Google Cloud & NCAA® ML Competition 2018-Women’s

Pranjal Vijay

March 6, 2018

Loading the required libraries

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(readr)
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
```

```
library(magrittr)
library(ModelMetrics)
```

```
##
## Attaching package: 'ModelMetrics'
```

```
## The following objects are masked from 'package:caret':
##
##     confusionMatrix, precision, recall, sensitivity, specificity
```

```
library(ggplot2)
library(stats)
```

Loading the WNCAATouneyCompactResults Data

```
WNCAAT_result <- read.csv("C:/Users/ddddd/Comptitions/Kaggle/Google Cloud & NCAA/WNCAATourneyCom
pactResults.csv",stringsAsFactors = FALSE)
```

Research question for stage 1

Stage 1 - You should submit predicted probabilities for every possible matchup in the past 4 NCAA® tournaments (2014-2017).

Data Exploration

```
names(WNCAAT_result)
```

```
## [1] "Season"  "DayNum"  "WTeamID" "WScore"  "LTeamID" "LScore"  "WLoc"
## [8] "NumOT"
```

```
str(WNCAAT_result)
```

```
## 'data.frame':    1260 obs. of  8 variables:
##  $ Season : int  1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
##  $ DayNum : int  137 137 137 137 137 137 137 137 137 137 ...
##  $ WTeamID: int  3104 3112 3163 3198 3203 3234 3242 3301 3304 3314 ...
##  $ WScore : int  94 75 93 59 74 77 72 89 76 91 ...
##  $ LTeamID: int  3422 3365 3193 3266 3208 3269 3408 3263 3307 3224 ...
##  $ LScore : int  46 63 52 45 72 59 68 64 59 71 ...
##  $ WLoc   : chr  "H" "H" "H" "H" ...
##  $ NumOT  : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
WNCAAT_result$WLoc <- as.factor(WNCAAT_result$WLoc )
```

```
str(WNCAAT_result)
```

```
## 'data.frame':    1260 obs. of  8 variables:
##  $ Season : int  1998 1998 1998 1998 1998 1998 1998 1998 1998 1998 ...
##  $ DayNum : int  137 137 137 137 137 137 137 137 137 137 ...
##  $ WTeamID: int  3104 3112 3163 3198 3203 3234 3242 3301 3304 3314 ...
##  $ WScore : int  94 75 93 59 74 77 72 89 76 91 ...
##  $ LTeamID: int  3422 3365 3193 3266 3208 3269 3408 3263 3307 3224 ...
##  $ LScore : int  46 63 52 45 72 59 68 64 59 71 ...
##  $ WLoc   : Factor w/ 3 levels "A","H","N": 2 2 2 2 1 2 2 2 3 2 ...
##  $ NumOT  : int  0 0 0 0 0 0 0 0 0 0 ...
```

Data Preprocessing

```
WNCAAT_result %>%
  filter( Season != "NA", DayNum != "NA", WTeamID != "NA", WScore != "NA", LTeamID != "NA", LSco
re != "NA", WLoc != "NA", NumOT != "NA") %>%
  select(Season, DayNum, WTeamID, WScore , LTeamID, LScore, WLoc, NumOT) %>%
  group_by(Season,DayNum, WTeamID, WScore , LTeamID, LScore, WLoc, NumOT) %>%
  summarise(count=n())
```

```
## # A tibble: 1,260 x 9
## # Groups:   Season, DayNum, WTeamID, WScore, LTeamID, LScore, WLoc [?]
##    Season DayNum WTeamID WScore LTeamID LScore   WLoc NumOT count
##     <int>  <int>   <int>  <int>   <int>  <int> <fctr> <int> <int>
##  1   1998    137    3104     94    3422     46      H     0     1
##  2   1998    137    3112     75    3365     63      H     0     1
##  3   1998    137    3163     93    3193     52      H     0     1
##  4   1998    137    3198     59    3266     45      H     0     1
##  5   1998    137    3203     74    3208     72      A     0     1
##  6   1998    137    3234     77    3269     59      H     0     1
##  7   1998    137    3242     72    3408     68      H     0     1
##  8   1998    137    3301     89    3263     64      H     0     1
##  9   1998    137    3304     76    3307     59      N     0     1
## 10   1998    137    3314     91    3224     71      H     0     1
## # ... with 1,250 more rows
```

As asked for stage 1, I filterized the data for season 2013 to 2017

```
WNCAAT_result %>%
  filter(Season >= 2013, Season != "NA", DayNum != "NA", WTeamID != "NA", WScore != "NA", LTeamI
D != "NA", LScore != "NA", WLoc != "NA", NumOT != "NA") %>%
  select(Season,DayNum, WTeamID, WScore , LTeamID, LScore, WLoc, NumOT) %>%
  #mutate(Seasons = Season >= 2013 )
  group_by(Season, DayNum, WTeamID, WScore , LTeamID, LScore, WLoc, NumOT) %>%
  summarise(count=n())
```

```
## # A tibble: 315 x 9
## # Groups:   Season, DayNum, WTeamID, WScore, LTeamID, LScore, WLoc [?]
##    Season DayNum WTeamID WScore LTeamID LScore   WLoc NumOT count
##     <int>  <int>   <int>  <int>   <int>  <int> <fctr> <int> <int>
##  1   2013    138    3143     90    3201     76      N     0     1
##  2   2013    138    3163    105    3225     37      H     0     1
##  3   2013    138    3166     61    3393     56      N     0     1
##  4   2013    138    3208     70    3285     50      N     0     1
##  5   2013    138    3235     72    3211     60      N     0     1
##  6   2013    138    3242     67    3160     52      A     0     1
##  7   2013    138    3268     72    3346     52      H     0     1
##  8   2013    138    3277     55    3265     47      N     0     1
##  9   2013    138    3304     73    3151     59      N     0     1
## 10   2013    138    3328     78    3141     73      N     0     1
## # ... with 305 more rows
```

Data Partitioning for training and testing

```
inTrain <- createDataPartition(y=WNCAAT_result$WTeamID,
                              p=0.90, list=FALSE)
training <- WNCAAT_result[inTrain,]
testing <- WNCAAT_result[-inTrain,]
dim(training)
```

```
## [1] 1135    8
```

```
dim(testing)
```

```
## [1] 125   8
```

Modelling

```
WNCAAT_Model = lm(WTeamID ~  Season+ WScore  + LScore +DayNum , data = training)
summary(WNCAAT_Model)
```

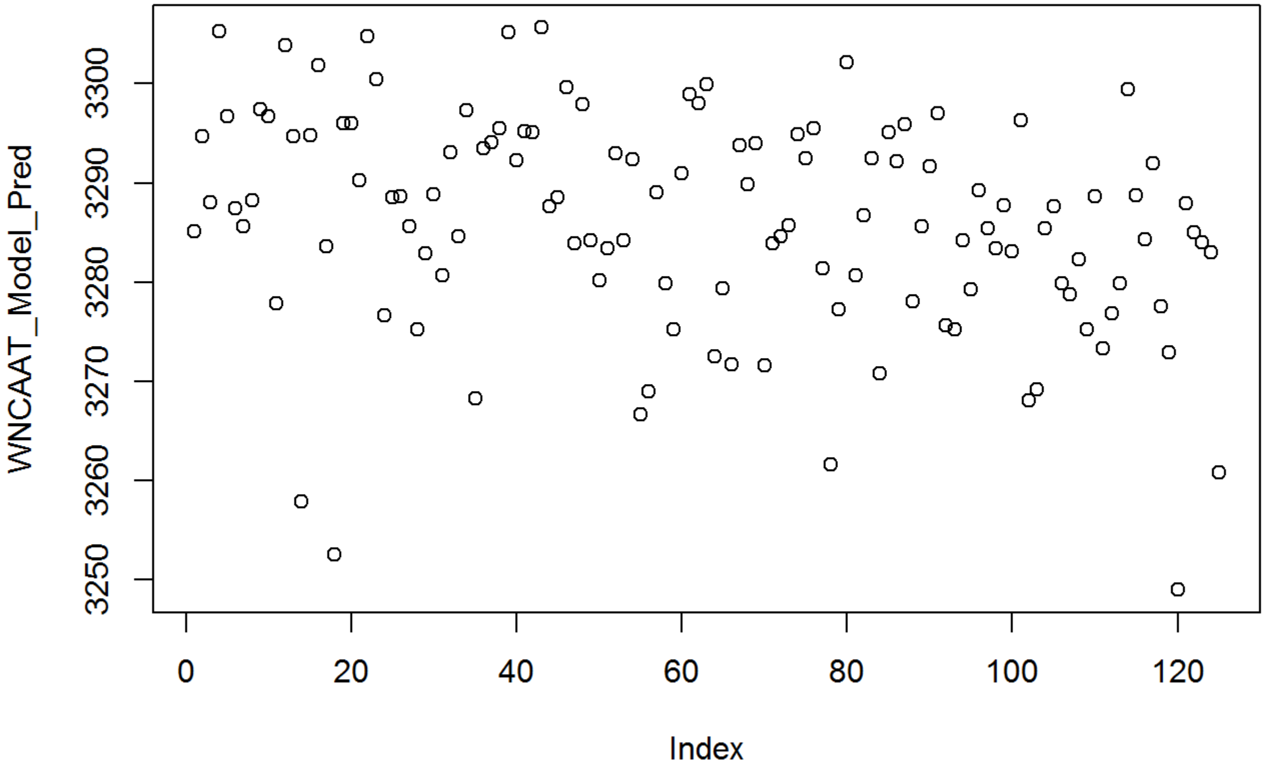
```
## 
## Call:
## lm(formula = WTeamID ~ Season + WScore + LScore + DayNum, data = training)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -194.718  -89.381    1.001   91.656  190.544 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4712.2401  1016.8356   4.634    4e-06 ***
## Season        -0.6171     0.5045  -1.223  0.22147    
## WScore        -0.9143     0.2843  -3.215  0.00134 ** 
## LScore         0.3267     0.2892   1.130  0.25881    
## DayNum        -0.9791     0.7093  -1.380  0.16777    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 97.44 on 1130 degrees of freedom
## Multiple R-squared:  0.01156,    Adjusted R-squared:  0.008063 
## F-statistic: 3.305 on 4 and 1130 DF,  p-value: 0.01055
```

Here we get 74% accuracy and p value <0.05

Prediction

```
WNCAAT_Model_Pred <- predict(WNCAAT_Model, testing)

plot(WNCAAT_Model_Pred)
```


WNCAAT_Model_Pred
3250
3260
3270
3280
3290
3300
0
20
40
60
80
100
120
Index