Capstone IDV Admission prediction

Bush Daniel Kwajaffa

6/3/2020

Introduction

Admission today is being affected by many factors and some of the things to be considered before being admitted are; CGPA, GRE.Score, TOEFL.Score, A statement of purpose (SOP), Letter of Recommendation (LOR), university rating and Research. In this project we will use the above mention criteria to create a model that will predict the admittance for a new student. Different models will be used to get the best out, the models that will be are linear regression, logistic regression and randomForest. The Root Mean Square Error (RMSE) will be used to evaluate the model performance. RMSE is a measure of how spread out the residuals are, it measures how concentrated the data is around the line of best fit. Models will be developed to compare RMSE in order to assess highest quality. The best resulting model will be used to predict the admittance.

Methodology and Analysis

Fisrtly we will load all required packages and get our data ready. Loading required packages.

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ----------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret
```

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
if(!require(data.table)) install.packages("data.table", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
if(!require(RCurl)) install.packages("RCurl", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: RCurl

##
## Attaching package: 'RCurl'

## The following object is masked from 'package:tidyr':
##
##     complete
```

```
if(!require(ggplot2)) install.packages("ggplot2", repos = "http://cran.us.r-project.org")
if(!require(corrplot)) install.packages("corrplot", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: corrplot

## corrplot 0.84 loaded
```

```
if(!require(gridExtra)) install.packages("gridExtra", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: gridExtra

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
if(!require(randomForest)) install.packages("randomForest", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: randomForest

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

Downloading and importing data.

```
urlfile <- "https://raw.github.com/bushdanielkwajaffa/edxcapstone2/master/Admission_Predict.csv"
Admission <- read.csv(urlfile)
```

Analysis and Data visualization

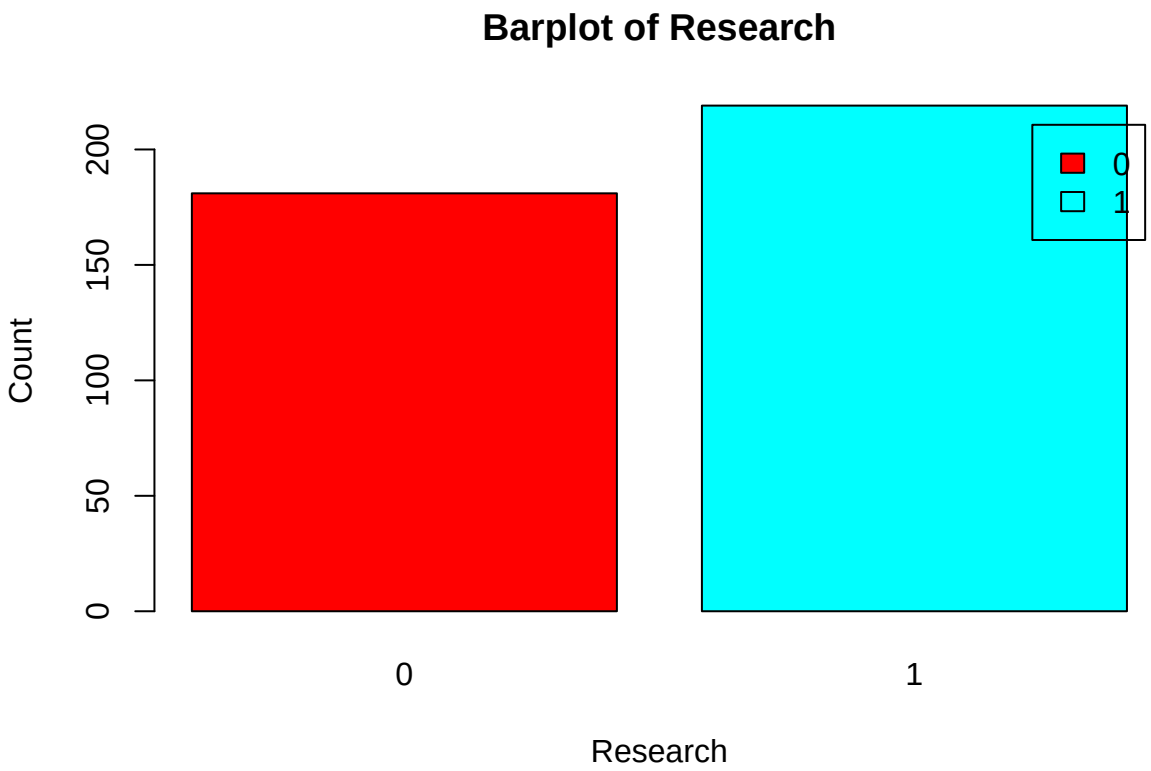
Summary

```
summary(Admission)
```

```
##    Serial.No.      GRE.Score      TOEFL.Score    University.Rating
##  Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000
##  1st Qu.:100.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000
##  Median :200.5   Median :317.0   Median :107.0   Median :3.000
##  Mean   :200.5   Mean   :316.8   Mean   :107.4   Mean   :3.087
##  3rd Qu.:300.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000
##  Max.   :400.0   Max.   :340.0   Max.   :120.0   Max.   :5.000
##       SOP             LOR             CGPA           Research
##  Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000
##  1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000
##  Median :3.5   Median :3.500   Median :8.610   Median :1.0000
##  Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475
##  3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000
##  Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000
##  Chance.of.Admit
##  Min.   :0.3400
##  1st Qu.:0.6400
##  Median :0.7300
##  Mean   :0.7244
##  3rd Qu.:0.8300
##  Max.   :0.9700
```

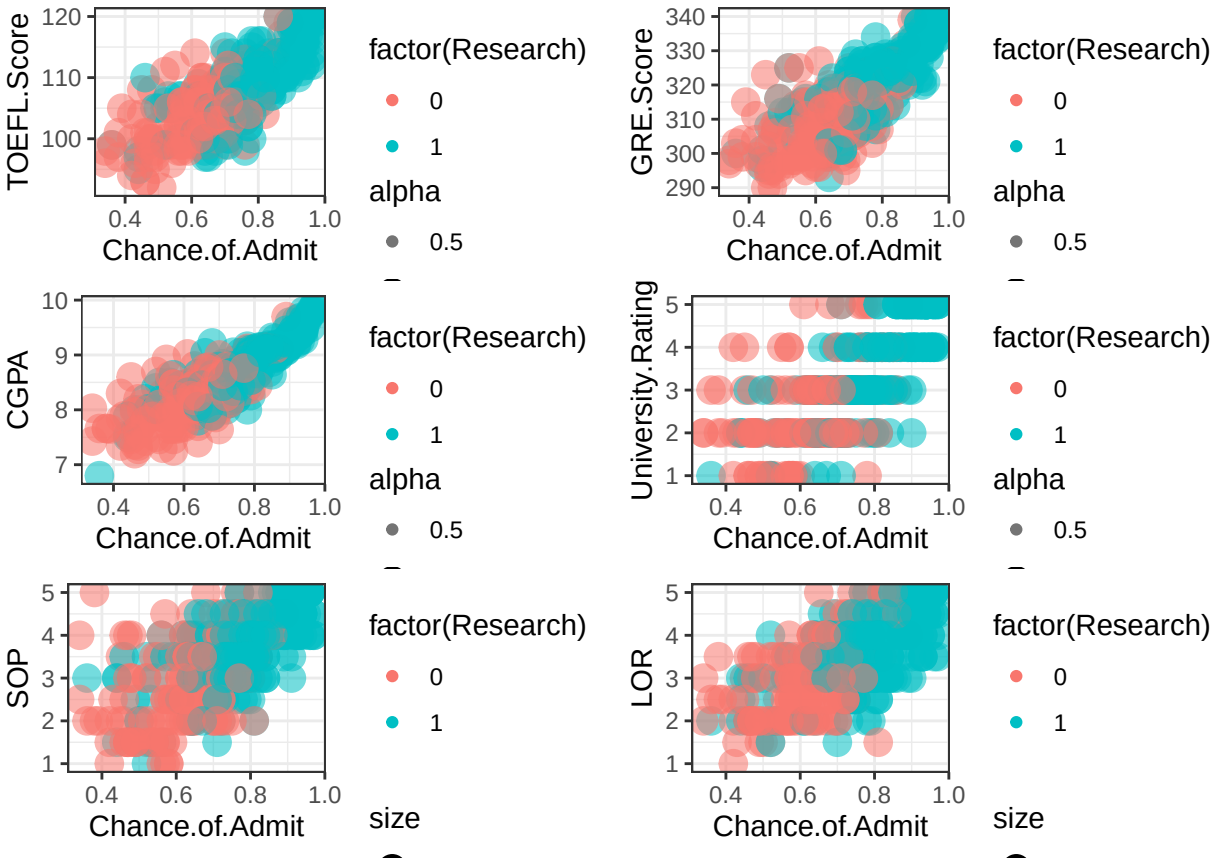
From the summary we have 400 rows, GRE.Score; 290-340, TOEFL.Score; 92-120, University rating; 1-5, SOP; 1-5, LOR; 1-5, CGPA; 6.8-9.920 and we see some of the canditates have no record of research.

Lets look at a bar chart of with research and no research.

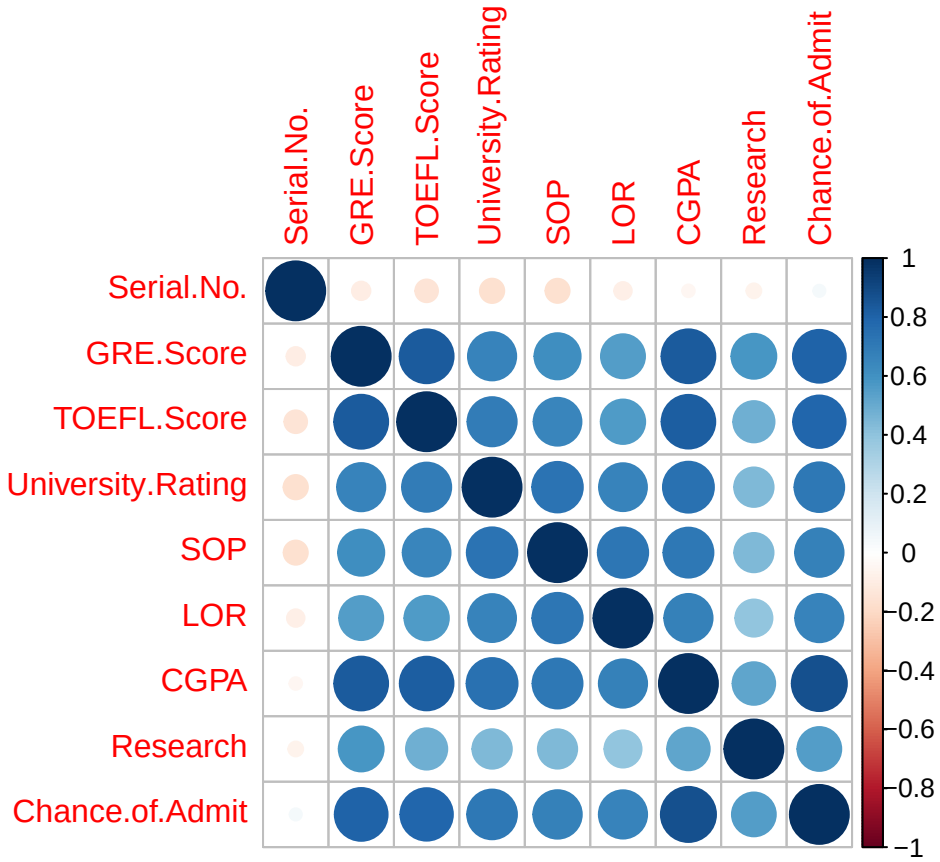


We see just a difference of 38 individuals. Lets look at plot of the other criteria in relation to Chance of

Admittance being distributed on research.



The corrolation of the criteria shows which of the criteria has strong relation with chance of admittance. Which in this case is CGPA.



Analysis

Data validation

```
set.seed(222, sample.kind = "Rounding")
```

```
## Warning in set.seed(222, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
test_index <- createDataPartition(Admission$GRE.Score, times = 1, p= 0.7, list = F)
train_set <- Admission %>% slice(-test_index)
test_set <- Admission %>% slice(test_index)
```

Modelling using linear regression

```
fit_lm <- lm(Chance.of.Admit ~ ., data = train_set)
y_hat_lm <- predict(fit_lm, test_set)

#calculating root mean square error for linear regression
rmse_lm <- sqrt(mean((y_hat_lm-test_set$Chance.of.Admit)^2))
rmse_lm
```

```
## [1] 0.06147279
```

Modelling using logistic regression

```
fit_glm <- glm(Chance.of.Admit ~ ., data = train_set)
y_hat_glm <- predict(fit_glm, test_set, type= "response")

#Calculating root mean square error for logistic regression
rmse_glm <- sqrt(mean((y_hat_glm-test_set$Chance.of.Admit)^2))
rmse_glm
```

```
## [1] 0.06147279
```

Modelling using randomForest

```
fit_rf <- randomForest(Chance.of.Admit~., train_set)
y_hat_rf <- predict(fit_rf, test_set)

#Calculating root mean square error for randomForest
rmse_rf <- sqrt(mean(y_hat_rf-test_set$Chance.of.Admit)^2)
rmse_rf
```

```
## [1] 0.002152295
```

Result

Looking at the table below, randomForest has the lowest root mean square error making it the best model

```
rmse <- matrix(c(rmse_lm, rmse_glm, rmse_rf),ncol=1,byrow=TRUE)
colnames(rmse) <- c("RMSE")
rownames(rmse) <- c("rmse_lm","rmse_glm","rmse_rf")
rmse <- as.table(rmse)
rmse
```

```
##                 RMSE
## rmse_lm  0.061472788
## rmse_glm 0.061472788
## rmse_rf  0.002152295
```

Conclusion

From the results we see that randomForest has the least RMSE which makes it the best model for predicting admittance into university.