FIFA World Cup Predictions: Utilizing Machine Learning to Determine the Accuracy of the Rank of Teams on World Cup Winners

Diana Batista Capellan, Kate Miller, Isabel Sumy

Introduction

The FIFA World Cup is an international competition like no other. In 2022, 1.5 billion people tuned in to watch the games and 88,966 were there live in Qatar. With an event of this size, fans go all out in support of their favorite team(s), but how can one know who will end up as the champion? This project will use machine learning and knowledge of data science to evaluate whether initial ranking of teams in the World Cup accurately predicts which team will win.

This project plans to predict the outcome of the next men's FIFA World Cup (2026) by using the previous year's team ranks from 1992 to 2023 and the outcomes of past games from 1930 to 2018, as well as evaluate the effectiveness of a machine learning model by testing on previous years' World Cups.

Machine learning is a practical approach to this because there is a large amount of data and machine learning is able to make complex predictions between variables in the data, as well as provide interpretations to the predictions of the data.

Illustration / Figure

Background and Related Work

With an event as large as the FIFA World Cup, many individuals have tried to determine who will win the next Cup for years. Individuals will develop brackets to attempt to predict who will win it all. Many times, people simply have gone off their intuition. However, those with experience in data science and machine learning have gone on to develop models similar to this one to predict the next winning team. Each model utilizes different data sets and features

to determine who will win the upcoming cup as well as different programming languages and visualizations to produce and showcase their work.

For example, there is a ProjectPro article depicting ways machine learning was utilized in FIFA 2022 and includes a project that tried to predict the outcome of the 2022 games using the results from 1870 to 2018 [1]. The article sets up a competition through Kaggle where teams or individuals can compete to produce the best model for predicting the winning teams. This is similar to what this project hopes to accomplish, though it will use different data sets and use the R programming language instead of Python.

Another example of similar work is outlined in a Medium article about predicting the 2022 FIFA World Cup [2]. The article goes into which features individuals found to be important in predicting the next winner and trying to simulate the results. Once the features were found, they were used to create different machine-learning models that analyze different team statistics to determine whether or not they could win the World Cup. This project will be similar to this idea as well but will use different data and work to predict the final winner of the competition solely based on the rank of the teams instead of multiple features.

Overall, there are many projects published that attempt to accomplish the same goal as this project. Although these projects exist, this specific project will differ in the data used to train the machine learning model and also differ in analyzing the most important features for the model.

Data Processing

About the Data

The collected data comes from three distinct sources for the exploratory data analysis seen below.

1. FIFA ranking data as of July 20, 2023 was acquired from a CSV file named `fifa_ranking-2023-07-20.csv` [3]. This dataset contains information about FIFA rankings for various countries, including the country name and its corresponding rank.
2. Data from FIFA World Cups was obtained from a CSV file named `worldcups.csv` [4]. This dataset encompasses details about different World Cup tournaments, such as the year, host country, winning team, runner-up, and other pertinent information.
3. Data on World Cup matches was sourced from a file named `wcmatches.csv` [5]. This dataset contains comprehensive information about matches played during FIFA World Cup tournaments, including the year, stage, participating teams, and match scores.

For the model used in the project, creation of `Excel csv` files was necessary. These three datasets were created by members of this project to analyze 2014, 2022, and 2024 men's FIFA team rankings. The datasets have two columns, `rank` and `Name`, where `Name` refers to the name

of the country participating in FIFA. The data for these csv files was obtained from the FIFA website [6]. To maintain accessability and reproducibility of this project, these data sets are available for viewing on GitHub and are linked in the References section at the end of this document. The model will be discussed in more detail later in this document.

Necessary packages for this project: dplyr, tidyverse, readr, glmnet, and ggplot2.

Data Cleaning

To ensure consistency and the facilitation of data integration, several cleaning and formatting steps were performed. The column names across all datasets were standardized by renaming the column for the winner of the World Cup to Country. Specifically, in the FIFA ranking data (data1), the column was renamed from country_full to Country. Similarly, in the World Cup data (data2), the column was renamed from winner to Country, and in the World Cup matches data (data3), the column was renamed from winning_team to Country. This uniform naming convention streamlines the subsequent data integration process.

Following the standardization of column names, the datasets were integrated through left joins based on the Country column. First, the World Cup data (data2) was merged with the FIFA ranking data (data1). The resultant dataset was merged with the World Cup matches data (data3). These sequential left joins to create the df dataframe enriched the World Cup data with FIFA rankings and match details, which facilitated effective Exploratory Data Analysis. To further clean the data, NA values, the win_conditions column, and the country column (a repeated column from data2) were all removed.

In order to properly train the machine learning model and be able to use it for future predictions, the creation of a new dataframe, new_df, was necessary. This new dataframe was created from the original df dataframe with the rank, Country, and year.x columns.

The df combined dataframe:

```
  year.x    host Country    second third     fourth goals_scored teams games
1   1930 Uruguay Uruguay Argentina   USA Yugoslavia           70    13    18
2   1930 Uruguay Uruguay Argentina   USA Yugoslavia           70    13    18
3   1930 Uruguay Uruguay Argentina   USA Yugoslavia           70    13    18
4   1930 Uruguay Uruguay Argentina   USA Yugoslavia           70    13    18
5   1930 Uruguay Uruguay Argentina   USA Yugoslavia           70    13    18
6   1930 Uruguay Uruguay Argentina   USA Yugoslavia           70    13    18
  attendance rank country_abrv total_points previous_points rank_change
1     434000   16          URU           48               0           0
2     434000   16          URU           48               0           0
3     434000   16          URU           48               0           0
4     434000   16          URU           48               0           0
5     434000   16          URU           48               0           0
```

```
6     434000   16          URU           48               0           0
  confederation  rank_date year.y           city       stage home_team
1      CONMEBOL 1992-12-31   1930     Montevideo     Group 3   Uruguay
2      CONMEBOL 1992-12-31   1930     Montevideo     Group 3   Uruguay
3      CONMEBOL 1992-12-31   1930     Montevideo  Semifinals   Uruguay
4      CONMEBOL 1992-12-31   1930     Montevideo       Final   Uruguay
5      CONMEBOL 1992-12-31   1950 Belo Horizonte     Group 4   Bolivia
6      CONMEBOL 1992-12-31   1950      São Paulo Final Round    Sweden
   away_team home_score away_score outcome losing_team       date month
1       Peru          1          0       H        Peru 1930-07-18   Jul
2    Romania          4          0       H     Romania 1930-07-21   Jul
3 Yugoslavia          6          1       H  Yugoslavia 1930-07-27   Jul
4  Argentina          4          2       H   Argentina 1930-07-30   Jul
5    Uruguay          0          8       A     Bolivia 1950-07-02   Jul
6    Uruguay          2          3       A      Sweden 1950-07-13   Jul
  dayofweek
1    Friday
2    Monday
3    Sunday
4 Wednesday
5    Sunday
6  Thursday
```

The `new_df` dataframe used for the model"

```
  Country rank year.x
1 Uruguay   16   1930
2 Uruguay   16   1930
3 Uruguay   16   1930
4 Uruguay   16   1930
5 Uruguay   16   1930
6 Uruguay   16   1930
```

Statistics

Summary statistics for the `df` dataframe:

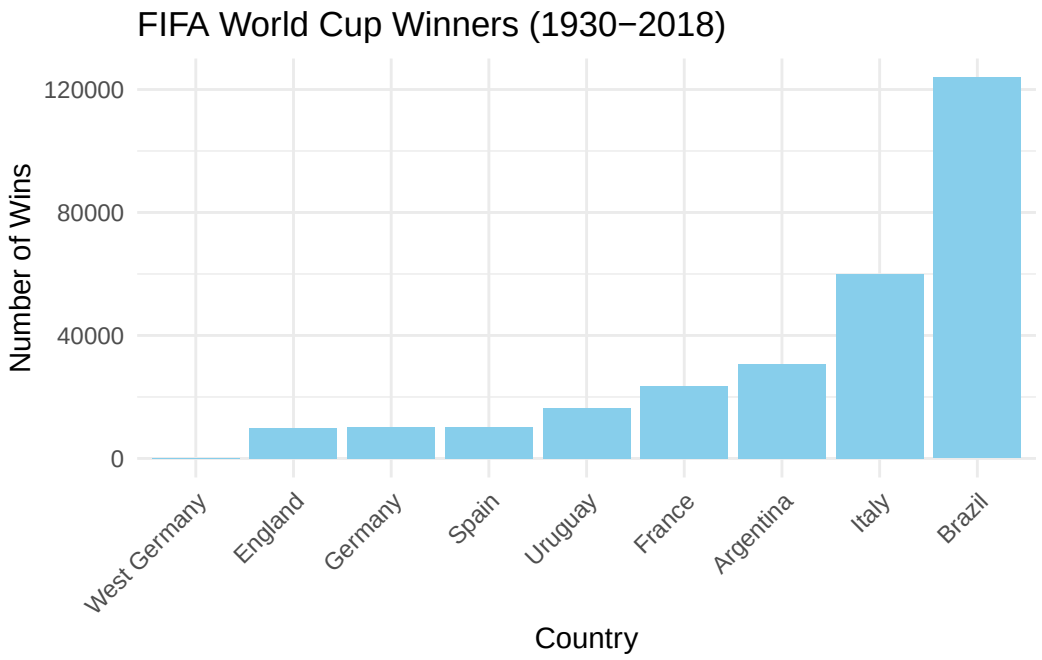
```
     year.x         host             Country             second
 Min.   :1930   Length:284272      Length:284272      Length:284272
 1st Qu.:1961   Class :character   Class :character   Class :character
 Median :1978   Mode  :character   Mode  :character   Mode  :character
 Mean   :1978
```

```
3rd Qu.:2002
Max.   :2018
   third              fourth           goals_scored       teams
Length:284272      Length:284272      Min.   : 70.0   Min.   :13.00
Class :character   Class :character   1st Qu.: 89.0   1st Qu.:16.00
Mode  :character   Mode  :character   Median :126.0   Median :16.00
                                      Mean   :122.4   Mean   :22.02
                                      3rd Qu.:147.0   3rd Qu.:32.00
                                      Max.   :171.0   Max.   :32.00
    games          attendance          rank         country_abrv
Min.   :17.00   Min.   : 395000   Min.   : 1.000   Length:284272
1st Qu.:32.00   1st Qu.: 868000   1st Qu.: 2.000   Class :character
Median :38.00   Median :1673975   Median : 4.000   Mode  :character
Mean   :43.64   Mean   :1925022   Mean   : 6.467
3rd Qu.:64.00   3rd Qu.:2859234   3rd Qu.: 9.000
Max.   :64.00   Max.   :3568567   Max.   :76.000
 total_points    previous_points   rank_change         confederation
Min.   :  36.0   Min.   :   0.0   Min.   :-32.00000   Length:284272
1st Qu.: 736.0   1st Qu.: 735.0   1st Qu.:  0.00000   Class :character
Median :1007.0   Median :1001.0   Median :  0.00000   Mode  :character
Mean   : 980.4   Mean   : 975.9   Mean   :  0.00221
3rd Qu.:1426.0   3rd Qu.:1426.0   3rd Qu.:  0.00000
Max.   :2172.0   Max.   :2187.0   Max.   : 29.00000
 rank_date            year.y         city              stage
Length:284272      Min.   :1930   Length:284272      Length:284272
Class :character   1st Qu.:1969   Class :character   Class :character
Mode  :character   Median :1986   Mode  :character   Mode  :character
                   Mean   :1984
                   3rd Qu.:2002
                   Max.   :2018
 home_team          away_team          home_score      away_score
Length:284272      Length:284272      Min.   :0.000   Min.   :0.00
Class :character   Class :character   1st Qu.:1.000   1st Qu.:0.00
Mode  :character   Mode  :character   Median :2.000   Median :1.00
                                      Mean   :1.976   Mean   :1.19
                                      3rd Qu.:3.000   3rd Qu.:2.00
                                      Max.   :8.000   Max.   :8.00
  outcome          losing_team           date              month
Length:284272      Length:284272      Length:284272      Length:284272
Class :character   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

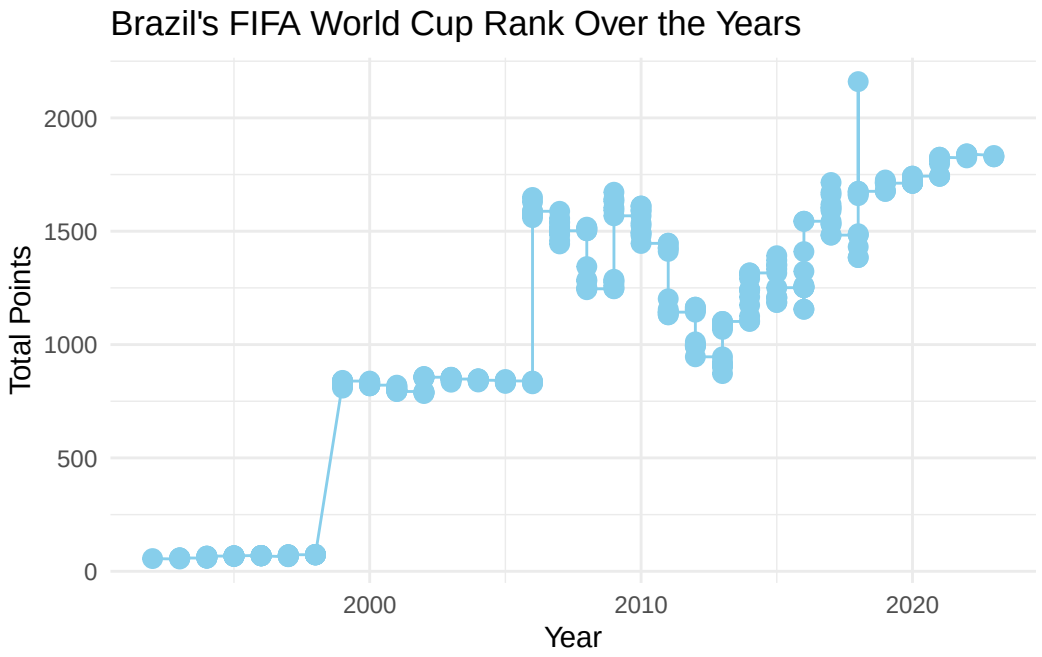
```
 dayofweek
Length:284272
Class :character
Mode  :character
```

Data Visualizations

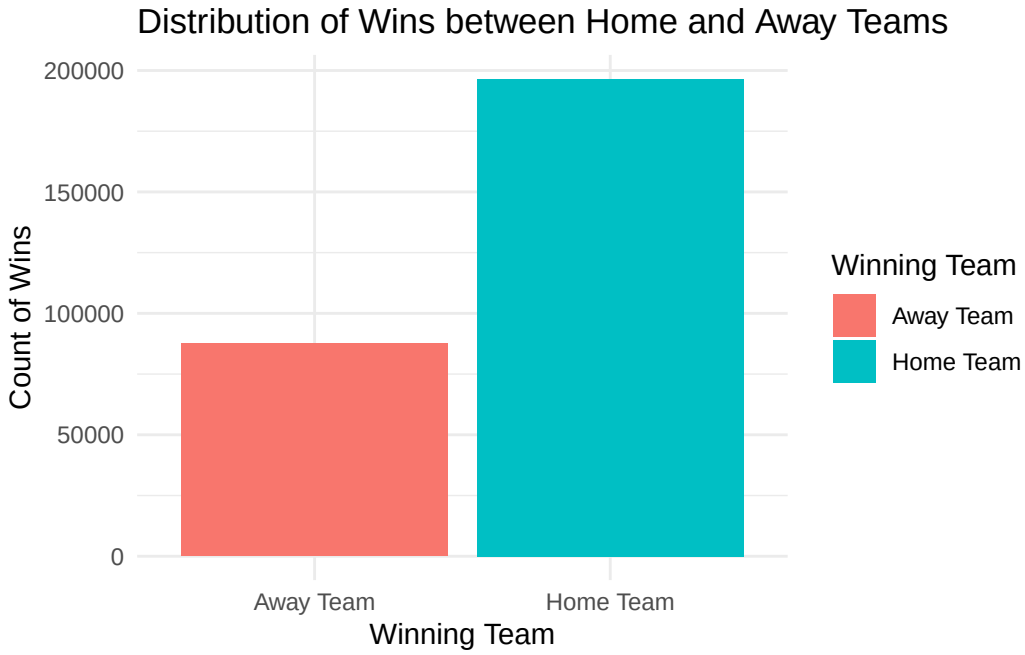
Three visualizations were created to help a general audience better understand the data used for this project. The visualizations utilize the `ggplot2` package.



The data visualization shows that Brazil has won the most World Cups throughout history.



Since Brazil has the most number of World Cup wins, the second visualization aims to show Brazil's ranking over time, which could be beneficial to the overall interpretation of the machine learning model's goal to show prediction of winners based on their ranking.



Another factor that could contribute to the model is a home team advantage. This visual aims to show that there may be a home team advantage, since the home team has won more often than the away team. One would expect the performance to be half and half if there was no correlation between the home team and the result of the match.

Architecture

The final model used for this project was the LASSO model. Experimentation took place to determine which model would be a better predictor for World Cup winner based on ranking, and the LASSO model had a lower mean square error than the ridge model, as seen below and described in detail in the next section.

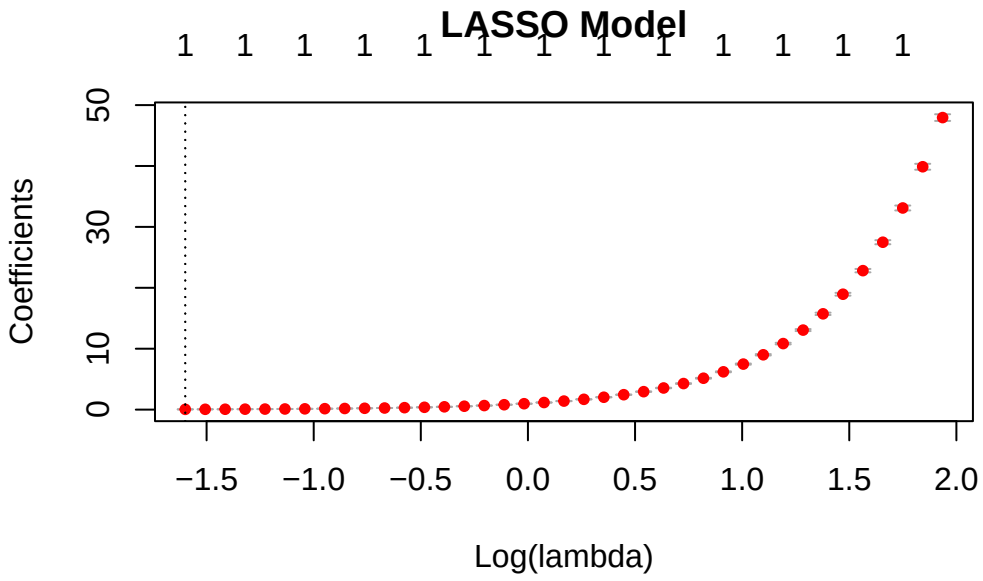
To create the LASSO model, the `glmnet` package was utilized. The `X` variable was created from the `new_df` dataframe and selected the `Country` column. the `y` variable was created from the `rank` column in `new_df`. Next, the X variable was converted to a matrix, and the y variable was converted to a vector to complete LASSO regression. Each of these were stored in a new variable, called `X_matrix` and `y_vector`, respectively. Next, the LASSO model was created utilizing `cv.glmnet` and using `X_matrix` and `y_vector`, with an alpha value of 1. This means that some of the coefficients are shrunk to zero and only the most important coefficients are selected. The mean square error from the model was extracted using the `cvm` column of the model and finding the minimum of that column.

To create a ridge model, one would perform the same steps as for the LASSO model but instead use an alpha value of 0, which only minimizes the sum of squared residuals. The mean square error was also extracted from the ridge model, but it was higher than that of the LASSO model, making the LASSO model the best performing model in this scenario.

A standard plot was created to visualize the LASSO model.

```
[1] "Mean Squared Error (Lasso): 0.0408740298234236"
```

```
[1] "Mean Squared Error (Ridge): 0.409325876812647"
```



Baseline Model

For a baseline model, a simple linear regression model was fit to the new dataframe `new_df` with `rank` and `Country` variables to predict which team will win the next World Cup. Linear regression minimizes the mean of the squared differences of the predicted and observed values. A drawback of this method is that overfitting can occur since linear regression does not impose any penalties on the coefficients. This is shown through this model's outputted mean square error (MSE) of 28.34 (rounded value), which is extremely high in this context. In order to decrease the MSE, other model types that used different strategies to decrease overfitting and multicollinearity were attempted, and the LASSO model was ultimately the best-performing model.

Quantitative Results

The main quantitative result used to measure and compare models is the mean squared error (MSE). The MSE is a commonly used measure to determine the effectiveness of a model and is calculated by taking the mean of the squared differences between the predicted and actual values. Since MSE is a quantitative measure, it can be used to objectively compare the performance of different regression models, as well as the specific performance of a single model.

MSE at its base is the difference between the actual and predicted values, so a lower MSE indicates that the predicted values are closer to the actual values. Contrarily, a higher MSE indicates that the predicted values are further from the actual values. In this way, models with lower MSEs usually have better performance than models with higher MSEs, which makes this measure a good representation of model quality in comparison situations.

Overall, mean square error is a helpful tool when determining the effectiveness of a regression model, and is an important comparison technique used in this project.

The mean square error for the LASSO model used in this project was 0.0408, which is a very low MSE. This means that the LASSO model was a good measure of whether ranking impacted the country that won the World Cup. This MSE shows that most teams who won the World Cup were ranked #1 going into it. Of course, there were still teams who won the World Cup who were ranked lower going into the World Cup. The mean square error for the Ridge model was 0.409, which is a slightly higher MSE than the LASSO model. It was ultimately determined to use the LASSO model to predict team performance in the World Cup in the future since it was the model with the lower MSE.

Qualitative Results

To determine the overall performance of the LASSO model, different years' World Cup ranking data were used. The predicted winners for 2014, 2022, and 2026 were predicted using the LASSO model, which was trained with the data of World Cup rankings and their winners from up until 2018. Though one of the datasets had data up until 2023, this is not included in the final model. The output of the predictions for these three years are shown below. The predictions were calculated using the base-R function `predict` and selecting the `rank` and `Name` columns from the specific year's dataset. Lastly, the minimum was selected from the rankings, since the minimum value would be the team that would be the LASSO model's prediction for winning the World Cup.

```
[1] "The predicted winner of the 2014 World Cup is: Germany"
```

```
[1] "The predicted winner of the 2022 World Cup is: Brazil"
```

```
[1] "The predicted winner of the 2026 World Cup is: Argentina"
```

The 2014 World Cup prediction was correct based on rank, since Germany did in fact win that World Cup. The 2022 rankings were not included in the original data, so this project aimed to evaluate the winner of the World Cup. Brazil was first ranked heading into the World Cup, but Argentina took the win against France, so Brazil wasn't even present in the final match. This is an example where the model might not work entirely to performance. The overall low mean square error of the model signifies that the majority of the time, the team that has been

ranked first heading into the World Cup will win the World Cup, making the predictions of the model favor those who are ranked higher, since their success has been seen over time.

Lastly, the 2026 predicted winner is Argentina. This is not guaranteed, since the data used is from 2024, and there are still two more years until the World Cup. The ranking could change in these two years, so to gain an accurate prediction, inputting the dataset with updated data in 2026 into the LASSO model could be beneficial.

Discussion

Overall, the project is performing at a subjectively mediocre performance. There was an accurate prediction for the future winners of the World Cup, but there was also an inaccurate prediction. The model in terms of ranking may need other predictors or factors, such as how well the team is performing as a whole, to determine the winner. However, that information is outside the scope of this particular project.

Something that is interesting about the results is that the model will usually choose the top ranked team to win the World Cup. This is interesting because in 2022, Brazil, the top ranked team, did not win and was not even in the final match, which was unexpected.

Our group learned how to clean a dataset and how to combine datasets to provide additional understanding of the data prior to building a machine learning model. We also learned how to communicate the difference between mean square errors of the different models to determine which would perform the best with the given data. Lastly, we learned that the model will not always be correct in predicting the output or more importantly the next winner of the World Cup, since the model is only based on rank.

Ethical Considerations

The World Cup prediction model faces potential biases and limitations that warrant consideration. One concern is fairness, as the historical data used might favor certain teams or regions, leading to unfair predictions that disadvantage others.

Another challenge is that the model relies on historical data, which might not fully reflect how the game has changed over time. For example, it might not account for new strategies or changes in player performance. Also, sports outcomes are uncertain, with factors like injuries and referee decisions making it hard to predict results accurately. Ranking might not be the only thing at play here, and the determination of each country's ranking may not take into account these other factors.

A specific limitation of the model itself is that it is limited to predicting the winner solely on rank. This project acknowledges that there may be more data necessary to determining the winner of the World Cup, such as team performance, total points, and historical performance in World Cup finals. In terms of limitations of the training data, there were multiple teams

who were not ranked first in the training data who won the World Cup, which was not reflected in the predictions of the model.

Conclusion

In conclusion, the LASSO model worked well to determine if rank is a good predictor of the overall winner of the World Cup, and the model accurately predicted the result of the 2014 winner based on the ranking of teams before heading into the World Cup playoffs. This model can be used in the future to hopefully predict the correct outcome of future World Cups based on ranking.

References

Code Appendix