Predicting the Winner of the 2026 Men's FIFA World Cup

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 do we know who will end up as the champion? With this project, we will use machine learning and our knowledge of data science to answer this question. We plan to predict the outcome of the next men's soccer 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.

Exploratory Data Analysis

```r
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.4
v ggplot2   3.4.4     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.0

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(readr)

data1 <- read.csv("fifa_ranking-2023-07-20.csv")

data2 <- read.csv("worldcups.csv")

data3 <- read.csv("wcmatches.csv")

# Use columns from data 3: Home score and away score
# Use columns from data 1: Rank
# Add these columns to data 2

# Make outcome a factor instead of a character
# One of columns was blank, remove na's

names(data1)[names(data1) == "country_full"] <- "Country"
names(data2)[names(data2) == "winner"] <- "Country"
names(data3)[names(data3) == "winning_team"] <- "Country"

first_join <- left_join(data2, data1, by = "Country")

second_join <- left_join(first_join, data3, by = "Country")

df <- second_join %>%
  select(-win_conditions) %>%
  na.omit()
```

```r
colnames(df)
```

```
 [1] "year.x"            "host"              "Country"           "second"
 [5] "third"            "fourth"           "goals_scored"     "teams"
 [9] "games"            "attendance"       "rank"             "country_abrv"
[13] "total_points"     "previous_points" "rank_change"      "confederation"
[17] "rank_date"        "year.y"           "country"          "city"
[21] "stage"            "home_team"        "away_team"        "home_score"
[25] "away_score"       "outcome"          "losing_team"      "date"
[29] "month"            "dayofweek"
```

```
colnames(data1)
```

```
[1] "rank"            "Country"         "country_abrv"    "total_points"
[5] "previous_points" "rank_change"     "confederation"   "rank_date"
```

```
colnames(data2)
```

```
 [1] "year"          "host"          "Country"       "second"        "third"
 [6] "fourth"        "goals_scored" "teams"         "games"         "attendance"
```

```
colnames(data3)
```

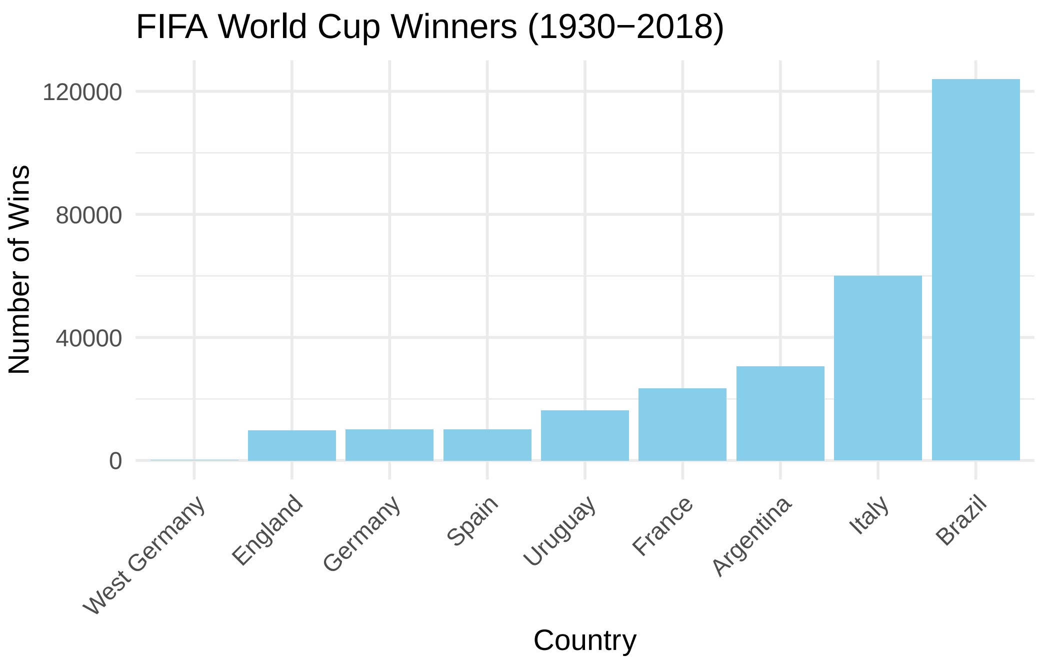
```
 [1] "year"            "country"         "city"            "stage"
 [5] "home_team"       "away_team"       "home_score"      "away_score"
 [9] "outcome"         "win_conditions" "Country"         "losing_team"
[13] "date"            "month"           "dayofweek"
```

Illustration / Figure

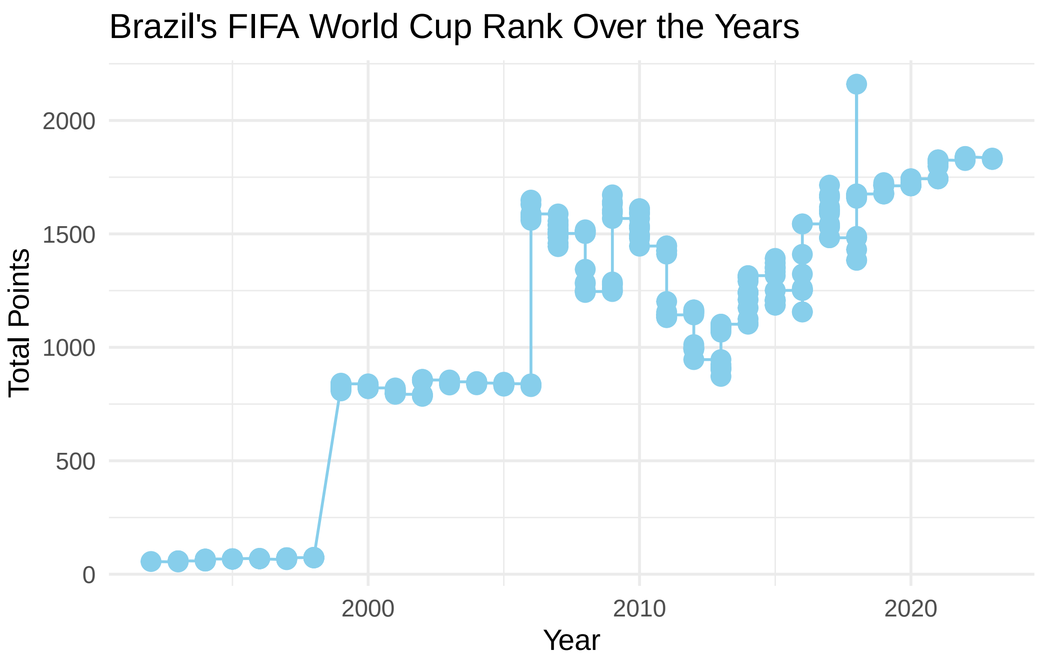
======= A figure or a diagram that illustrates the overall model or idea of your project. The idea is to make your report more accessible, especially to readers who are starting by skimming your work. For the project, taking a picture of a hand-drawn diagram is fine, as long as it's legible. PowerPoint is another option. You will not be penalized for hand-drawn illustrations – you are graded on the design and illustrative power

```
library(dplyr)
library(ggplot2)

# Filter matches where the home team won
```

```r
home_wins <- filter(second_join, outcome == "H")

# Filter matches where the away team won
away_wins <- filter(second_join, outcome == "A")

# Combine the home and away wins
all_wins <- bind_rows(home_wins, away_wins)

# Aggregate the data to count the number of wins for each country
world_cup_wins <- all_wins %>%
  group_by(Country) %>%
  summarise(Wins = n())

# Sort the data by the number of wins in descending order
world_cup_wins <- world_cup_wins[order(-world_cup_wins$Wins),]

# Create the bar plot
ggplot(world_cup_wins, aes(x = reorder(Country, Wins), y = Wins)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "FIFA World Cup Winners (1930-2018)",
       x = "Country",
       y = "Number of Wins") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
library(dplyr)
library(ggplot2)

# Filter Brazil's rank data
brazil_rank <- filter(data1, Country == "Brazil")

# Extract the year from the rank_date column
brazil_rank$year <- substr(brazil_rank$rank_date, 1, 4)

# Convert year to numeric
brazil_rank$year <- as.integer(brazil_rank$year)

# Create the line plot
ggplot(brazil_rank, aes(x = year, y = total_points)) +
  geom_line(color = "skyblue") +
  geom_point(color = "skyblue", size = 3) +
  labs(title = "Brazil's FIFA World Cup Rank Over the Years",
       x = "Year",
       y = "Total Points") +
  theme_minimal()
```



```r
class(df$outcome)
```

```
[1] "character"
```

```r
# Print first few rows of the data frame
head(df)
```

```
  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 country           city       stage home_team
1      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo     Group 3   Uruguay
2      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo     Group 3   Uruguay
3      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo  Semifinals   Uruguay
4      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo       Final   Uruguay
5      CONMEBOL 1992-12-31   1950  Brazil Belo Horizonte     Group 4   Bolivia
6      CONMEBOL 1992-12-31   1950  Brazil      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
```

Background & Related Work (2 points)

With an event as big as the Fifa World Cup, many people have tried to determine who will win the next Cup for years. People will develop brackets and make bets with their friends, or for money, about who will win it all. Many times people simply go off their intuition but those with experience in data science and machine learning have gone on to develop models similar to ours to predict the next winning team. Each model utilizes different datasets 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, from ProjectPro there is an article depicting ways machine learning was utilized in Fifa 2022 and includes a project which tries to predict the outcome of the 2022 games using the results from 1870 to 2018. 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 we wish to do in this project although we will be using different data sets and doing our work using R instead of Python like the competition suggests. Another example of similar work is outlined in a Medium article about predicting the 2022 Fifa World Cup. The article goes into which features they found to be important in predicting the next winner and trying to simulate the results. Once the features from their datasets were found, they were used to create different machine-learning models that take in team aspects to determine whether or not they could win the World Cup. Our project will be similar to this one as well but will use different data and work to predict the final winner of the competition based on existing teams.

Overall, there are other projects out there that attempt to accomplish the same goal as ours, meaning that it is possible. Although these projects exist, ours will differ in the data used and therefore also differ in which features are important for our particular models.

Data Processing

Describe the data that you have collected and cleaned. Be clear and specific when describing what you've done, so that a classmate can reproduce your work. Show some statistics and examples of your data.

- 4/4 Clearly describes sources of data, and the steps you took to clean and format your data. Statistics and data example are well-chosen, and gives readers a "feel" for your data.
- 3/4 Mostly clear description, but some aspects of the data processing steps are vague. Statistics and data example are somewhat illustrative/helpful.
- 2/4 Vague description or missing key information about where your data comes from or what you did. No example data shown, or the ones shown are not illustrative.
- 1/4 Incomplete information.

Architecture

```r
library(glmnet)
```

```
Warning: package 'glmnet' was built under R version 4.3.3

Loading required package: Matrix


Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack

Loaded glmnet 4.1-8
```

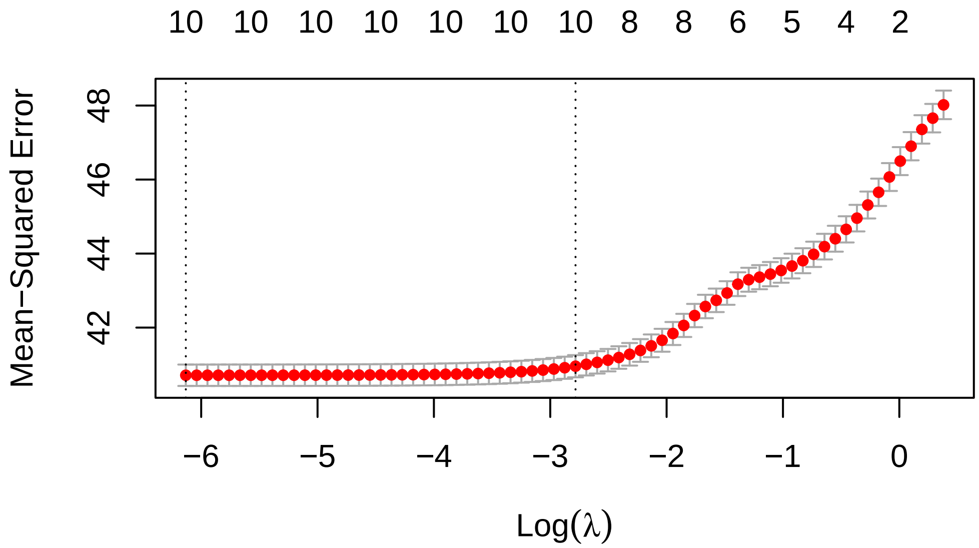
```r
df <- df[, !colnames(df) %in% c("outcome")]

# Extract rank variable
rank <- df$rank

# Remove rank variable from predictors
X <- df[, -which(names(df) == "rank")]

# Lasso regression
lasso <- cv.glmnet(x = as.matrix(X), y = rank, alpha = 1)
```

```
Warning in storage.mode(xd) <- "double": NAs introduced by coercion
```

```
Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion
```

```
Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion
Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion
```

```
library(ggplot2)

# Extract lambda values and corresponding mean squared errors
plot(lasso)
```

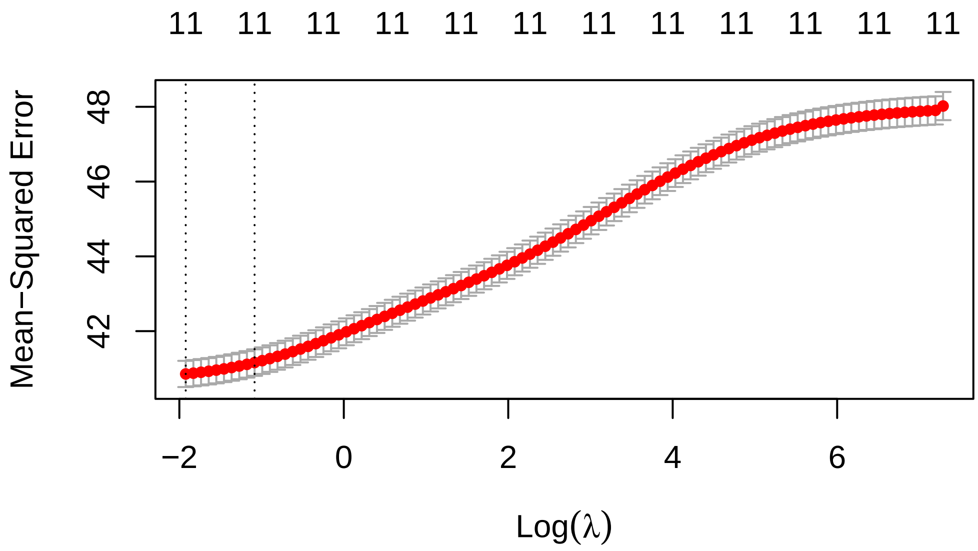


```
lambda <- log(lasso$lambda)
mse <- lasso$cvm

# Ridge regression
```

```r
ridge <- cv.glmnet(x = as.matrix(X), y = rank, alpha = 0)
```

```
Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in storage.mode(xd) <- "double": NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion

Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion
Warning in cbind2(1, newx) %*% nbeta: NAs introduced by coercion
```

```
plot(ridge)
```



```
lambda <- log(ridge$lambda)
mse <- ridge$cvm

# Extract MSE for Lasso model
lasso_mse <- min(lasso$cvm)

# Extract MSE for Ridge model
ridge_mse <- min(ridge$cvm)

# Output MSE for both models
print(paste("Mean Squared Error (Lasso):", lasso_mse))
```

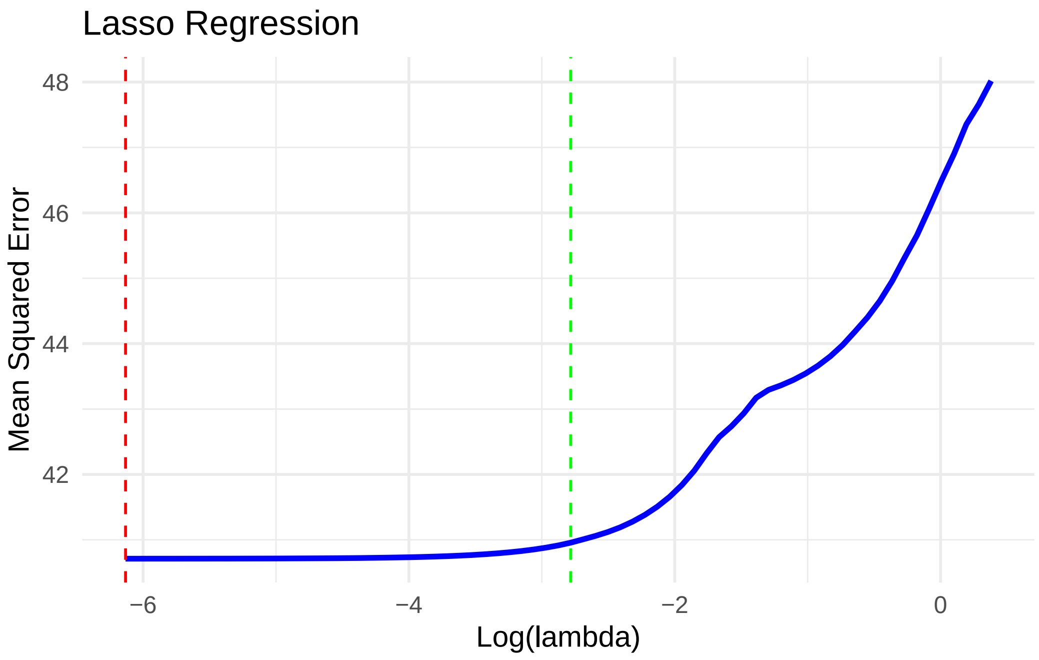
```
[1] "Mean Squared Error (Lasso): 40.7116236317294"
```

```
print(paste("Mean Squared Error (Ridge):", ridge_mse))
```

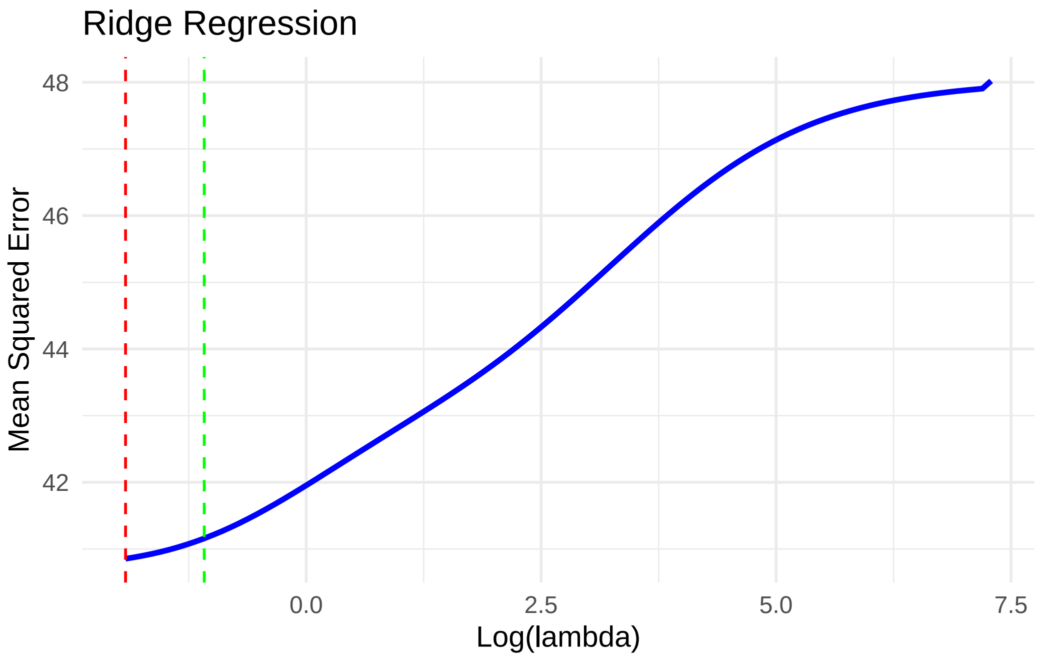
```
[1] "Mean Squared Error (Ridge): 40.8543266734552"
```

```
# Lasso regression
lambda_lasso <- log(lasso$lambda)
mse_lasso <- lasso$cvm
lambda_min_lasso <- log(lasso$lambda.min)
lambda_1se_lasso <- log(lasso$lambda.1se)

plot_data_lasso <- data.frame(lambda = lambda_lasso, mse = mse_lasso)

ggplot(plot_data_lasso, aes(x = lambda, y = mse)) +
  geom_line(color = "blue", size = 1) +
  geom_vline(xintercept = lambda_min_lasso, linetype = "dashed", color = "red") +
  geom_vline(xintercept = lambda_1se_lasso, linetype = "dashed", color = "green") +
  labs(x = "Log(lambda)", y = "Mean Squared Error", title = "Lasso Regression") +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



```
# Ridge regression
lambda_ridge <- log(ridge$lambda)
mse_ridge <- ridge$cvm
lambda_min_ridge <- log(ridge$lambda.min)
lambda_1se_ridge <- log(ridge$lambda.1se)

plot_data_ridge <- data.frame(lambda = lambda_ridge, mse = mse_ridge)

ggplot(plot_data_ridge, aes(x = lambda, y = mse)) +
  geom_line(color = "blue", size = 1) +
  geom_vline(xintercept = lambda_min_ridge, linetype = "dashed", color = "red") +
  geom_vline(xintercept = lambda_1se_ridge, linetype = "dashed", color = "green") +
  labs(x = "Log(lambda)", y = "Mean Squared Error", title = "Ridge Regression") +
  theme_minimal()
```



```
head(data3)
```

```
  year country       city   stage home_team     away_team home_score away_score
1 1930 Uruguay Montevideo Group 1    France        Mexico          4          1
2 1930 Uruguay Montevideo Group 4   Belgium United States          0          3
3 1930 Uruguay Montevideo Group 2    Brazil    Yugoslavia          1          2
```

```
4 1930 Uruguay Montevideo Group 3      Peru       Romania          1          3
5 1930 Uruguay Montevideo Group 1 Argentina        France          1          0
6 1930 Uruguay Montevideo Group 1     Chile        Mexico          3          0
  outcome win_conditions       Country losing_team       date month dayofweek
1       H                       France      Mexico 1930-07-13   Jul    Sunday
2       A                United States     Belgium 1930-07-13   Jul    Sunday
3       A                   Yugoslavia      Brazil 1930-07-14   Jul    Monday
4       A                      Romania        Peru 1930-07-14   Jul    Monday
5       H                    Argentina      France 1930-07-15   Jul   Tuesday
6       H                        Chile      Mexico 1930-07-16   Jul Wednesday
```

```
head(df)
```

```
  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 country           city       stage home_team
1      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo     Group 3   Uruguay
2      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo     Group 3   Uruguay
3      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo  Semifinals   Uruguay
4      CONMEBOL 1992-12-31   1930 Uruguay     Montevideo       Final   Uruguay
5      CONMEBOL 1992-12-31   1950  Brazil Belo Horizonte     Group 4   Bolivia
6      CONMEBOL 1992-12-31   1950  Brazil      São Paulo Final Round    Sweden
   away_team home_score away_score losing_team       date month dayofweek
1       Peru          1          0        Peru 1930-07-18   Jul    Friday
2    Romania          4          0     Romania 1930-07-21   Jul    Monday
3 Yugoslavia          6          1  Yugoslavia 1930-07-27   Jul    Sunday
4  Argentina          4          2   Argentina 1930-07-30   Jul Wednesday
5    Uruguay          0          8     Bolivia 1950-07-02   Jul    Sunday
6    Uruguay          2          3      Sweden 1950-07-13   Jul  Thursday
```

A description of the final model. Do not describe all the intermediate models that you have tried. Instead, present the model (or models) whose quantitative results you will show. These should be your most interesting models. Be as specific as you can while being concise. Readers should be able to reproduce a model similar enough to yours and obtain a similar performance.

Baseline Model

Describe a simple, baseline model that you will compare your neural network against. This can be a simple model that you build.

Quantitative Results

A description of the quantitative measures of your result. What measurements can you use to illustrate how your model performs?

Qualitative Results

Include some sample outputs of your model, to help your readers better understand what your model can do. The qualitative results should also put your quantitative results into context (e.g. Why did your model perform well? Is there a type of input that the model does not do well on?)

Discussion

Discuss your results. Do you think your model is performing well? Why or why not? What is unusual, surprising, or interesting about your results? What did you learn?

Ethical Considerations

Description of a use of the system that could give rise to ethical issues. Are there limitations of your model? Your training data?

(Note that the expectations are higher here than in the project proposal.)

Conclusion(Optional)

Summarize the whole report.