80 / 100

Önder Türe - 2170215

2013 Grand Slams Tennis Data Analysis Interim Report

## 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. To reach this aim, firstly, data will be cleaned and tidied for the EDA and CDA. Then, explanatory data analysis and confirmatory data analysis will be done to see how data distributed and to interpret variables effectively. Then, missing values will be handled and data manipulation and feature engineering will be done if it is necessary. Finally, for statistical modeling to predict match result cross-validation techniques will be used. And, statistical modeling will be performed and performance of statistical modeling will be investigated.

## 2. Source of The Data, Variables and 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. Variable descriptions can be seen on [APPENDİX A](#appendix-a). Dependent variable is ‘Result’. Our data have 943 match observations.

## 3. Data Cleaning and Tidying

Check variables and data types for character variables and some variables that should be factor. Others are already numeric type as desired.

## 'data.frame': 943 obs. of 6 variables:  
## $ Result : int 0 1 0 1 0 0 0 1 0 1 ...  
## $ 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 ...  
## $ tournament: chr "AusOpen" "AusOpen" "AusOpen" "AusOpen" ...  
## $ gender : chr "M" "M" "M" "M" ...

Character values, “Round” and “Result” should be a factor. And name of the variables should be all upper or lower letter. So, all names of variables changed to upper letter.

## [1] "PLAYER1" "PLAYER2" "ROUND" "RESULT" "FNL.1"   
## [6] "FNL.2" "FSP.1" "FSW.1" "SSP.1" "SSW.1"   
## [11] "ACE.1" "DBF.1" "WNR.1" "UFE.1" "BPC.1"   
## [16] "BPW.1" "NPA.1" "NPW.1" "TPW.1" "ST1.1"   
## [21] "ST2.1" "ST3.1" "ST4.1" "ST5.1" "FSP.2"   
## [26] "FSW.2" "SSP.2" "SSW.2" "ACE.2" "DBF.2"   
## [31] "WNR.2" "UFE.2" "BPC.2" "BPW.2" "NPA.2"   
## [36] "NPW.2" "TPW.2" "ST1.2" "ST2.2" "ST3.2"   
## [41] "ST4.2" "ST5.2" "TOURNAMENT" "GENDER"

Before change character values, “Round” and “Result” to factor let’s check head and tail of the data set. There is 44 variable, so visualizing the head and tail is not feasible. Therefore, just problematic variable parts of the data shown.

Head and tail of the PLAYER1 and PLAYER2:

| PLAYER1 | PLAYER2 |
| --- | --- |
| Lukas Lacko | Novak Djokovic |
| Leonardo Mayer | Albert Montanes |
| Marcos Baghdatis | Denis Istomin |

| PLAYER1 | PLAYER2 |
| --- | --- |
| M.Bartoli | K.Flipkens |
| S.Lisicki | A.Radwanska |
| S.Lisicki | M.Bartoli |

As seen on the “PLAYER1” and “PLAYER2”, there is problem about how player names saved to data. In some rows player names saved as full name (first name + surname). However, in other rows, player names saved as shortened way like “M.Bartolli”. This problem fixed by convert all names shortened.

After fixed player names, there is 358 player on this data set.

| PLAYER1 | PLAYER2 |
| --- | --- |
| L.Lacko | N.Djokovic |
| L.Mayer | A.Montanes |
| M.Baghdatis | D.Istomin |

Except player name columns, data seems well tabulated data. There is no unnecessary columns. There is 0 duplicated observations in the data set. Player names fixed too.

Now character values, “Round” and “Result” can be changed to factor. And finalize structure of the data is:

Also there is problem about BPC(Break Points Created) and BPW(Break Points Win). Logically, BPC must be greater or equal to BPW but in some rows BPC is lower than BPW like below example. This fixed by exchanging BPC values by BPW and BPW values by BPC when BPW > BPC.

| BPC.1 | BPW.1 | BPC.2 | BPW.2 |
| --- | --- | --- | --- |
| 1 | 3 | 4 | 8 |
| 7 | 14 | 0 | 0 |
| 1 | 9 | 4 | 13 |

There are similar problems with NPA and NPW for Aus open and French Open tournaments, NPA should be greater or equal to NPW. This fixed by exchanging NPA values by NPW and NPW values by NPA when NPW > NPA.

|  | NPA.1 | NPW.1 | NPA.2 | NPW.2 |
| --- | --- | --- | --- | --- |
| 1 | 8 | 11 | 8 | 9 |
| 3 | 16 | 23 | 12 | 16 |
| 5 | 9 | 13 | 16 | 28 |

After fix the problem, all BPC values are greater than BPW. Same for NPA and NPW.

## 4. Exploratory Data Analysis(EDA) and Confirmatory Data Analysis(CDA)

### Summary Statistics

Below there are frequency tables for tournament, round, result and gender. Ausopen and French open tournaments have most number of match. Also, as expected most matches played at round 1. Number of male matches are greater than female matches. Moreover, Most round 1 matches played at US open, but at the remaining round US Open matches are not much when considering other tournaments. There may be incomplete data for US Open.

|  | TOURNAMENT | ROUND | RESULT | GENDER |
| --- | --- | --- | --- | --- |
|  | AusOpen :253 | 1:536 | 1:465 | F:452 |
|  | FrenchOpen:252 | 2:197 | 2:478 | M:491 |
|  | USopen :202 | 3:107 | NA | NA |
|  | Wimbledon :236 | 4: 55 | NA | NA |
|  | NA | 5: 27 | NA | NA |
|  | NA | 6: 14 | NA | NA |
|  | NA | 7: 7 | NA | NA |

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| AusOpen | 128 | 64 | 31 | 16 | 8 | 4 | 2 |
| FrenchOpen | 127 | 63 | 32 | 16 | 8 | 4 | 2 |
| USopen | 153 | 20 | 16 | 7 | 3 | 2 | 1 |
| Wimbledon | 128 | 50 | 28 | 16 | 8 | 4 | 2 |

Summary statistics for numeric variables are below. Average performances are almost equal for player 1 and player 2 statistics. Average performance of tennis players can be seen in this summary statistics. Some of them are given below. When we look some statistics like ACE.1 and ACE.2, seem like there is outlier values but each match has different number of games, so if the number of games is too high, then number of some statistics can be high.

|  | FSP.1 | FSP.2 | FSW.1 | FSW.2 | ACE.1 | ACE.2 | BPC.1 | BPC.2 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min. :38.00 | Min. : 0.00 | Min. : 3.00 | Min. : 0.00 | Min. : 0.00 | Min. : 0.000 | Min. : 0.000 | Min. : 0.000 |
|  | 1st Qu.:57.00 | 1st Qu.:57.00 | 1st Qu.: 26.00 | 1st Qu.: 25.00 | 1st Qu.: 2.00 | 1st Qu.: 1.000 | 1st Qu.: 5.000 | 1st Qu.: 5.000 |
|  | Median :62.00 | Median :62.00 | Median : 37.00 | Median : 36.00 | Median : 4.00 | Median : 4.000 | Median : 9.000 | Median : 8.000 |
|  | Mean :62.26 | Mean :61.97 | Mean : 38.79 | Mean : 38.69 | Mean : 6.14 | Mean : 5.983 | Mean : 8.792 | Mean : 8.648 |
|  | 3rd Qu.:67.00 | 3rd Qu.:67.00 | 3rd Qu.: 49.00 | 3rd Qu.: 48.50 | 3rd Qu.: 8.00 | 3rd Qu.: 9.000 | 3rd Qu.:12.000 | 3rd Qu.:12.000 |
|  | Max. :86.00 | Max. :93.00 | Max. :109.00 | Max. :132.00 | Max. :41.00 | Max. :39.000 | Max. :28.000 | Max. :26.000 |
|  | NA | NA | NA | NA | NA’s :22 | NA’s :22 | NA’s :1 | NA’s :1 |

Also, set result summary is like below, as seen, there are a lot of NA values for ST4 and ST5 especially. This is because most of the match is not last to 5 or 4 set. It is over in 2 or 3 sets.

|  | ST1.1 | ST2.1 | ST3.1 | ST4.1 | ST5.1 |
| --- | --- | --- | --- | --- | --- |
|  | Min. :0.000 | Min. :0.000 | Min. :0.000 | Min. :0.000 | Min. :0.000 |
|  | 1st Qu.:3.000 | 1st Qu.:4.000 | 1st Qu.:3.000 | 1st Qu.:3.000 | 1st Qu.:3.000 |
|  | Median :6.000 | Median :6.000 | Median :6.000 | Median :6.000 | Median :5.000 |
|  | Mean :4.801 | Mean :4.781 | Mean :4.728 | Mean :4.787 | Mean :4.651 |
|  | 3rd Qu.:6.000 | 3rd Qu.:6.000 | 3rd Qu.:6.000 | 3rd Qu.:6.000 | 3rd Qu.:6.000 |
|  | Max. :7.000 | Max. :7.000 | Max. :9.000 | Max. :7.000 | Max. :9.000 |
|  | NA’s :1 | NA’s :4 | NA’s :321 | NA’s :722 | NA’s :857 |

### Research Questions

#### How does number of break points win (BPW) and break points created (BPC) corralate between 2 player according to result?

As seen in the scatter plot, there is correlation between result and break points win. If player has more break points than opponent player, then, player that has more BPW and BPC probably win the match.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

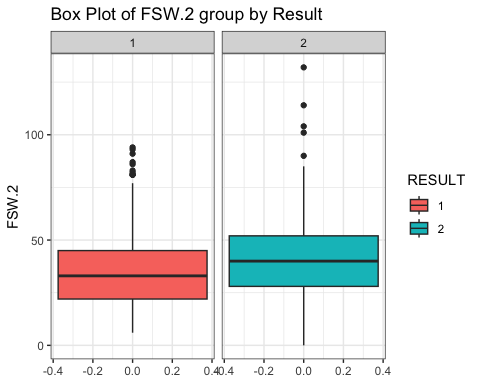
Hypothesis t-test can be conduct to show avarage BPC or BPW is greater for winning player.

##   
## Two Sample t-test  
##   
## data: BPC.1.P1win and BPC.2.P1win  
## t = 16.758, df = 928, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 4.285722 Inf  
## sample estimates:  
## mean of x mean of y   
## 11.210753 6.458065

H0 is BPC.1 is equal or less than BPC.2 for RESULT is equals 1. H1 is BPC.1 is greater than BPC.2 for RESULT is equals 1. From t test confidence interval is 4.2857224, to infinite. So we can reject null hypothesis. Mean of BPC.1 is greater than mean of BPC.2 for RESULT equals 1. Similarly, mean of BPW.1 is greater than mean of BPW.2 for RESULT equals 1 according to t-test. This situation is valid for RESULT equals 2 too. t-test results does not shown here to simplicity of document, they can be seen on r code file.

#### How does change distribution of first serve percentage (FSP) and First serve winning (FSW) for winning and losing player?

As seen on the density plot, when player win the match, player’s first serve percentage is increase. Therefore, first serve percentage can be effective for winning the match. Similarly, box plot of FSP show that winner player has more FSW than loser one. So, having more FSW can affect the result.

metin, diyagram, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

For confirmatory data analysis for this reasearch question, two paired t-test can be conduct too. As seen on the t-test result, mean of FSP.1 is changed significantly when result change. t-tests for FSP.2 and FSW.2 are like FSP.1 and FSP.2, t-test results can be seen on r code file.

## [1] "Two-sample t-test for First Serve Percentage (FSP.1):"

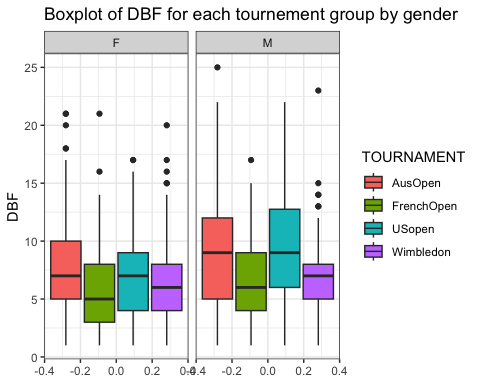
##   
## Two Sample t-test  
##   
## data: fsp1\_P1win and fsp1\_P2win  
## t = 3.756, df = 941, p-value = 0.0001833  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 0.9190557 2.9303976  
## sample estimates:  
## mean of x mean of y   
## 63.23226 61.30753

## [1] "Two-sample t-test for First Serve Winning (FSW.1):"

##   
## Two Sample t-test  
##   
## data: fsw1\_P1win and fsw1\_P2win  
## t = 3.6426, df = 941, p-value = 0.0002847  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1.901268 6.342994  
## sample estimates:  
## mean of x mean of y   
## 40.87527 36.75314

#### How does distribution of number of double faul (DBF) and aces (ACE) change for different tournaments grouped by gender?

metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturulduOn this part, ACE variable is created by summing of ACE.1 and ACE.2. Similarly other variables are summing and total of player 1 and player 2 statistic variables are created. Aim of this research is investigating effect of different tournament on player performance. Each tournament have different surface type so this should be effect the match statistics. As seen in the boxplots, at French open tournament, players do less ace than other tounaments. Clay surface (French open surface) can lead to this, because clay surface slow down the ball. Also, in the boxplot of DBF, we can say that players do more double foul at the US and AUS open tournaments. Both tournaments has hard surfaces. Additionally, there is significant diffrence on male and female statistics for ACE, male can do more ACE than womwn according to this graph but number of sets are not same for male and female matches. Female matches are best of 3, but male matches are best of 5. So, this can be reason for that.

With analysis of variance (ANOVA) for DBF and ACE, We can see there are significant difference between tournaments and gender for 0.05 significance level.

## [1] "ANAVO of DBF"

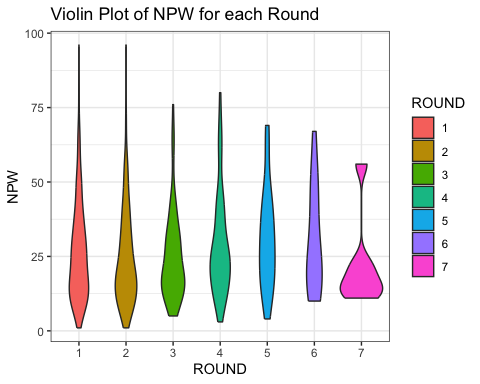
## Df Sum Sq Mean Sq F value Pr(>F)   
## TOURNAMENT 3 1134 378.1 24.13 4.82e-15 \*\*\*  
## GENDER 1 310 310.0 19.78 9.72e-06 \*\*\*  
## Residuals 930 14572 15.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 8 observations deleted due to missingness

## [1] "ANAVO of ACE"

## Df Sum Sq Mean Sq F value Pr(>F)   
## TOURNAMENT 3 4114 1371 24.5 2.96e-15 \*\*\*  
## GENDER 1 36744 36744 656.5 < 2e-16 \*\*\*  
## Residuals 916 51265 56   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## 22 observations deleted due to missingness

#### Are there differences in the net points attempt (NPA) and net points win (NPW) by players across different rounds of the tournament by result?

metin, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturulduNPA and NPW has right skewed distribution for each round except 7th round, but in 7th round we have 8 observation so it can mislead the interpretation of this plot for 7th round. Therefore player NPA and NPW performances are similar for all rounds except 7th.

Again by using ANOVA, we can show there is significant difference or not in NPW and NPA. As seen in ANOVA, for 0.05 significance level, there is no significant difference between NPA and NPW observations of different rounds.

## [1] "ANAVO of NPA"

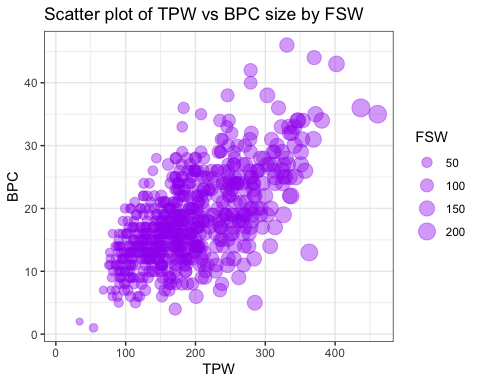
## Df Sum Sq Mean Sq F value Pr(>F)  
## ROUND 6 5553 925.5 1.558 0.156  
## Residuals 852 506175 594.1   
## 84 observations deleted due to missingness

## [1] "ANAVO of NPW"

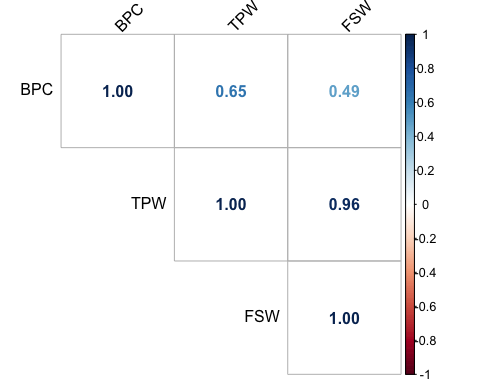
## Df Sum Sq Mean Sq F value Pr(>F)  
## ROUND 6 2253 375.6 1.453 0.191  
## Residuals 853 220433 258.4   
## 83 observations deleted due to missingness

#### Is there any correlation between number of break points created (BPC), total points win (TPW) and first serve win (FSW)?

As seen on bubble plot, whwn BPC increase, TPW increase because when player break the opponent player serve, match can be extend. Also number of first serve win tend to increase when TPW and BPC increase.



For CDA, we can check correlations between BPC, TPW and FSW. As seen on correlation plot,



Also, checking p values of correlation test show that there is significant correlation between BPC, TPW and FSW for 0.05 significance level.

## cor\_test p\_values  
## 1 BPC\_TPW 4.643419e-76  
## 2 BPC\_FSW 1.610577e-49  
## 3 TPW\_FSW 0.000000e+00

## 5. Missing Observation Cleaning, Imputation and Data Manipulation

Firstly, we should find NA values. As we look at the number of NA observations of variables that include NA (below), we can see there are 5338 NA values

## FNL.1 FNL.2 ACE.1 DBF.1 WNR.1 UFE.1 BPC.1 BPW.1 NPA.1 NPW.1 TPW.1 ST1.1 ST2.1   
## 1 1 22 8 126 126 1 1 84 83 312 1 4   
## ST3.1 ST4.1 ST5.1 ACE.2 DBF.2 WNR.2 UFE.2 BPC.2 BPW.2 NPA.2 NPW.2 TPW.2 ST1.2   
## 321 722 857 22 8 126 126 1 1 84 83 312 1   
## ST2.2 ST3.2 ST4.2 ST5.2   
## 4 321 722 857

For, FNL.1 and FNL.2 NA observation are because of cancelled match. At the [link](https://www.tennisabstract.com/cgi-bin/wplayer-classic.cgi?p=AlizeCornet&f=A2014qqC2), you can see this match (date 13.01.2014) is cancelled. This row is removed from data set.

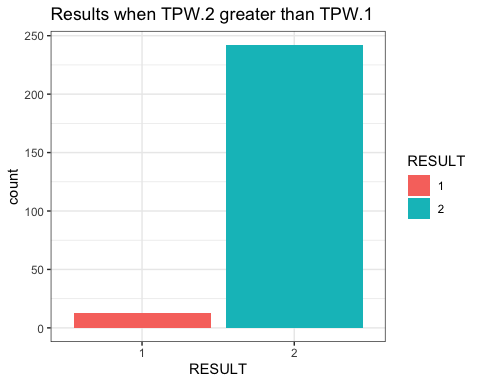
There is 22 NA in ACE.1 and ACE.2 columns at same observations, this is MCAR type missing observations, we can fill this values by median value of ACE. Median of ACE.1 and ACE.2 are 4 and 4.

Similarly DBF NA values are MCAR, we can fill them by median value of DBF.1 and DBF.2. Median of DBF.1 and DBF.2 are 3 and 3.

UFE and WNR NA observations belong the same rows. Those NA values are in male US Open tournament. There is no UFE and WNR observation for male US open tournament. This is Missing at Random mechanism. Rows that has NA UFE and WNR values removed from data set because this variables can be effective to result and removing the column may not be good idea.

After removing NA UFE and WNR values, There are 46 and 45 NA values for NPA and NPW are MCAR case because there is no any pattern. Those NA values can be filled by median of NPA and NPW values.

There are 312 NA values for total points win (TPW) observations. As seen in the bar graphs, when player’s TPW is greater than oppenent player, this player wins the match, it is almost like result. Therefore TPW is totaly dependent to result. TPW variable should be removed from data set. Similarly, FNL is directly give the result, so we should remove FNL variables too.



Set scores (ST) are similar to TPW, if we know 2 players set score, match result can be known. Therefore ST values should be removed from data set. However, number of total game (TG) variable can be created from sum of all ST values for each row. Number of total game can be use for standardize the match statistics. For example we can find number of ace per game from ACE.1 over number of total game. So, after created TG, all ST values can be removed. FSW, SSW, ACE, DBF, WNR, UFE, BPC, BPW, NPA, NPW variables are divided to TG to standardize match statistics.

There is 350 unique player on this data set. So using players as predictor is not good idea because there are 350 levels, this is not convenient for modeling. Also, our aim in this project is predict result by using player performance like ACE, BPC etc., and show the most important statistics. Because of that PLAYER1 and PLAYER2 columns are removed from data set.

After missing data cleaning and imputation, we have 816 observations and 28 variables.

## Statistical Modelling

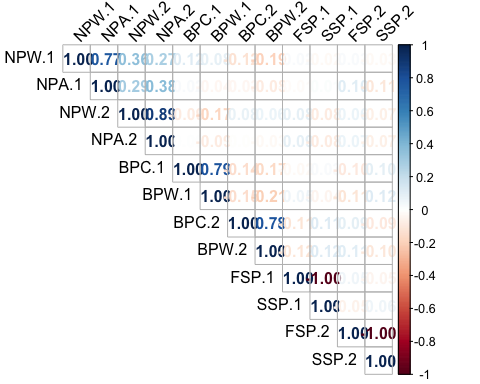
### Train-Test Data Set Preparation

GrandSlam data set splited by 20% test and 80% training by random selection. Before modelling, numeric variables in the train and test data set should be scaled. Firstly, train set is scaled and then, by using train set mean and standart deviation test set should be scaled too. Test and train sets are scaled after splitting, because, test observations should not be effect anything on the train data set in order to maintain integrity of model evaluation.

### Logistic Regression

Logistic regression is used to predict the RESULT of the tennis matches on this project. As a cross validation method, k-fold cv is used for k equals 10. This cv is made by using “caret” package in r. Our dependent variable should be binary. So, RESULT is binary with 1 or 2.

Independence: predictors should be independent from eachother, We can check that by checking correlation between numeric predictors. As seen on the correlation plot. There are high correlation between NPW and NPA (0.77), BPC and BPW (0.79), FSP and SSP(1.0). So, NPW, BPW and SSP variables are removed from data set.



Now, logistic regression can be applied.

FSW, SSW and BPC statistics are more effective than other to predict winner. Also, p-values show that FSW, SSW and BPC are highly significant predictors. FSW and BPC influence can be seen in EDA part too. That shows that, player can practice more serve shoots to increase the win rate. Also, model indicate that breaking the opponent serve with fast shoots is really effective on win rate.

On the other hand, ACE.1 and UFE.2 seems marginally significant and can be effective to predict result.

##   
## Call:  
## NULL  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.35344 0.65675 -0.538 0.5905   
## FSP.1 -0.71382 0.55058 -1.296 0.1948   
## FSW.1 -3.31050 0.54717 -6.050 1.45e-09 \*\*\*  
## SSW.1 -2.29100 0.50632 -4.525 6.05e-06 \*\*\*  
## ACE.1 -0.75834 0.40777 -1.860 0.0629 .   
## DBF.1 0.37041 0.32017 1.157 0.2473   
## WNR.1 -0.61818 0.42202 -1.465 0.1430   
## UFE.1 0.41812 0.41340 1.011 0.3118   
## BPC.1 -3.53518 0.53104 -6.657 2.79e-11 \*\*\*  
## NPA.1 0.33799 0.32144 1.051 0.2930   
## FSP.2 -0.33636 0.52969 -0.635 0.5254   
## FSW.2 4.28693 0.69301 6.186 6.17e-10 \*\*\*  
## SSW.2 2.27867 0.48378 4.710 2.47e-06 \*\*\*  
## ACE.2 0.02546 0.35475 0.072 0.9428   
## DBF.2 -0.15095 0.33655 -0.449 0.6538   
## WNR.2 0.62334 0.39991 1.559 0.1191   
## UFE.2 -0.52951 0.31712 -1.670 0.0950 .   
## BPC.2 2.37091 0.48352 4.903 9.42e-07 \*\*\*  
## NPA.2 -0.35142 0.34789 -1.010 0.3124   
## ROUND2 -0.47880 0.54270 -0.882 0.3776   
## ROUND3 -0.29845 0.73599 -0.406 0.6851   
## ROUND4 -0.31759 0.91067 -0.349 0.7273   
## ROUND5 -1.80490 1.26076 -1.432 0.1523   
## ROUND6 0.13923 2.52302 0.055 0.9560   
## ROUND7 -1.09237 18.49824 -0.059 0.9529   
## TOURNAMENTFrenchOpen -0.23430 0.61739 -0.380 0.7043   
## TOURNAMENTUSopen 1.93073 1.09716 1.760 0.0784 .   
## TOURNAMENTWimbledon 0.18404 0.64878 0.284 0.7767   
## GENDERM 0.69909 0.78145 0.895 0.3710   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 905.06 on 652 degrees of freedom  
## Residual deviance: 156.42 on 624 degrees of freedom  
## AIC: 214.42  
##   
## Number of Fisher Scoring iterations: 9

You can find vif matrix of logistic regression at appendix b, There are no higher than 10 vif value. So, there is no highly correlated variables that lead multicollinearity. There is confusion matrix for test data and performance statistics for test and train data. Accuracy is 0.9 for test data which is acceptable and effective result. However, train data has 0.95 accuracy because logistic regression model made by train data. So, there are some overfitting. Other performance parameters are good as well. Other train data performance parameter are greater than test data too.

Confusion Table of Test Data

| 1 | 2 |
| --- | --- |
| 64 | 5 |
| 10 | 84 |

Test data Performance of Logistic Regression. Train data Performance of Logistic Regression

| Metric | Value |
| --- | --- |
| Accuracy | 0.9079755 |
| Precision | 0.9275362 |
| Recall | 0.8648649 |
| F1-Score | 0.8951049 |

| Metric | Value |
| --- | --- |
| Accuracy | 0.9555896 |
| Precision | 0.9522388 |
| Recall | 0.9608434 |
| F1-Score | 0.9565217 |

## APPENDİCES

### APPENDİX A

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 |

### APPENDİX B

VIF matrix of logistic regression:

## FSP.1 FSW.1 SSW.1   
## 6.386922 6.273041 6.721242   
## ACE.1 DBF.1 WNR.1   
## 4.044830 1.942179 4.057837   
## UFE.1 BPC.1 NPA.1   
## 3.975883 4.313217 2.484020   
## FSP.2 FSW.2 SSW.2   
## 5.454639 9.565770 5.163032   
## ACE.2 DBF.2 WNR.2   
## 2.894532 2.029614 2.909925   
## UFE.2 BPC.2 NPA.2   
## 3.107296 2.842413 2.426543   
## ROUND2 ROUND3 ROUND4   
## 1.428437 1.411339 1.256708   
## ROUND5 ROUND6 ROUND7   
## 1.516416 1.053369 1.001458   
## TOURNAMENTFrenchOpen TOURNAMENTUSopen TOURNAMENTWimbledon   
## 1.824837 1.653996 2.203210   
## GENDERM   
## 3.690153