Classsification of 2013 Grand Slam Tennis Tournaments Result

Önder Türe   
Middle East Technical University *Ankara, Türkiye*tureonder@gmail.com

*Abstract*—This paper mainly focusses on result classification of 2013 grand slam dataset by using different modeling techniques and performance comparison of models. Moreover, result of best model is investigated to make interpretation. Also, Data investigation and missingness procedures are reported.

Keywords—logistic regression, classification, svm, xgboost, neural network, random forest

# Introduction

In this project, dataset is 4 different tennis singles 2013 grand slams match statistics. In the data, there are different statistics about match like foul, number of first 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 are cleaned and tidied for the explanatory data analysis. Then, EDA and CDA are applied to see how data distributed and to interpret variables effectively. Then, missing values and data manipulation are handled. In order to apply classification modeling, cross-validation techniques are used and modeling assumptions are checked. And, 5 different classification model are performed and test performance of models are investigated.

# Methodology

## 2013 Grand Slam Dataset

Source of the data is [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/dataset/300/tennis+major+tournament+match+statistics) [1]. 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 Appendix A. Dependent variable is ‘Result’. Our data have 943 match observations.

Before giving descriptive statistics and EDA, data are cleaned and tidied. 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.

TABLE I Head and Tail

|  |  |  |  |
| --- | --- | --- | --- |
| Head of The Data | | Tail of The Data | |
| PLAYER1 | PLAYER2 | PLAYER1 | PLAYER2 |
| Lukas Lacko | Novak Djokovic | M.Bartoli | K.Flipkens |
| Leonardo Mayer | Albert Montanes | S.Lisicki | A.Radwanska |
| Marcos Baghdatis | Denis Istomin | S.Lisicki | M.Bartoli |

Character values, "Round" and "Result" are converted to the factor type.

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. 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.

## Descriptive Statistics

Below there are frequency tables for tournament, round, result and gender. Result distribution is almost 50-50, so , there is balanced data for classification. 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.

TABLE II Frequecy Table For Categorical Variables

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

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.

TABLE III Numeric Variable Summary

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Metric | FSP.1 | FSP.2 | BPC.1 | BPC.2 | ACE.1 | ACE.2 |
| Min | 38.00 | 0.00 | 0.000 | 0.000 | 0.00 | 0.000 |
| 1.Q | 57.00 | 57.00 | 5.000 | 5.000 | 2.00 | 1.000 |
| Median | 62.00 | 62.00 | 9.000 | 8.000 | 4.00 | 4.000 |
| Mean | 62.26 | 61.97 | 8.792 | 8.648 | 6.14 | 5.983 |
| 3.Q | 67.00 | 67.00 | 12.000 | 12.000 | 8.00 | 9.000 |
| Max | 86.00 | 93.00 | 28.000 | 26.000 | 41.00 | 39.000 |

## Exlanatory Data Analysis (EDA) with Confirmatory Data Analysis (CDA)

### How does number of 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 oluşturuldu

Figure I Scatter Plot of BPC.1 and BPC.2 by Result

To approve this situation. T-test may be conduct, but BPC is not normally distributed (p<<0.05). Therefore, non-parametric test mann-whitney u test applied, and test show that mean of the BPC.1 is greater than BPC.2 (p<0.05) for result equal 1.

### How does change distribution of number of First serve win (FSW) for winning and losing player?

Box plot of FSW show that winner player has more FSW than loser one. So, having more FSW can affect the result.

metin, ekran görüntüsü, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure II Box Plot of FSW.1 group by Result

For CDA, mann-whitney u test applied again because FSW data are not normally distributed (p << 0.05). Test shoe that FSW.1 is greater than FSW.2 for result equal 1.

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

NPA and NPW has right skewed distribution for each round except 7th round, but in 7th round we have 8 observations 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.

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

Açıklama otomatik olarak oluşturuldu

Figure III Violin Plot of NPA for each Round

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

Açıklama otomatik olarak oluşturuldu

Figure IV Violin Plot of NPW for each Round

To approve this EDA, ANOVA can be conduct but, NPA and NPW are not normally distributed (p<<0.05). Non-parametric Kruskal-Wallis test is conducted to investigate difference between rounds. Test show that there is no statistically significant difference for NPA and NPW for different rounds (p>0.05).

### 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, when BPC increase, TPW increase because when player break the opponent player serve, match can be extended. Also, number of first serve win tend to increase when TPW and BPC increase.

metin, ekran görüntüsü, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure V Scatter Plot of TPW vs BPC size by FSW

For CDA, we can check correlations between BPC, TPW and FSW. As seen on correlation plot, There are correlation between those 3 variable.

metin, ekran görüntüsü, diyagram, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

Also, checking p values (p<<0.05) of correlation test show that there is significant correlation between BPC, TPW and FSW.

## Missingness and Data Manipulation.

Number of NA values are shown below table. There are 5338 NA value in the data.

TABLE IV # of NA Values

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

For, FNL.1 and FNL.2 NA observation are because of cancelled match. This row is removed from data set.

There are 312 NA values for total points win (TPW) observations. As seen in the bar graphs, when player’s TPW is greater than opponent player, this player wins the match, it is almost like result. Therefore, TPW is totally 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.

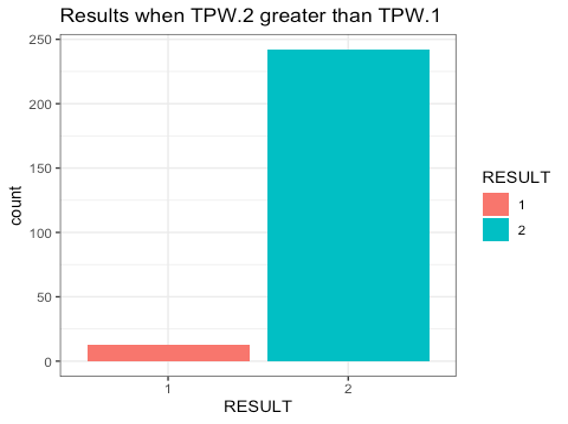


Figure VI Results when TPW.2 greater than TPW.1

There are a lot of NA values on set scores (ST) because all matches are not last 5 sets, so there are NA values naturally. Also, set scores are totally dependent with result as expected. Therefore, set scores are 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 games can be used for standardizing the match statistics. For example, we can find number of aces per game from ACE.1 over number of total games. So, after created TG, all ST values are removed. FSW, SSW, ACE, DBF, WNR, UFE, BPC, BPW, NPA, NPW variables are divided to TG to standardize match statistics. There are 358 unique different players on the data set, this is too much for modeling, so player columns are removed from data set.

ACE, DBF, NPW and NPA NA values are MCAR (Missing completely at random) mechanism. All of them are imputated by using mice package in r.

At the end of NA imputation, change of distribution of variables are checked by using Kolmogorov-Smirnov (K-S) test. Test show that there is no statistically significant change on the variable distributions.

## Modelling

Before applying models, GrandSlam data set split 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 standard 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.

Also, 5-fold cross validation are used for tuning hyperparameter processes for all models.

### Logistic Regression

To apply logistic regression, predictors should be independent from each other, 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.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure VII Correlation Plot Of Variables

After removing correlated variables, logistic regression is applied, and in the first logistic regression FSW, SSW, BPC, WNR and UFE variables are significant. So other predictors are removed from dataset, and logistic regression is applied again.

TABLE V Logistic Regression Summary

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Estimate Coef. | Std. Error | P-value |
| FSW.1 | -3.3222 | 0.4456 | 8.93e-14 |
| SSW.1 | -1.5061 | 0.2746 | 4.12e-08 |
| BPC.1 | -3.2159 | 0.4142 | 8.26e-15 |
| WNR.1 | -0.8519 | 0.3250 | 0.00876 |
| UFE.1 | 0.6404 | 0.3074 | 0.03722 |
| FSW.2 | 3.6181 | 0.4606 | 4.00e-15 |
| SSW.2 | 2.1677 | 0.3389 | 1.59e-10 |
| BPC.2 | 2.2941 | 0.4304 | 9.84e-08 |
| WNR.2 | 0.3612 | 0.2918 | 0.21581 |
| UFE.2 | -0.4994 | 0.2203 | 0.02341 |

The expected decrease in log-odds is -3.32 for one unit increase in FSW.1, FSW.2 has reverse effect as expected. Other effects can be seen on the TABLE V, estimated coefficients column show that effect of variables on the log-odds for result equal 2.

Test accuracy for logistic regression model is 0.926, also test data performance is 0.95.

### Support Vector Machine

### Random Forest

### Artificial Neural Network

### XgBoost

## Performance Comparison with Test Data

# Results

# Conclusion

# References

# APPENDICES

## Appendix A.

TABLE VI Variables

| 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 |