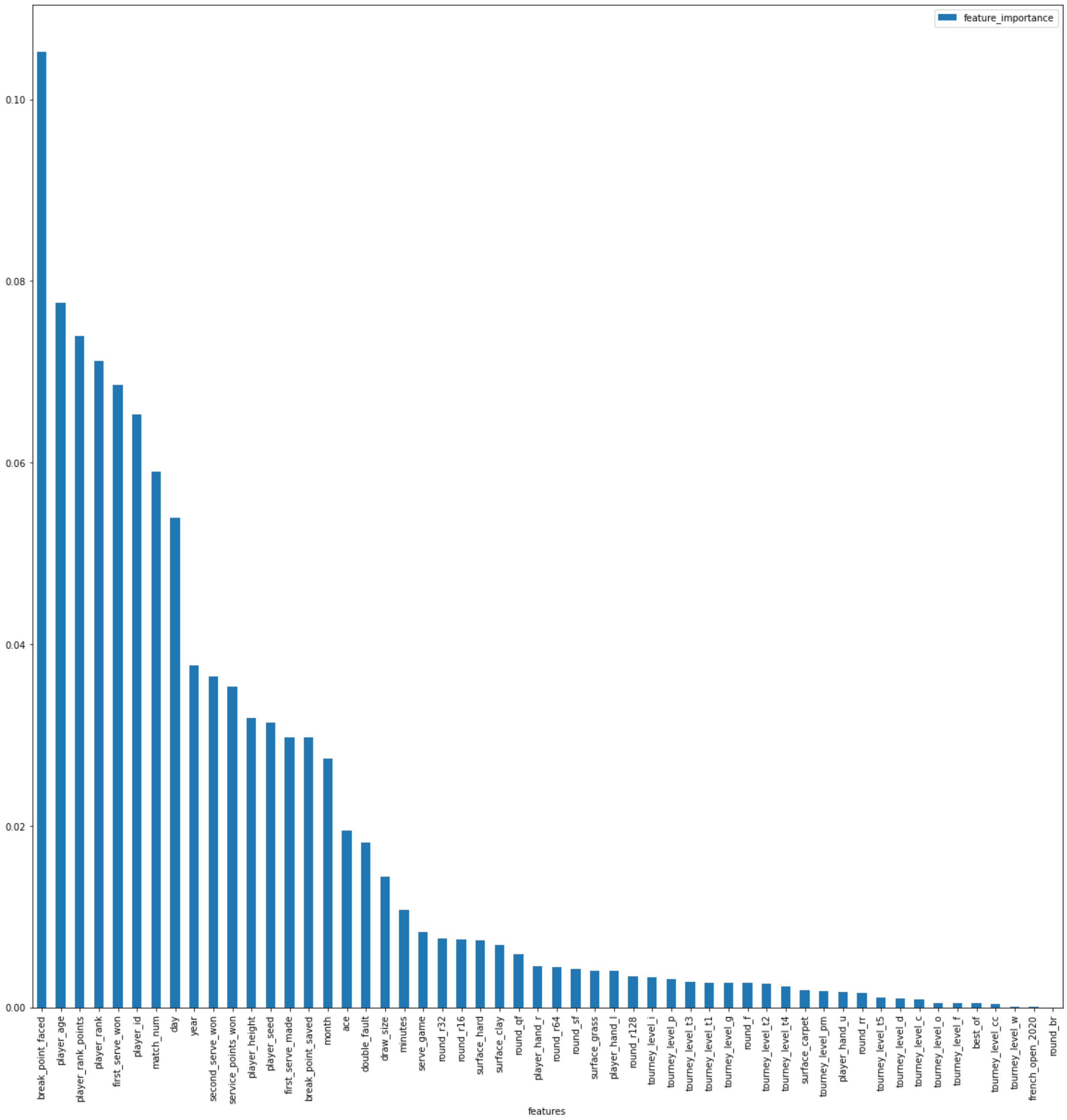

0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
PC0
PC1
PC2
PC3
PC4
PC5
0.15
0.1
0.05
0
PC0
PC1
PC2
PC3
PC4
PC5
trace 0
trace 1



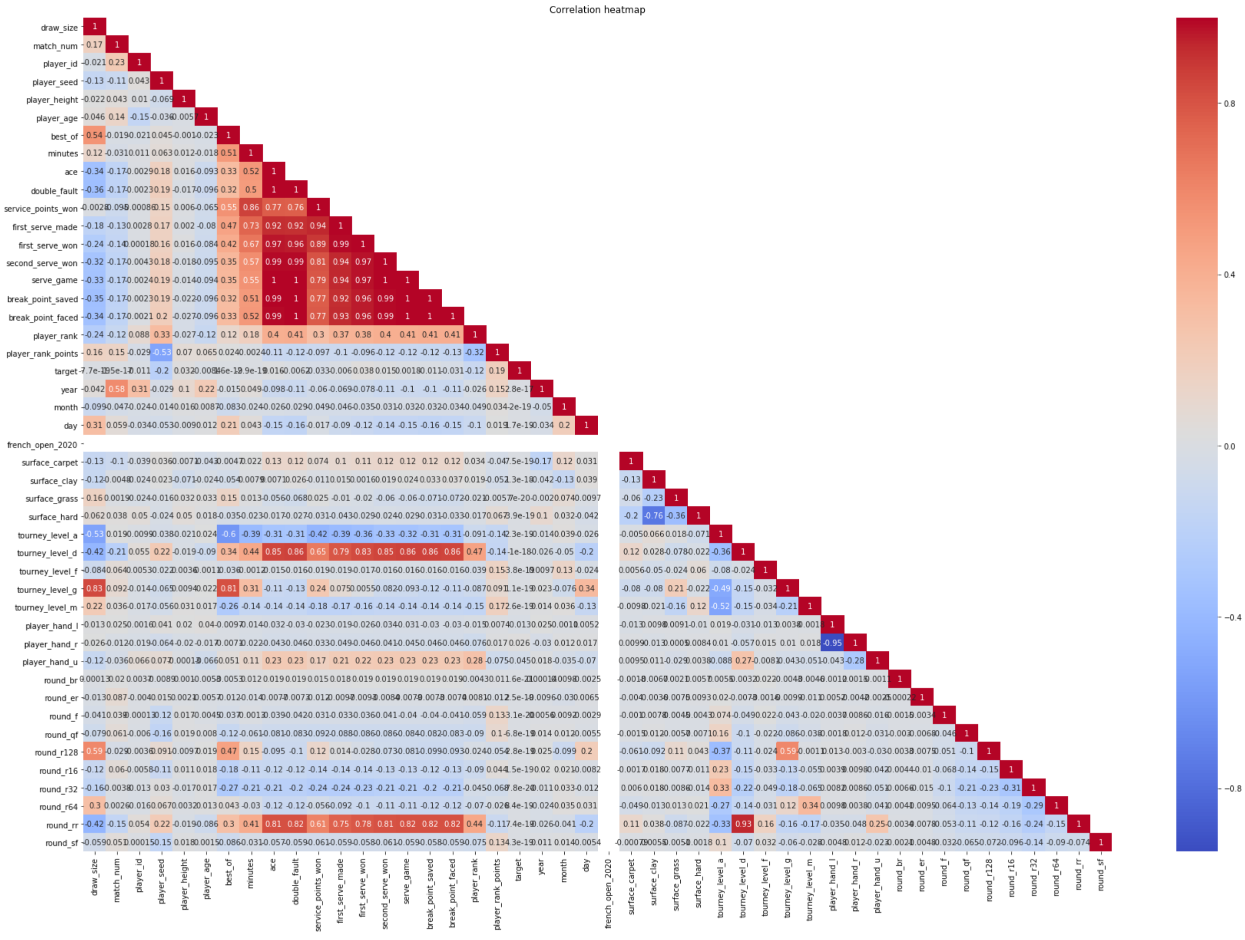
## Feature Importance for ATP



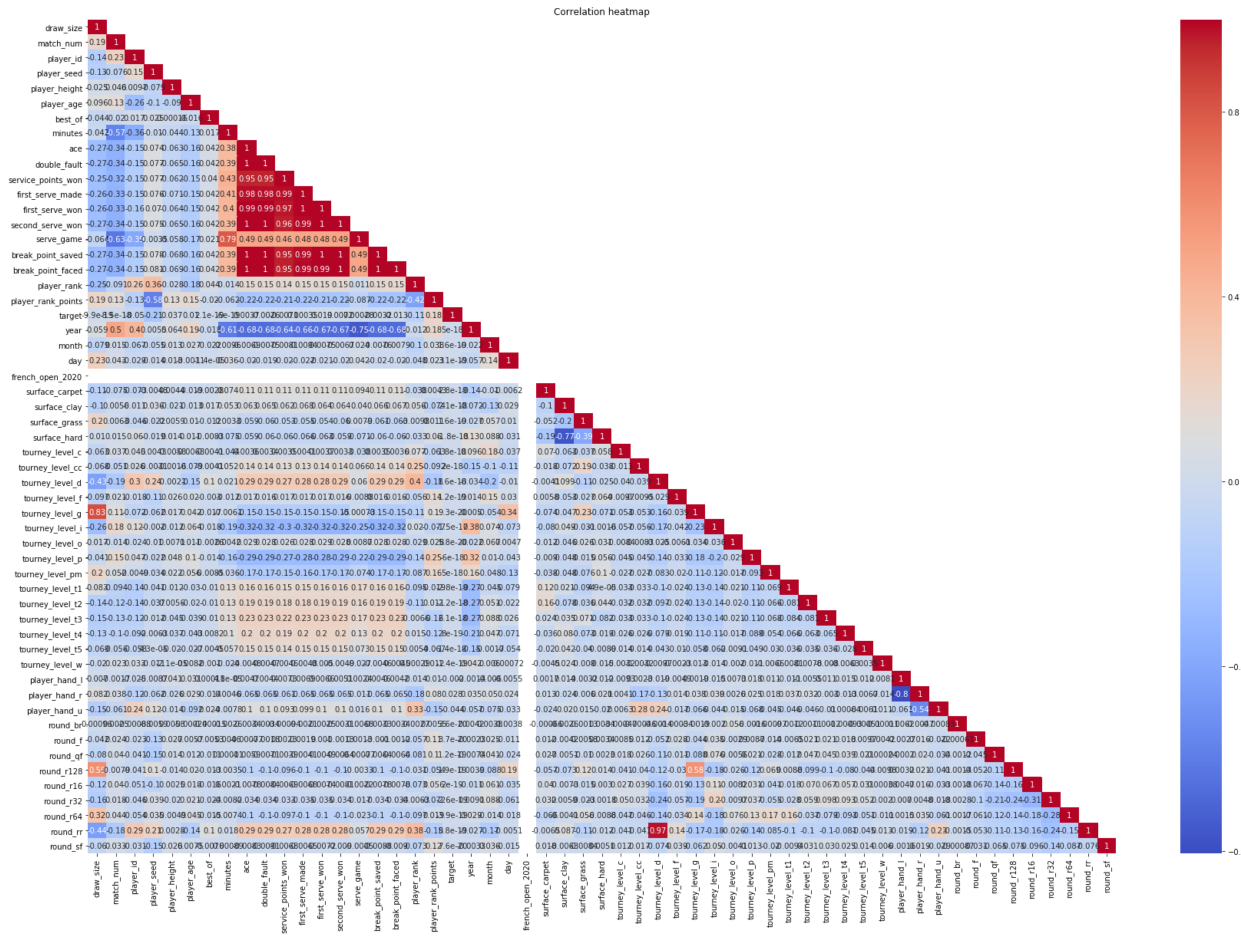
## Feature Importance for WTA


feature_importance
0.10
0.08
0.06
0.04
0.02
0.00
break_point_faced
player_age
player_rank_points
player_rank
first_serve_won
player_id
match_num
day
year
second_serve_won
service_points_won
player_height
player_seed
first_serve_made
break_point_saved
month
ace
double_fault
draw_size
minutes
serve_game
round_r32
round_r16
surface_hard
surface_clay
round_qf
player_hand_r
round_r64
round_sf
surface_grass
player_hand_l
round_r128
tourney_level_i
tourney_level_p
tourney_level_t3
tourney_level_t1
tourney_level_g
round_f
tourney_level_t2
tourney_level_t4
surface_carpet
tourney_level_pm
player_hand_u
round_rr
tourney_level_t5
tourney_level_d
tourney_level_c
tourney_level_o
tourney_level_f
best_of
tourney_level_cc
tourney_level_w
french_open_2020
round_br
features

# ATP Correlation Matrix



# WTA Correlation Matrix



Correlation heatmap