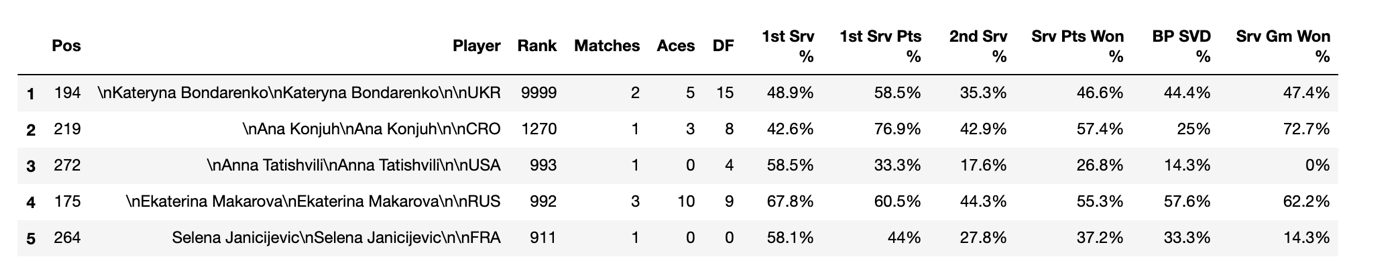
Predicting The Rank of Women’s Tennis Players Based on Their Serve

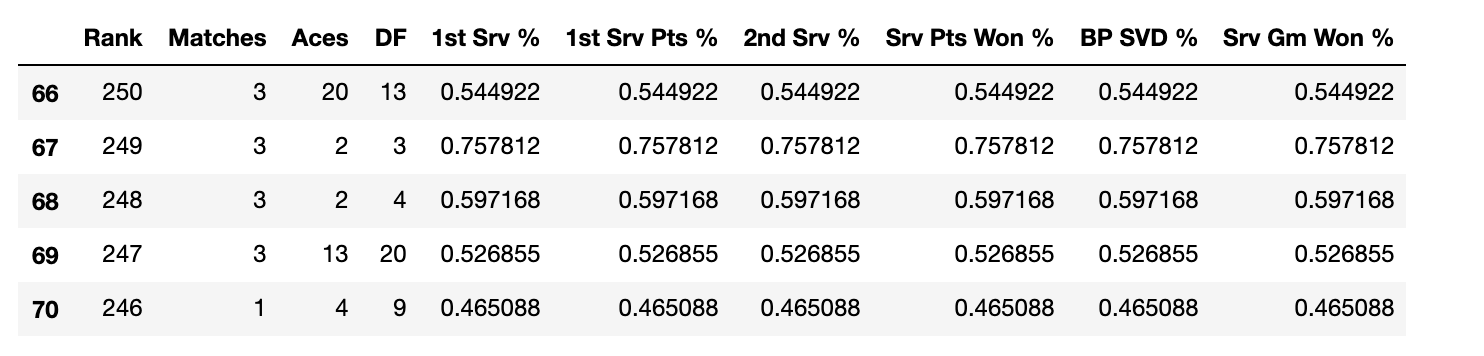
# Introduction

Tennis is one of the world most watched and enjoyed sports. Part of the enjoyment fans have is the ranking system to see who is the best in the world and who is more likely to a win a match. The WTA (Women’s Tennis Association) rankings are founded off a 52 week system where a players gathers points based on their results in tournaments, with up to 16 singles tournaments and 11 double tournaments able to contribute to their points score throughout the year. The different level of tournaments grant different number of points e.g. Grand Slams would be the highest. This system is thought to be highly accurate for estimating the WTA rankings, however as the rankings culminate over a 52 week period, it is unlikely that the rankings in week 1 will be the same as week 52. Therefore, it may be beneficial to be able to predict the end of season rankings using a much shorter time-framed system. In every top level tennis match the serve statistic are always collected and displayed as a metric of a players capability. This study will aim to use these serve statistics across a season to predict the end of season rankings. If predictions are accurate then it would be sensible to take this further and test modelling of individual tournament serve statistics to determine the end of season rankings, however lets walk before we can run.

# Data

As previously stated, the serve statistics for all WTA ranked players are always collected and are readily available for download of the WTA tennis website for each year (<https://www.wtatennis.com/stats>). There are some missing data, however for 1 to 250 the data seems to be almost fully complete, hence only those ranks were taken forward. Figure 1 shows the top 5 rows of the dataset prior to cleaning. The name column for the data contained duplicates in every cell, however as the study is looking at empirical ranks and not specific players, these names can be dropped from the dataset. The position cell is not needed, hence was dropped The dataset contain percentages as ‘strings’ hence these were update to be decimals so that the machine learning algorithm would work. Figure 2 shows the top 5 rows after cleaning. The all the features left after cleaning were determined to be suitable for modelling as they all pertain some information about the matches and serve statistics.





# Exploratory Data Analysis

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedThe easiest way to explore data is to visualise it. Each feature was plotted against the rank to determine if we could see any relationships forming.

A screenshot of a social media post

Description automatically generatedA screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a social media post

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 3 – Exploratory plots

There are a few main takeaways from the plots:

1. A squared relationship between number of matches and Rank
2. A squared relationship between number of aces and Rank
3. A squared relationship between number of double faults and Rank
4. No strong relationship for any of the serve statistics and rank

From the plots it is clear that the statistics that will inherently increase as more matches are played (i.e. Aces & Double faults) have a positive relationship with a better rank. We can therefore infer that the more matches a player plays the higher there rank is likely to be. It is not clear from the serve statistic plots if there is any relationship at all. It could be possible that the relationship is not obvious in a 2D sense and may be revealed when the data is dissected from a machine learning algorithm, therefore is important that the correct algorithm is chosen.

# 4. Machine Learning Methodology

Lets forget about relationships we can see in the data for the time being. We need to think what is our data. The data is numeric, there are no categorical variables. We need to think about what we are trying to do with the data. We are trying to predict a continuous variable by modelling the relationship between other, already measured, variables. Therefore a supervised algorithm such as regression seems sensible. We are aiming to use multiple parameters therefore a multiple linear regression may bee suitable. However, we have seen that there are some squared relationships within our data. The optimal algorithm for this would be a multiple non-linear regression, however that is not within the scope of this course. We could use a simple non-linear regression, however that would not allow us to build multiple parameters into our model. A multiple linear regression seems some-what sensible, therefore will be used for modelling.

In order to model and evaluate the model, the data must be split into a train and test dataset. This will use an 80:20 split. The model coefficients will be derived and then applied on the datasets from which we can evaluate using a variance and residual squares method to determine if the model is any good. From there we can try different combinations of the input features to determine which provides us with the best model.

# Results

Three models were created using different input features. Table 1 summarises these models and their evaluation results. The results show that models 1 and 2 have similar results, with model 2 providing a slightly better variance and residual sum of squares. Model 3 and show poor variances and residual sum of squares. Model 5 shows the best results.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Input Features** | **Variance** | **Residual sum of squares** |
| **1** | Matches, Aces, DF, 1st Srv %, 1st Srv Pts %, 2nd Srv %, Srv Pts Won %, BP SVD %, Srv Gm Won % | 0.67 | 1649.47 |
| **2** | Matches, Aces, DF | 0.68 | 1607.82 |
| **3** | 1st Srv %, 1st Srv Pts %, 2nd Srv %, Srv Pts Won %, BP SVD %, Srv Gm Won % | -0.01 | 4724.81 |
| **4** | Aces, DF | 0.49 | 2377.77 |
| **5** | Matches, DF | 0.73 | 1282.87 |

# Discussion

From the results we can see that the models are not optimal and most have a poor prediction rates. Model 3 provides no correct predictions, hence can be excluded. Models 1 and 2 show us that the Matches, Aces and Double faults are the most useful parameters for prediction. The serve percentage values may not provide any useful information for predicting rank, however none of the models provide high prediction confidence. Model 5 is a simple model and may suggest that the more matches you play with less double faults, the higher your ranking may be This is most likely due to the fact that the serve is such a small part of the sport and that the model should include other parameters such as return statistics, unforced errors, winners etc.

# Conclusions

To conclude, the results show that the regression worked somewhat, however the model prediction confidence is low for most of the models, with model 5 showing average results. We assume this is due to other factors within a tennis match such as return statistics, unforced errors, winners and that further investigation for prediction tennis rankings based on measurable statistics should include more parameters from the wider game.