Candidate No.	2206664
Year:	3
Course Code :	MN3515
Course Tutor:	Dr Catherine Harbor
Assignment No.:	2
Degree Title:	BSc Economics and Management
Title:	2206664_MN3515

Business Understanding

This research paper focuses on the effect of play features in women's tennis on the outcome of a match. Tennis is now being examined, and data is being used to guide the process. The data has been acquired from a legal data source - UCI Machine Learning Repository. The project is intelligible with a comprehensible explanation and the quality of the data is well suited for answering the research question to the extent of ensuring generalizability. In tennis, people compete by using not only their own abilities but also a specific set of skills that contribute to increasing their chances of winning. The main objective of this report is to find out what it takes one player to win. Translating this into a data mining goal gives the objective of identifying important factors of the game that would affect positively or negatively the outcome. The execution of the project plan consists of implying a logistic regression on a binary dependent variable, using decision tree techniques for classification, and generating a random forest to compare performance across models.

Data Understanding

The primary source of the data which consists of multiple .xls file. It was obtained from https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics. The research is focused only on the first three major ladies' tournaments because they are played on the three unique surfaces in tennis. Each row in the datasets AusOpen-women-2013.csv, FrenchOpen-women-2013.csv, and Wimbledon-women-2013.csv represents a single match, and each column represents player's information and matches statistics. After an initial view of the data, The French Open and Australian Open data sets consist of 127 observations and 44 variables, but Wimbledon has 122 observations and 44 variables. It was later found out that the raw data was missing out on five Round-2 matches. In addition, it was missing data points on two variables – TPW.1 and TPW.2. This data could only be added if one is to calculate all points for every match which is time-consuming and complex. The structure shows that most variables are factors or time-series numbers. The 44 variables are split between the features of play of the two players playing each round. The dependent variable 'Result' is binary. A score of 1 is displayed when Player 1 won and a score of 0 when Player 2 won.

The hexagon graph helps with visualizing a possible relationship between two variables in the Australian Open data set. The relationship between the winners made by the two players in a match is tested, as this feature is one of the major measurements of match outcome. The .cor function displays

a correlation coefficient of 0.40, suggesting a positive correlation. The number of examples in each category is counted, and the size of the points is scaled using this number.

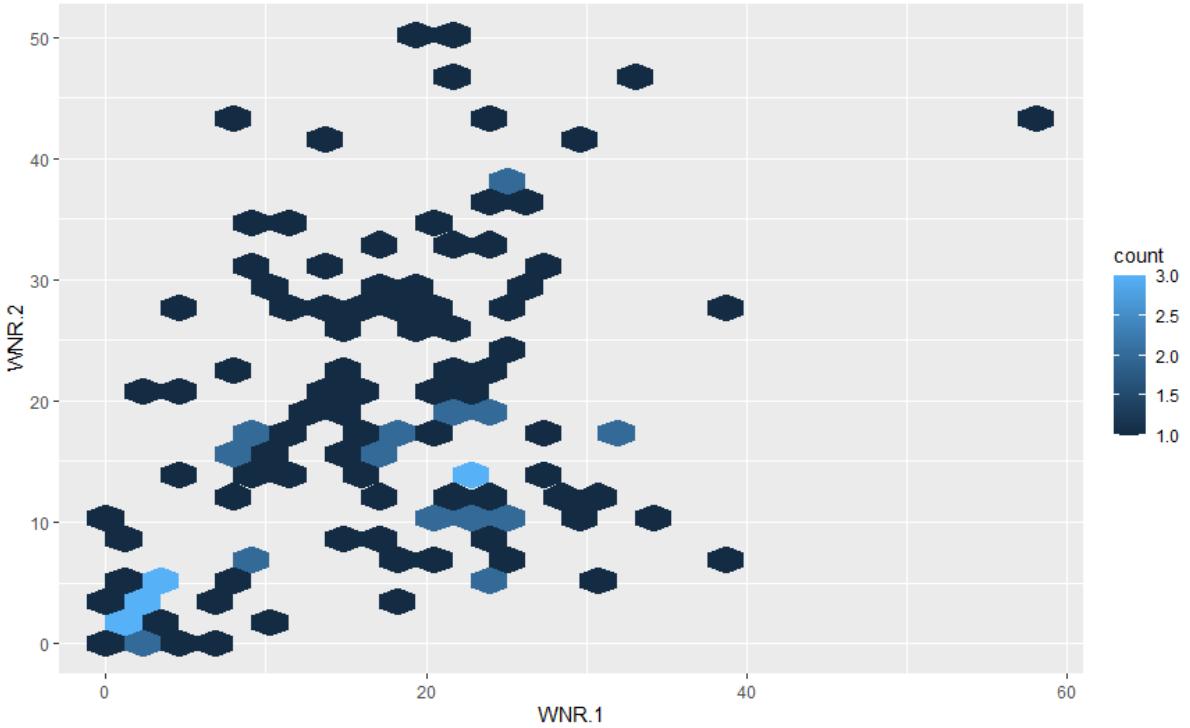

Figure 1.

The next graph plots how the relationship between winners made by Player 1 and Player 2 varies throughout the tournament. Scatter plot number 7 shows that the winner of this tournament was Player 1 and the number of winners made by her is much bigger than Player 2. This implies that this feature of play might influence the outcome of the match.

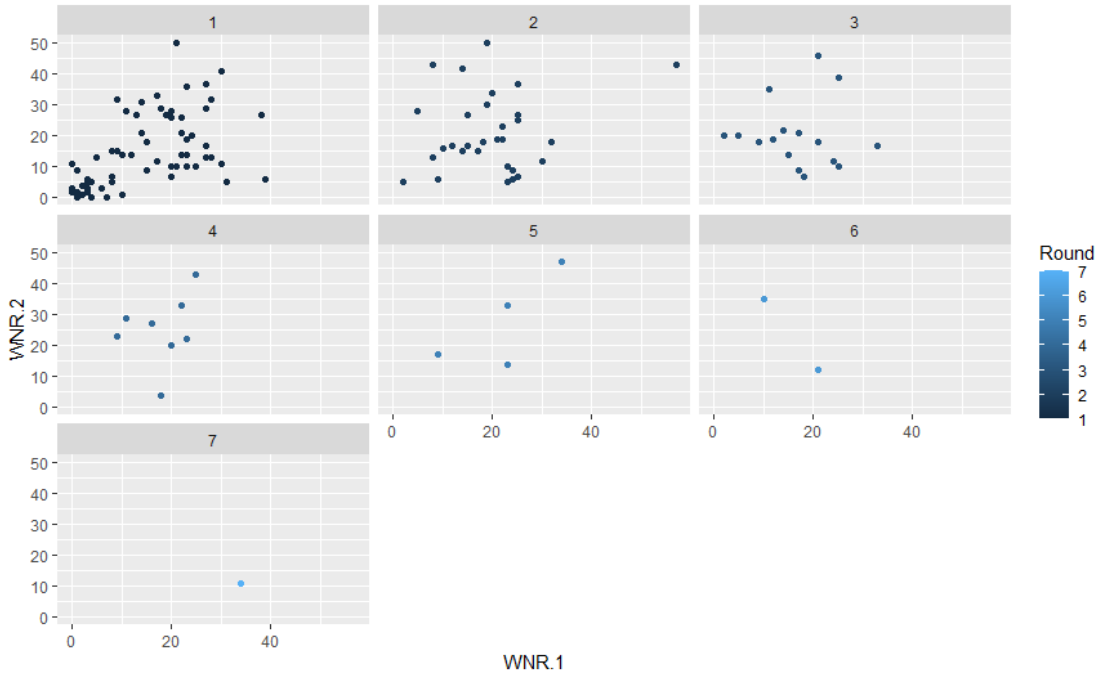

Figure 2.

However, the final of Wimbledon provides some interesting insight. The female player who won the final had made fewer winners than the one who lost. The 7th scatter plot below shows that WNR.1 is higher than WNR.2, thus this element of the game may not always guarantee the win and there are other aspects that must be taken into consideration.

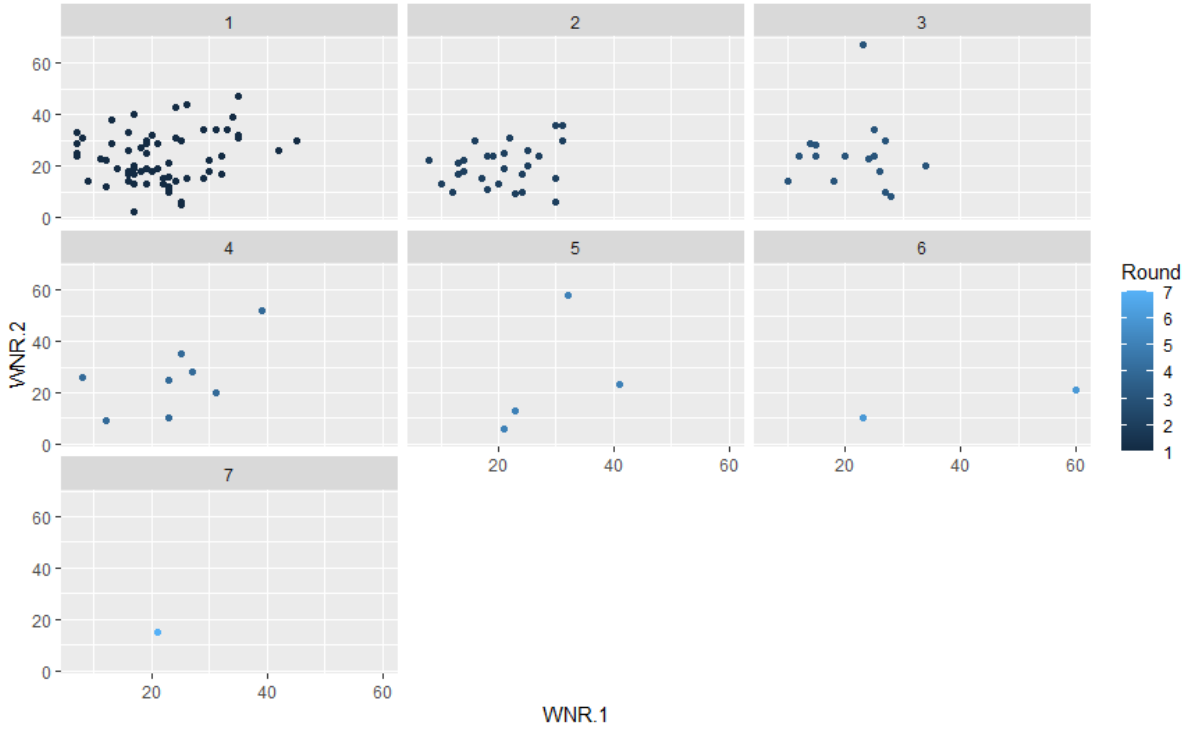

Figure 3.

Data Preparation

Having investigated the nature of the data, the next step is preparing the data. The data is prepared by taking basic data checks, dealing with missing values in the data sets, adding an additional variable, dropping unnecessary variables, combining the data sets, and splitting the data into a training and a testing set. The data was loaded correctly into RStudio and it was assessed that there is no need for transforming variables or changing their type. The additional variable that has been added is 'Surface'. It is considered an important factor because different play styles fit the different surfaces. Variables FNL and ST1-5 were removed because they act as an equivalent to the result of the match by calculating the winner. The report aims to provide generalised results, therefore variables 'Player 1' and 'Player 2' were cut off as well. The number of null variables found in the merged data set is 388, meaning that the better approach to solving this problem was to fill in the missing data points based on average values. It has been assessed that multiple imputation would be the best approach of the three available as most observations would have had to be erased.

Modelling

Trying to find the best way to represent and model the data was not met with much difficulties. There was no overlap between the modelling stage and the data cleaning stage when attempting to find the best way to represent it and the best form in which to model it because the data is more suitable for building a logistic regression model, rather than a linear regression model. In this case, the dependent variable is 'Result' and it takes the value of 1 for 'Player 1' winning the match and 0 for 'Player 2' winning. The task in this case is to identify when Player 1 wins and look at the variables that follow

this pattern which leads to a classification problem. Two approaches were used for analysing - logistic regression and building a decision tree. The pros of using a decision tree are the following:

- It is efficient and easy to understand
- The splitting is easier to interpret because of using a binary dependent variable;

Cons:

- The results after pruning the tree were not efficient as it produced the same ones as before.

The pros of using logistic regression:

- It deals with the classification problem
- It is suitable for the essence of the data.

Cons:

- The data has too many variables and it could not fit the model at first.

Evaluation

The first step toward evaluation was to compute the correlation amongst the independent variables of the training set. After inspecting the variables, taking the most significant ones out of the 29 available and gradually dropping them until the AIC started increasing again, this represents the final variables in the logistic regression and their coefficients. It turns out that break points won, unforced errors made, and total points by a player are most likely to make a player win, with total points won being the most significant one.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.38655    1.87175  -2.344 0.019101 *
BPC.2        0.09867    0.18341   0.538 0.590599
BPW.2        0.39179    0.19975   1.961 0.049841 *
TPW.1        0.60582    0.16385   3.698 0.000218 ***
TPW.2       -0.60108    0.16967  -3.543 0.000396 ***
UFE.1       -0.09514    0.06328  -1.504 0.132704
UFE.2        0.12901    0.06120   2.108 0.035035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 260.623  on 187  degrees of freedom
Residual deviance:  29.798  on 181  degrees of freedom
AIC: 43.798

Number of Fisher Scoring iterations: 9

> |
```

Figure 4.

Building the decision tree proved the logistic model. The pruned decision tree tells a lot about what variables are the most significant when it came to winning the match. In the first play aspect of the game, break points tend to put a lot of pressure on the opponent and breaking your opponent would result in one game being won. It is interesting to point out that the tree also includes unforced errors made. So far key aspects are features of the play where the player has a chance to break their opponent or wait for them to make an unforced error. However, a player is more likely to win when they generate a higher portion of total points won than their opponent. However, it is important to note down that after pruning the tree, it didn't make any significant changes. The best logistic regression model had an accuracy of 0.6941176. The baseline model always predicts Reverse, has an accuracy of 0.5 and the pruned CART model beats neither with an accuracy of 0.4840426.

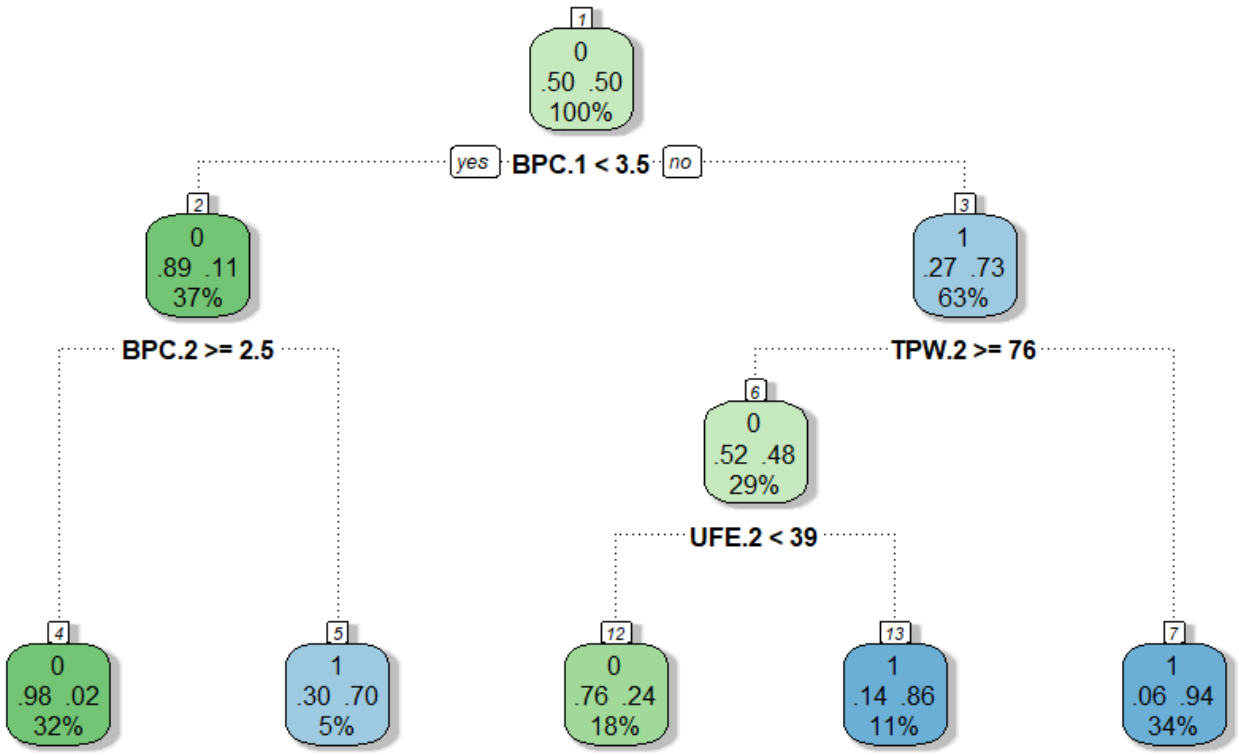

Figure 5.

Another method used to evaluate the model is a random forest. This method shows once again the high significance of total points won, break points created/won and unforced errors made by their opponent made to win a tennis match. One interesting finding is that winners made by a player didn't turn out to be as important as stated in the beginning of this report.

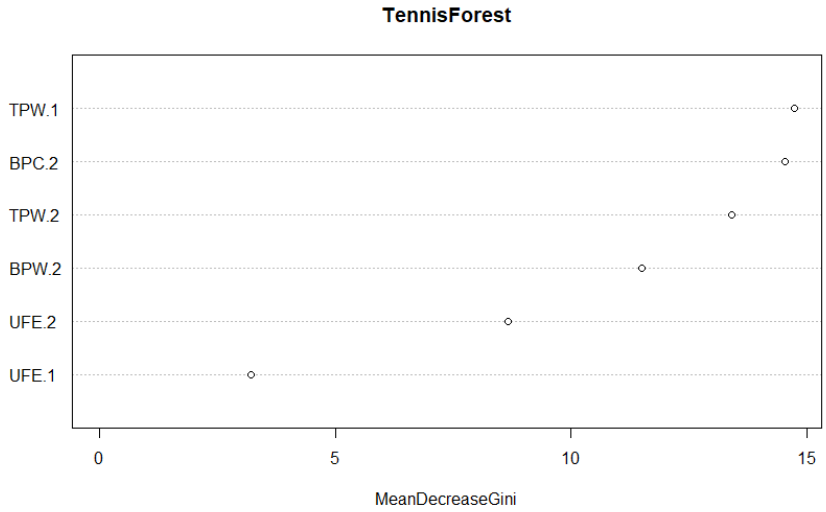

Figure 6.

Deployment

On both the ATP and WTA tours, the majority of the Top 20 players now work with analytics as it is becoming more standardized (). The WTA's relationship with SAP offers coaches with increasingly specific analytics that may be used during on-court coaching across the women's tour. In addition, private organisations have sprung up to give players, federations, and college teams with even more video and data. It could strengthen the partnership between the association and the software company as SAP will have access to more data from the WTA and this would result in players receiving more insightful analysis of their game and competition. Further improvement could be made on this classification problem. Additional research on data sets from the tournaments held in previous years in order to answer for generalisability in the features of play and decrease to an extent the uncertainty in the measures and give better predictor estimates.

Reference List:

Tandon, K., 2020. *THE ROLE OF ANALYTICS IN TENNIS IS ON A LONG, SLOW RISE*. [online] Available at: <https://www.tennis.com/news/articles/the-role-of-analytics-in-tennis-is-on-a-long-slow-rise> [Accessed 22 April 2022].