**IST 707 Final Project: Using Game Data to Predict the Winner of the Rugby World Cup 2023**

**Introduction**

The Rugby World Cup (RWC) is an international tournament that is held every four years where the top 20 international men’s rugby union teams compete to be the world rugby champion. Twelve of the teams qualify based on their previous Rugby World Cup performance while the remaining eight teams are decided in region-based qualification games. The tournament began in 1987 and since then there have only been four winners: New Zealand and South Africa with three championships, Australia winning twice, and England winning once.

The winners of the tournament are awarded the Webb Ellis Cup, named after the person who invented rugby William Webb Ellis, who according to popular legend got frustrated during a football game, picked up the ball and bulldozed through his opponents beginning the sport known as rugby. The accolade of world champion is sought after by all international teams and is highly respected to the extent that only those who have won the Webb Ellis Cup are allowed to physically touch the trophy, a tradition that expresses the reverence within rugby culture.

Respect is pervasive in the culture of the sport, stemming from the recognition of the fortitude and work ethic of not only one’s teammates but also one’s opponents. It’s the reason why the winner of the championship is revered within the international teams, as only those who perform at the highest level of the sport can earn the honor. Winners of the championship are truly regarded as the best in the world for a period of four years until the next world cup is held and the title is up for retaining or relinquishing.

The rewards of winning go further than just respect of the peers and players within rugby, but the tournament being an international event brings ancillary prestige and awards to countries who win. The winning country not only receives a monetary prize but being the third largest sporting event in the world, the winners receive international exposure only comparable to the Olympics and the FIFA World Cup. International teams who participate in the tournament prepare for four years for the sole purpose of winning and those who miss qualifying make it their goal to qualify for the next tournament as all teams and fans recognize the prestige that comes with winning the World Rugby Championship.

**Analysis and Models**

**About the Data:**

The data comes from Kaggle and contains a record of all men’s international rugby union games since 1871 and goes through to mid-2023. The data frame contains 2727 observations (rows) and 11 variables (columns) where each row is a game or match, and the columns contain information with respect to each game. The match information in the columns of the data frame are Date corresponding to the date of the rugby match, Home Team, Away Team, Home Score, Away Score, Competition corresponding to the tournament or friendly game the match was under, Stadium which is the venue of the match, City corresponding to the city where the match was held, Country corresponding to the country the city and stadium is located, Neutral (True/ False) a logical column that states if the game was on neutral ground, and World Cup (True/ False) a logical column that states if the match was a apart of a previous world cup.

Table 1: First Five Rows of the Initial Data Set

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| date | home\_team | away\_team | home\_score | away\_score | competition | stadium | city | country | neutral | world\_cup |
| 1871-03-27 | Scotland | England | 1 | 0 | 1871 Scotland v England International | Raeburn Place | Edinburgh | Scotland | FALSE | FALSE |
| 1872-02-05 | England | Scotland | 2 | 1 | 1871-72 Home Nations International | The Oval | London | England | FALSE | FALSE |
| 1873-03-03 | Scotland | England | 0 | 0 | 1872-73 Home Nations International | West of Scotland F.C. | Glasgow | Scotland | FALSE | FALSE |
| 1874-02-23 | England | Scotland | 1 | 0 | 1873-74 Home Nations International | The Oval | London | England | FALSE | FALSE |
| 1875-02-15 | England | Ireland | 2 | 0 | 1874-75 Home Nations rugby union matches | The Oval | London | England | FALSE | FALSE |

Additional columns were added to improve the analysis where the first column called Winning Side, which determines the winner from the home or away team using those columns and the home and away score columns to determine which side won. The second column added is called Winning Team which uses the same columns as before to determine the name of the team who won the match. The third column added to the data is called Outcome which combines the previous two columns as a concise summary. Finally, the last column added was called Winning Margin, which calculates the score difference between the winning team and the losing team.

Table 2: Additional Columns Added with First Five Rows of the initial Data Set

|  |  |  |  |
| --- | --- | --- | --- |
| winning\_side | winning\_team | outcome | winning\_margin |
| home | Ireland | Ireland-home | 7 |
| home | Wales | Wales-home | 42 |
| home | France | France-home | 7 |
| away | England | England-away | 7 |
| home | Ireland | Ireland-home | 10 |

When viewing the distribution of games across time, most of the matches occur from the late 1900’s to today.

Figure 1: Distribution of International Games per Year

A graph of a number of blue and white bars

Description automatically generated

Another aspect when trying to predict the outcome of a sports tournament is only the most recent data matters for an accurate analysis. This is because the rules can change, the effective strategies change and the players in the team are changing over time, all of which affect the performance of the team. In order to get the most accurate analysis out of the data to determine the top performer which likely becomes the champion of the Rugby World Cup the data was focused on matches since the last RWC reducing the number of rows to 151 representing the matches since RWC 2019. This was done after initial analysis on the historical data where the data was found to be useful in determining the top performing team historically but not useful for predicting the upcoming winner of the current RWC tournament (2023).

Figure 2: Distribution of International Games per Year since Previous RWC

A graph of blue rectangular bars

Description automatically generated

To further reinforce why the data was focused on data since the last RWC, the team’s winning records since 2000 and the last RWC can be compared. When looking at the winning teams since the turn of the century, New Zealand seems to be the top performing team followed by England, South Africa and Australia. When looking at the winning teams since the last RWC, the top performing teams are France followed by Ireland and then New Zealand providing a different view of the performance of each team. This can be seen in the first Association Rule Minning models because if the data was focused on all the historical data or the data since the turn of the century, one would expect New Zealand, England, South Africa and Australia to be the top performing teams, but when focusing the data since the last RWC, the recent performance of teams like Fance, New Zealand, and Ireland can be seen giving a more complete understanding of each team’s performance.

Figure 3: Winning Record of each International Team Since 2000

A graph of blue rectangular bars

Description automatically generated

Figure 4: Winning Record of each International Team Since RWC 2019

A graph of blue bars

Description automatically generated

**Models:**

The models used in the analysis are Association Rule Mining, Classification using Naïve Bayes, and Decision Trees. Association rule mining (ARM) is used for discovering interesting relationships or associations between items in a data frame. In this analysis, ARM is used to try and see if any factors contribute to teams winning. Naïve Bayes is used to determine if instances in the data to predefined classes or categories. In this analysis, Naïve Bayes was used to see if the data would be ideal for determining the winning team as the target variable with the deciding data features being winning margins, winning side, home team, away team, home score, and away score. Lastly, Decision Trees use tree-like models to make predictions based on input features. In this analysis, Decision Trees are used to try and predict the winner of teams within the dataset using winning team as the target variable with the deciding data features being winning side, home team, away team, home score, and away score.

**Association Rule Mining:**

The first analysis using Association Rules Mining used just the data since the turn of the century to find relationships in the teams who won their matches. For the model, the columns world cup, neutral, country, city, stadium, home team and away team were removed so the outcome, only the winning side at home/away could be focused on. The top rule found was that the teams won at home with the teams with the most wins at home being New Zealand, England, South Africa and then Australia.

Figure 5: Association Rules Mining Plot using Data since 2000

A diagram of a network

Description automatically generated

Figure 6: Item Frequency Plot from the Association Rule Minning with Data since 2000

A graph of a number of people

Description automatically generated with medium confidence

Another analysis using Association Rule Minning was performed to see if there were any teams that performed well away. The difference between the first and second model is the winning team column was removed to further focus on the outcome which gives a more comprehensive view of which rules affect the overall result of a match. It was found that New Zealand stood out as an outlier for winning away almost as much as winning at home.

Figure 7: Association Rules Mining Plot using All Data since 2000 Teams Winning Away

A group of words and lines

Description automatically generated

Figure 8: Item Frequency Plot from the Association Rule Minning RHS = Winning Away with Data since 2000

A line of words and symbols

Description automatically generated with medium confidence

For reasons discussed earlier, the data was focused on matches since the last RWC. The next Association Rule Minning models using the most recent data also tell a different story than the data since the turn of the century. The top rule even in recent years since the last RWC found that the home team still won more than the away team, but it also found that France and Ireland performed the best both at home and away closely followed by Wales and New Zealand. It was also found that even at home Italy lost to the away team as an outlier to teams won at home more than away.

Figure 9: Association Rules Mining Plot using Data Since Last RWC

A network of words and lines

Description automatically generated with medium confidence

Figure 10: Association Rules Mining Plot using Data Since Last RWC

A row of grey squares with black text

Description automatically generated

In a similar attempt to finding which teams performed well away from home, the right-hand side of the model was set to find rules where the winning side was the away team. The support and confidence parameters were reduced to 0.05 and 0.01 respectively to find rules. Again, Italy lost to away teams more than any other team, but both France and New Zealand are shown to perform well away with Ireland being a close third.

Figure 11: Association Rules Mining Plot using Data Since Last RWC Winning Team Away

A group of words and a circle

Description automatically generated with medium confidence

Figure 12: Association Rules Mining Plot using Data Since Last RWC Winning Team Away

A row of grey squares with black text

Description automatically generated

**Naïve Bayes Analysis:**

The Naïve Bayes analysis began using the data since the last RWC, with creating a data frame using the winning team as the target variable with the deciding data variables being winning margins, winning side, home team, away team, home score, and away score. The data set contained 151 observations and 6 variables, the data set was split into a training set and testing set where the training set contained 66% of the data and the remaining went to the test set. A correlation heat map was also created to investigate if the numerical data had any correlation. Only a slight correlation was found but nothing above 0.5.

Table 3: Data Used in the Naïve Bayes Model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| winning\_team | home\_team | away\_team | home\_score | away\_score | winning\_margin |
| Ireland | Ireland | Scotland | 19 | 12 | 7 |
| Wales | Wales | Italy | 42 | 0 | 42 |
| France | France | England | 24 | 17 | 7 |
| England | Scotland | England | 6 | 13 | 7 |
| Ireland | Ireland | Wales | 24 | 14 | 10 |

Figure 13: Correlation Heat Map using Numeric Data in Naïve Bayes

A diagram of a heat map

Description automatically generated

The initial results of the Naïve Bayes model showed an accuracy of 53.33% when trying to predict the winning team when using data such as who the home team is, who the away team is, the scores for each team and the point difference. Laplace smoothing was introduced to the model and when the Laplace variable was set to a value of 1 the model had an improved accuracy of 60%. To try and further improve the model’s accuracy, a 10-fold cross validation was performed where the Laplace value was sequenced from 0.1 to 1.0 by 0.1. The cross-validation model using the best tunning parameters returned the same accuracy 53.33% showing that using who the home team is, who the away team is, the scores for each team and the point difference between the winning and losing sides gives us about 53.33% prediction power in determining the winning team in the data since the last RWC. More match data could improve the model, such as the number of penalties, cards (yellow/red cards) handed out for extreme infractions, or injuries before or during the match.

**Decision Trees:**

A decision tree model was used to try and predict the winning team using the home team, away team, home score, away score, winning side and winning margin from the data since the last RWC. The data was split using a training set of 66% of the data and a testing set of the remaining data. The first tree model had an overall accuracy of 48.9% where France’s winning record established France as the root node where the branches in the determine the opposition and if they were home or away. For example, in the root node, if France were to play Ireland at home, then going to the left down the branch Ireland would win 11% of the time, and if France were to not play Ireland, going down the right branch, at home then France would win 89% of the time. The rest of the nodes are winners from the previous nodes using the branching criteria. The terminal nodes show the final decisions determining the winners down each branch of the tree.

Figure 14: Initial Decision Tree Model using Data Since Last RWC

A diagram of a network

Description automatically generated with medium confidence

The results in the terminal nodes of the tree showed France at home winning 11% of the time and France away winning 9% of the time giving a combined 20% winning percentage. Ireland winning 11% (against France at home) and Ireland winning 8% away from home for a combined 19% winning percentage. Wales winning 13%, South Africa winning 12%, Australia winning 11% of the time and a draw occurring 7% of the time.

There seems to be some overfitting of the data, in the bottom right the tree models England playing against England when England is away, so clearly some pruning was needed to improve the accuracy and reduce the overfitting of the data. To improve the model, a 10-fold cross validation was performed where the ideal complexity parameter (cp) was determined along a sequence from 0.01 to 0.1 increasing by 0.01. The complexity parameter is a hyperparameter within the model that controls the trade-off between the tree complexity and the accuracy when pruning the tree. Adjusting the cp of the model essentially represents the cost assigned to adding a new node to the tree.

The optimal complexity parameter was found to be 0.02 increasing the model accuracy to 57.03% and changing the internal nodes and final decisions in the terminal nodes. There also did not seem to be any overfitting of the data.

Figure 15: Pruned Decision Tree Model using Data Since Last RWC (cp = 0.02)

A diagram of a computer network

Description automatically generated with medium confidence

The results in the pruned model terminal nodes showed France still have the highest combined home and away winning percentage of 20%, and the first change seen where New Zealand came in second with a 19% combined winning percentage. Ireland came in third with a 17% combined winning percentage followed by both South Africa and Scotland with a combined 12% winning percentage, England with the same winning percentage at 12% with Wales coming in last at 5% and a draw occurring at 5%. In the RWC championship matches (quarter, semifinal and final matches) the outcome being a draw is unrealistic as the there would be a period of sudden death where the first person to score would win the match, but it could occur in pool play which decides the quarter final match placements.

**Results**

The results for the Association Rules Minning using data from the turn of the century showed the top rule found was that the teams won at home with the teams with the most wins at home being New Zealand, England, South Africa and then Australia. When controlling for teams winning away from home using the data from the turn of the century, New Zealand was found to win away from home more than any other team. The results for the Association Rules Minning using data since the last Rugby World Cup, it was found that the home team still won more than the away team, but it also found that France and Ireland performed the best both at home and away closely followed by Wales and New Zealand. The outlier being Italy who lost as home more than any other team When controlling for teams winning away from home, again Italy lost to away teams more than any other team, but both France and New Zealand are shown to perform well away with Ireland being a close third and Scotland performing well away in the 2023 Six Nations Championship competition.

The initial results of the Naïve Bayes model showed an accuracy of 53.33% when trying to predict the winning team when using data such as who the home team is, who the away team is, the scores for each team and the point difference. When Laplace smoothing was introduced and set to a value of 1, the accuracy of the Naïve Bayes model increased to 60%. The results after attempting to optimize the model using a 10-fold cross validation where the Laplace value was sequenced from 0.1 to 1.0 by 0.1 showed an accuracy of 53.33% matching the initial model’s accuracy.

The initial tree model had an overall accuracy of 48.9% and showed France at home winning 11% of the time and France away winning 9% of the time giving a combined 20% winning percentage. Ireland winning 11% (against France at home) and Ireland winning 8% away from home for a combined 19% winning percentage. Wales winning 13%, South Africa winning 12%, Australia winning 11% of the time and a draw occurring 7% of the time. After performing a 10-fold cross validation where the ideal complexity parameter (cp) was determined along a sequence from 0.01 to 0.1 increasing by 0.01, the model’s accuracy increased to 57.03%. The new terminal nodes showed France still have the highest combined home and away winning percentage of 20%, and the first change seen where New Zealand came in second with a 19% combined winning percentage. Ireland came in third with a 17% combined winning percentage followed by both South Africa and Scotland with a combined 12% winning percentage, England with the same winning percentage at 12% with Wales coming in last at 5% and a draw occurring at 5%.

**Conclusions**

The analysis showed that France and Ireland are the top performing teams since the last Rugby World Cup. Through the Association rules mining it was found that the home team has a significant advantage to win over the visiting team. When shifting the data to focus on the most recent top performing teams since the last RWC, France and Ireland performed the best at home and France and New Zealand performed the best away closely followed by Ireland. The Naïve Bayes model showed that using the criteria in the decision trees had a 53.33% accuracy in predicting the winning team, and the decision tree model after pruning showed the top performing teams to be France, New Zealand and Ireland in that order.

In regard to improving analysis, more match information in the data could be used to improve the model’s accuracies as that has been an issue in the current analysis. This could be in the form of numerical data such as the number of penalties for each side in the match, the number of cards for each team, yellow or red, or the number of injured players before or during the match as all would theoretically affect the outcome of the match. Other categorical information that could be included could be if the first- or second-string team was playing, the weather on the day of the match (raining, heat, wind speed) or the field formation the team adopts when playing in a match. This would not only aid in improving the current models but also widen the number of models one can apply to the data.

Using this analysis, its predicted that France will win the Rugby World Cup 2023 since they not only have been the best performing team since RWC 2019, but they are also the host nation for RWC 2023. Essentially all of France’s games this tournament will be at home and the number one rule found in Association rule mining showed home teams had a significant advantage over away teams. This is reinforced with the decision tree models showing France as the top performer in both the initial and improved models. Other teams like Ireland and New Zealand both perform well away from home with Ireland slightly outperforming New Zealand could be in contention for winning. South Africa also has incentive to defend their world title being the champions from RWC 2019 and should not be written off as well, but the prediction is France winning the Rugby World Cup 2023.

**Reference**

Begbie, L. (2023). *International Rugby Union results from 1871-2023* (CC BY-NC-SA 4.0; Version 16) [Data set]. Kaggle. <https://www.kaggle.com/datasets/lylebegbie/international-rugby-union-results-from-18712022>