2013 Grand Slams Tennis Data Interim Report

Önder Türe

2024-05-04

## 1. Brief Statement of the aim of the project

In this project, data is 4 different tennis singles 2013 grand slams match statistics. In the data, there are different statistics about match like foul, number of firs serve win, and others that will be explained in the next part. Aim of the project is predict the result of the tennis match by using different match statistics. Another aim is to find most effective match statistics for the win, and how is change for different tournament and gender. To reach this aim, firstly, data will be cleaned and tidied for the EDA and CDA. Then, explanatory data analysis will be done to see how data distributed and to interpret variables effectively. After EDA, missing values will be handled and data manipulation and feature engineering will be done if it is necessary. After data preparation is done, Confirmatory data analysis will be done by considering EDA. Finally, for statistical modeling to predict match result cross-validation techniques will be used. Finally, statistical modeling will be performed and performance of statistical modeling will be investigated.

## 2. Source of the data variables, dependent variable

Source of the data is [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics). In the source of the data, there is 8 different csv file. Each one represent different tournament from 2013 (Aus Open-Men,AusOpen-Women, USOpen-men,…). Data has 4 different tournament group by gender, so we have 8 different csv file data. All of the 8 csv data has same 42 variables, so all of them are combined to 1 csv file and 2 variable added as tournament and gender. At the end, data has 44 variable. Dependent variable is ‘Result’. Our data have 943 match observations.

Variable Name and Definition

| Name of Variable | Definition of Variable |
| --- | --- |
| Player 1 | Name of Player 1 (Nominal) |
| Player 2 | Name of Player 2 (Nominal) |
| Round | Round of the tourneament (Nominal Ordinal) (1: 1st elimination round, 7: Final of the tournament ) |
| Result | Result of the match (0/1) - Referenced on Player 1 is Result = 1 (Nominal) |
| FSP.1 and .2 | First Serve Percentage for player 1 (if .2, player 2) (Integer) |
| FSW.1 and .2 | First Serve Won by player 1 (if .2, player 2) (Integer) |
| SSP.1 and .2 | Second Serve Percentage for player 1 (if .2, player 2) (Integer) |
| SSW.1 and .2 | Second Serve Won by player 1 (if .2, player 2) (Integer) |
| ACE.1 and .2 | Aces won by player 1 (if .2, player 2) (Integer) |
| DBF.1 and .2 | Double Faults committed by player 1 (if .2, player 2) (Integer) |
| WNR.1 and .2 | Winners earned by player 1 (if .2, player 2) (Integer) |
| UFE.1 and .2 | Unforced Errors committed by player 1 (if .2, player 2) (Integer) |
| BPC.1 and .2 | Break Points Created by player 1 (if .2, player 2) (Integer) |
| BPW.1 and .2 | Break Points Won by player 1 (if .2, player 2) (Integer) |
| NPA.1 and .2 | Net Points Attempted by player 1 (if .2, player 2) (Integer) |
| NPW.1 and .2 | Net Points Won by player 1 (if .2, player 2) (Integer) |
| TPW.1 and .2 | Total Points Won by player 1 (if .2, player 2) (Integer) |
| ST1.1 and .2 | Set 1 result for Player 1 (if .2, player 2) (Integer) |
| ST2.1 and .2 | Set 2 Result for Player 1 (if .2, player 2) (Integer) |
| ST3.1 and .2 | Set 3 Result for Player 1 (if .2, player 2) (Integer) |
| ST4.1 and .2 | Set 4 Result for Player 1 (if .2, player 2) (Integer) |
| ST5.1 and .2 | Set 5 Result for Player 1 (if .2, player 2) (Integer) |
| FNL.1 and .2 | Final Number of Games Won by Player 1 (if .2, player 2) (Integer) |
| tournament | Tournament name |
| gender | Tournament gender |

## 3 Data Cleaning and Tidying

Check variables and data types:

kable(str(GrandSlam))

## 'data.frame': 943 obs. of 44 variables:  
## $ Player1 : chr "Lukas Lacko" "Leonardo Mayer" "Marcos Baghdatis" "Dmitry Tursunov" ...  
## $ Player2 : chr "Novak Djokovic" "Albert Montanes" "Denis Istomin" "Michael Russell" ...  
## $ Round : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Result : int 0 1 0 1 0 0 0 1 0 1 ...  
## $ FNL.1 : int 0 3 0 3 1 1 2 2 0 3 ...  
## $ FNL.2 : int 3 0 3 0 3 3 3 0 3 2 ...  
## $ FSP.1 : int 61 61 52 53 76 65 68 47 64 77 ...  
## $ FSW.1 : int 35 31 53 39 63 51 73 18 26 76 ...  
## $ SSP.1 : int 39 39 48 47 24 35 32 53 36 23 ...  
## $ SSW.1 : int 18 13 20 24 12 22 24 15 12 11 ...  
## $ ACE.1 : int 5 13 8 8 0 9 5 3 3 6 ...  
## $ DBF.1 : int 1 1 4 6 4 3 3 4 NA 4 ...  
## $ WNR.1 : int 17 13 37 8 16 35 41 21 20 6 ...  
## $ UFE.1 : int 29 1 50 6 35 41 50 31 39 4 ...  
## $ BPC.1 : int 1 7 1 6 3 2 9 6 3 7 ...  
## $ BPW.1 : int 3 14 9 9 12 7 17 20 7 24 ...  
## $ NPA.1 : int 8 NA 16 NA 9 6 14 6 5 NA ...  
## $ NPW.1 : int 11 NA 23 NA 13 12 30 9 14 NA ...  
## $ TPW.1 : int 70 80 106 104 128 108 173 78 67 162 ...  
## $ ST1.1 : int 3 6 4 6 6 3 6 6 5 3 ...  
## $ ST2.1 : int 6 6 5 6 4 2 6 6 4 6 ...  
## $ ST3.1 : int 1 6 4 6 6 6 7 NA 1 3 ...  
## $ ST4.1 : int NA NA NA NA 2 6 3 NA NA 6 ...  
## $ ST5.1 : int NA NA NA NA NA NA 3 NA NA 6 ...  
## $ FSP.2 : num 68 60 77 50 53 63 60 54 67 60 ...  
## $ FSW.2 : int 45 23 57 24 59 60 66 26 42 68 ...  
## $ SSP.2 : int 32 40 23 50 47 37 40 46 33 40 ...  
## $ SSW.2 : int 17 9 15 19 32 22 34 13 14 25 ...  
## $ ACE.2 : int 10 1 9 1 17 24 2 0 12 8 ...  
## $ DBF.2 : int 0 4 1 8 11 4 6 11 NA 12 ...  
## $ WNR.2 : int 40 1 41 1 59 47 57 11 32 8 ...  
## $ UFE.2 : int 30 4 41 8 79 45 72 46 20 12 ...  
## $ BPC.2 : int 4 0 4 1 3 4 10 2 7 6 ...  
## $ BPW.2 : int 8 0 13 7 5 7 17 6 10 14 ...  
## $ NPA.2 : int 8 NA 12 NA 16 14 25 8 8 NA ...  
## $ NPW.2 : int 9 NA 16 NA 28 17 36 12 11 NA ...  
## $ TPW.2 : int 101 42 126 79 127 122 173 61 94 141 ...  
## $ ST1.2 : int 6 1 6 2 1 6 3 3 7 6 ...  
## $ ST2.2 : int 7 3 7 2 6 6 7 2 6 3 ...  
## $ ST3.2 : int 6 1 6 3 7 3 6 NA 6 6 ...  
## $ ST4.2 : int NA NA NA NA 6 7 6 NA NA 3 ...  
## $ ST5.2 : int NA NA NA NA NA NA 6 NA NA 4 ...  
## $ tournament: chr "AusOpen" "AusOpen" "AusOpen" "AusOpen" ...  
## $ gender : chr "M" "M" "M" "M" ...

Head and tail of the data

## Player1 Player2 Round Result FNL.1 FNL.2 FSP.1 FSW.1 SSP.1  
## 1 Lukas Lacko Novak Djokovic 1 0 0 3 61 35 39  
## 2 Leonardo Mayer Albert Montanes 1 1 3 0 61 31 39  
## 3 Marcos Baghdatis Denis Istomin 1 0 0 3 52 53 48  
## 4 Dmitry Tursunov Michael Russell 1 1 3 0 53 39 47  
## 5 Juan Monaco Ernests Gulbis 1 0 1 3 76 63 24  
## 6 Santiago Giraldo Sam Querrey 1 0 1 3 65 51 35  
## SSW.1 ACE.1 DBF.1 WNR.1 UFE.1 BPC.1 BPW.1 NPA.1 NPW.1 TPW.1 ST1.1 ST2.1 ST3.1  
## 1 18 5 1 17 29 1 3 8 11 70 3 6 1  
## 2 13 13 1 13 1 7 14 NA NA 80 6 6 6  
## 3 20 8 4 37 50 1 9 16 23 106 4 5 4  
## 4 24 8 6 8 6 6 9 NA NA 104 6 6 6  
## 5 12 0 4 16 35 3 12 9 13 128 6 4 6  
## 6 22 9 3 35 41 2 7 6 12 108 3 2 6  
## ST4.1 ST5.1 FSP.2 FSW.2 SSP.2 SSW.2 ACE.2 DBF.2 WNR.2 UFE.2 BPC.2 BPW.2 NPA.2  
## 1 NA NA 68 45 32 17 10 0 40 30 4 8 8  
## 2 NA NA 60 23 40 9 1 4 1 4 0 0 NA  
## 3 NA NA 77 57 23 15 9 1 41 41 4 13 12  
## 4 NA NA 50 24 50 19 1 8 1 8 1 7 NA  
## 5 2 NA 53 59 47 32 17 11 59 79 3 5 16  
## 6 6 NA 63 60 37 22 24 4 47 45 4 7 14  
## NPW.2 TPW.2 ST1.2 ST2.2 ST3.2 ST4.2 ST5.2 tournament gender  
## 1 9 101 6 7 6 NA NA AusOpen M  
## 2 NA 42 1 3 1 NA NA AusOpen M  
## 3 16 126 6 7 6 NA NA AusOpen M  
## 4 NA 79 2 2 3 NA NA AusOpen M  
## 5 28 127 1 6 7 6 NA AusOpen M  
## 6 17 122 6 6 3 7 NA AusOpen M

## Player1 Player2 Round Result FNL.1 FNL.2 FSP.1 FSW.1 SSP.1 SSW.1  
## 938 S.Stephens M.Bartoli 5 0 0 2 63 24 37 5  
## 939 A.Radwanska N.Li 5 1 2 1 77 52 23 9  
## 940 S.Lisicki K.Kanepi 5 1 2 0 59 26 41 10  
## 941 M.Bartoli K.Flipkens 6 1 2 0 61 21 39 10  
## 942 S.Lisicki A.Radwanska 6 1 2 1 63 53 37 19  
## 943 S.Lisicki M.Bartoli 7 0 0 2 65 22 35 9  
## ACE.1 DBF.1 WNR.1 UFE.1 BPC.1 BPW.1 NPA.1 NPW.1 TPW.1 ST1.1 ST2.1 ST3.1  
## 938 3 3 21 19 8 4 10 8 NA 4 5 NA  
## 939 0 5 32 18 16 5 32 21 NA 7 4 6  
## 940 3 4 23 19 5 4 12 9 NA 6 6 NA  
## 941 5 3 23 10 7 5 11 11 NA 6 6 NA  
## 942 9 7 60 46 14 6 44 32 NA 6 2 9  
## 943 6 5 21 25 8 2 18 11 NA 1 4 NA  
## ST4.1 ST5.1 FSP.2 FSW.2 SSP.2 SSW.2 ACE.2 DBF.2 WNR.2 UFE.2 BPC.2 BPW.2  
## 938 NA NA 58 26 42 15 0 3 6 12 12 6  
## 939 NA NA 77 57 23 13 1 4 58 40 10 4  
## 940 NA NA 67 25 33 9 3 3 13 23 2 1  
## 941 NA NA 49 8 51 7 0 1 10 6 2 1  
## 942 NA NA 70 42 30 15 1 0 21 10 14 6  
## 943 NA NA 67 31 33 7 2 6 15 14 13 5  
## NPA.2 NPW.2 TPW.2 ST1.2 ST2.2 ST3.2 ST4.2 ST5.2 tournament gender  
## 938 10 4 NA 6 7 NA NA NA Wimbledon F  
## 939 71 48 NA 6 6 2 NA NA Wimbledon F  
## 940 19 9 NA 3 3 NA NA NA Wimbledon F  
## 941 21 8 NA 1 2 NA NA NA Wimbledon F  
## 942 31 16 NA 4 6 7 NA NA Wimbledon F  
## 943 11 9 NA 6 6 NA NA NA Wimbledon F

Character values, “Round” and “Result” should be factor.

Also ‘Result’ and ‘Round’ should be factor.