Placement Data Analysis

## Abstract:

The project is based on a binary classification problem where our main goal is to predict whether a student will be hired or not in the campus recruitment drive being held. The data is taken from Kaggle’s website and could be found at [1]. Exploratory data analysis is performed a binary classifier is built using logistic regression with family set to binomial. The model gave us an accuracy of 87% on the test set. For exploratory data analysis box plots and pie charts are used. Based on the performance of our model, we can easily deploy this model to production so that based on academics of a person it could give us a prediction of whether the person will be hired or not. The project focuses on building a logistic regression classifier to predict whether a person will be recruited or not based on input features that are present in the data set. We have a binary classification problem. The target variable has two values, the student is either placed (Recruited) or not placed (Not Recruited).

## Introduction:

Having enough information in hand to know which type of specialty will lead to higher chances of getting a job, does percentage / GPA matters when applying for a job, is there any biasness based on gender, or what is the salary offered to different candidates is no less than a blessing. Having these insights one could easily prepare himself / herself for a particular job. This project is based on the analysis of campus recruitment that is being done. The target variable is binary variable which tells us that whether the person is hired or not. The rest of the document explains the data set, objectives and the libraries to be used.

## Literature Review:

As far as the school and the students are concerned, the campus placement activity is quite important. To validate the techniques, a work was assessed and forecasted using the classification algorithms Decision Tree and Random Forest algorithm to enhance the students’ performance (Hassan et al., 2018). This article also worked on the prediction of whether a person is hired or not based on features. The problem statement of this article is also the same as ours. The article has performed exploratory data analysis, and built two different classifiers; one is Random Forest Classifier and the second one is using logistic regression. The accuracy thus obtained on the test set is 85% and 80% for logistic regression and random forest classifier respectively (Shaikhina et al., 2018). The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for Decision tree is 84% and for the Random Forest is 86% (Douglas et al., 2011). Hence, from the above said analysis and prediction its better if the Random Forest algorithm is used to predict the placement results.

## Theory:

In order to be clear about the objectives of this project, a couple of questions have been devised which will be answered using the analysis performed.

• What factors determined a candidate’s placement? The summary of logistic regression will provide the answer to this objective.

• Is it important to have a certain percentage to acquire a job?

• Which degree of specialty is most in demand by businesses? It will be found using the percentage of hires based on specialty.

• Build a classifier using logistic regression which could classify whether a student will be recruited or not.

## Dataset:

This data collection contains student placement statistics from an XYZ campus. It covers percentages and specializations from secondary and higher secondary schools. The data set could be found at <https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement>. It also contains degree specialty, work experience, and pay offers to students who have been placed.

• Sr.No: It is an identifier for each record in data set.

• Gender: Represents the gender of a person.

• SSC-P: The percentage obtained in secondary school certification.

• SSC-B: What is education board? Central or Others.

• HSSC-P: The percentage obtained in higher secondary education.

• HSC-B: What is education board? Central or others.

• HSC-S: Specialization in higher secondary education.

• Degree Percentage: Percentage obtained in degree.

• Work Experience: Work experience of candidate in year.

• Specialization: The field of specialization.

• Test Score: The Entry Test Score Obtained.

• MBA: Percentage in MBA.

• Status: Target variable whether the person is recruited or not.

• Salary: Salary offered.

The data set is cleaned prior to performing any type of data analysis. The descriptive statistics of the cleaned data is attached below:

## 'data.frame': 215 obs. of 15 variables:  
## $ sl\_no : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ gender : chr "M" "M" "M" "M" ...  
## $ ssc\_p : num 67 79.3 65 56 85.8 ...  
## $ ssc\_b : chr "Others" "Central" "Central" "Central" ...  
## $ hsc\_p : num 91 78.3 68 52 73.6 ...  
## $ hsc\_b : chr "Others" "Others" "Central" "Central" ...  
## $ hsc\_s : chr "Commerce" "Science" "Arts" "Science" ...  
## $ degree\_p : num 58 77.5 64 52 73.3 ...  
## $ degree\_t : chr "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Sci&Tech" ...  
## $ workex : chr "No" "Yes" "No" "No" ...  
## $ etest\_p : num 55 86.5 75 66 96.8 ...  
## $ specialisation: chr "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...  
## $ mba\_p : num 58.8 66.3 57.8 59.4 55.5 ...  
## $ status : chr "Placed" "Placed" "Placed" "Not Placed" ...  
## $ salary : int 270000 200000 250000 NA 425000 NA NA 252000 231000 NA ...

Check for any missing values in the columns. Salary has 67 missing values.

## sl\_no gender ssc\_p ssc\_b hsc\_p   
## 0 0 0 0 0   
## hsc\_b hsc\_s degree\_p degree\_t workex   
## 0 0 0 0 0   
## etest\_p specialisation mba\_p status salary   
## 0 0 0 0 67

Remove rows with missing values.

## gender ssc\_p ssc\_b hsc\_p hsc\_b   
## 0 0 0 0 0   
## hsc\_s degree\_p degree\_t workex etest\_p   
## 0 0 0 0 0   
## specialisation mba\_p status salary   
## 0 0 0 0

Check for any duplicate values in the dataset. There are no duplicates in the data

duplicated(data.frame(data))

## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

summary(data)

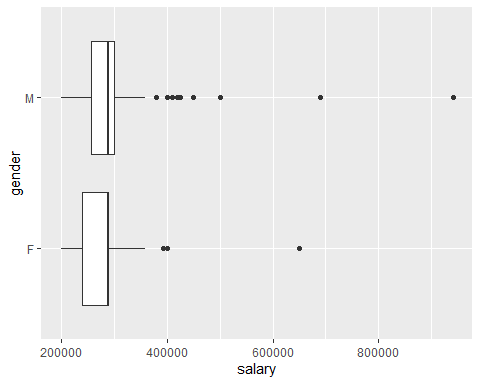
## gender ssc\_p ssc\_b hsc\_p   
## Length:215 Min. :40.89 Length:215 Min. :37.00   
## Class :character 1st Qu.:60.60 Class :character 1st Qu.:60.90   
## Mode :character Median :67.00 Mode :character Median :65.00   
## Mean :67.30 Mean :66.33   
## 3rd Qu.:75.70 3rd Qu.:73.00   
## Max. :89.40 Max. :97.70   
## hsc\_b hsc\_s degree\_p degree\_t   
## Length:215 Length:215 Min. :50.00 Length:215   
## Class :character Class :character 1st Qu.:61.00 Class :character   
## Mode :character Mode :character Median :66.00 Mode :character   
## Mean :66.37   
## 3rd Qu.:72.00   
## Max. :91.00   
## workex etest\_p specialisation mba\_p   
## Length:215 Min. :50.0 Length:215 Min. :51.21   
## Class :character 1st Qu.:60.0 Class :character 1st Qu.:57.95   
## Mode :character Median :71.0 Mode :character Median :62.00   
## Mean :72.1 Mean :62.28   
## 3rd Qu.:83.5 3rd Qu.:66.25   
## Max. :98.0 Max. :77.89   
## status salary   
## Length:215 Min. :200000   
## Class :character 1st Qu.:250000   
## Mode :character Median :288655   
## Mean :288655   
## 3rd Qu.:288655   
## Max. :940000

## Methodology:

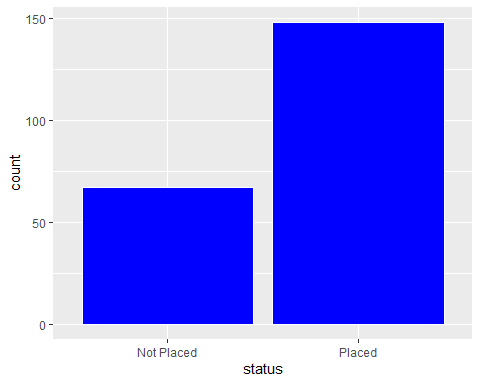
The flow of our project includes three stages.

* The first one is data pre-processing. This step includes preparing the data set for use. The data set is checked for missing values and it is found that the salary column has 67 null values. There are multiple ways to impute missing values. In this case, the method being used is replacing the missing values with mean of that particular column. The next step includes removal of column sl\_no which is unnecessary and is just an identifier which provides us no use full information. The final pre-processing step performed is to divide the data into training and testing set for training the classifier and testing the accuracy of a classifier.
* The second step is exploratory data analysis. In order to explore the dataset, table function, box plots, pie charts are used to identify interesting insights and to understand the data set.
* The final step is to build a logistic regression classifier which predicts whether a student is recruited or not based on his educational data points mentioned in data set description above. After building the classifier, the classifier is tested on the testing set and performance of model is evaluated using accuracy score, specificity, sensitivity, kappa statistics and confusion matrix. Since the first stage has been complete in the Data section, this section we will perform Exploratory data analysis and build the logistic regression classifier. First we create a boxplot to visualize thr distribution of salary against gender, that is Male and Female.

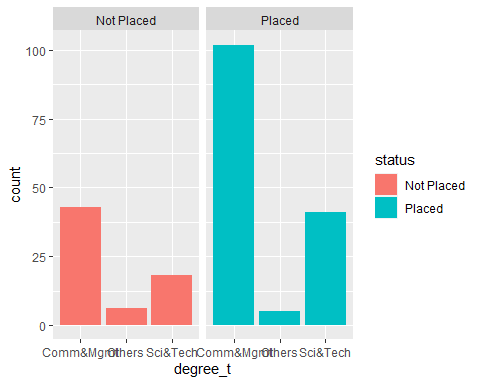
dataVis <- filter(data, salary != 0) #filter out 0 as data with 0 values have no placement yet  
  
options(scipen=999)  
  
ggplot(dataVis, aes(salary, gender)) + geom\_boxplot()

 Next, we perform an analysis of the distribution of students placed and those that have not been placed using a bar chart.

ggplot(data, aes(status)) + geom\_bar(fill = "blue", color = "white")

 Next, using a stacked barchart, we perform an analysis on the status of students, that is thos placed and those that were not placed, against the degree type.

ggplot(data, aes(degree\_t, fill = status)) + geom\_bar() + facet\_wrap(~status)

 Next step involve building the logistic regression classifier. First, we split data to 70% training set and 30% testing set.

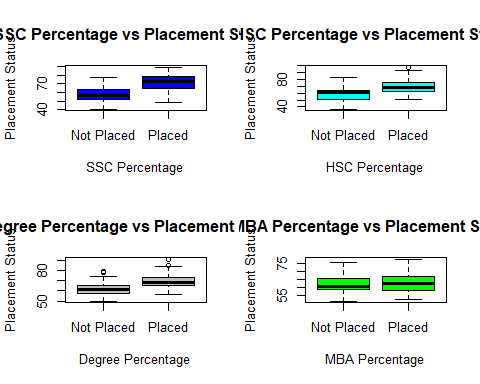
set.seed(888)  
data$status <- as.factor(data$status)  
split1<- sample(c(rep(0, 0.7 \* nrow(data)),   
 rep(1, 0.3 \* nrow(data))))  
train <- data[split1 == 0, ]   
test <- data[split1== 1, ]

Next, we fit the logistic regression model.

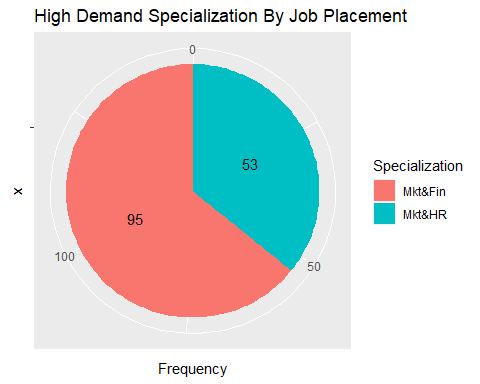
set.seed(123)  
model <- glm(status ~.,   
 data = train,   
 family = "binomial")

## Results:

Figure attached below shows a couple of box plots which explores different percenatges that are present in the dataset. Starting with the first one starting from left, we can see that the average SSC percentage of those students / persons who are recruited is higher as compared to those who are not recruited. The average SSC percentage of those recruited is around 70 while for those who are not recruited is below 60. Same is the case with HSC and Degree Percentage. However, the MBA percentage doesn’t show any differences. We can see that the average percentage for MBA for those who are recruited and those who are not recruited is approximately same. Although for those who are recruited, the average is highest to those who are not recruited but the difference in percenatage is not so high as compared to differences in percenatges observed between HSC, SSC and Degree Percenatges.

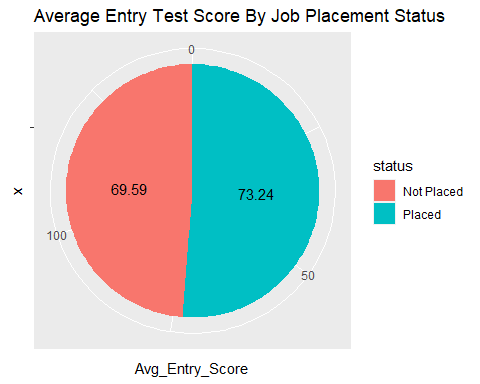


The pie chart below tells us that out of total people 148 who were recruited, 95 of them were having the specialization in Marketing and Finance while 53 of those who are recruited have a specialization in Marketing and HR. The percentage of people recruited with Marketing and finance percentage is 64% while for those who are recruited with Marketing and HR, the percentage is 35% approximately. This means Marketing and Finance is in higher demand as compared to Marketing and HR.



## # A tibble: 2 × 2  
## status Avg\_Salaries  
## <fct> <dbl>  
## 1 Not Placed 288655   
## 2 Placed 288655.

Exploring the average entry test of people from both categories that are those who are recruited and those who are not recruited, a pie chart is created. It can be seen that the average entry score for those who are recruited is around 73 approximately while for those who are not recruited, the average score is 69 approximately. This means, score of a person also affects the placement decision. Higher the test score increases the chances of getting hired. Lower score, decreases the chances of being hired.



Below is the result of logistic regression model attached with family set to binomial so that the problem is treated as a classification problem rather than regression problem. According to the summary, the SSC percentage, Degree Percentage, Work Experience, MBA Percentage and Gender Male are significant as their p values are below the significance level of 0.05. The residuals of the model are symmetric around 0 which gives us an indication that the model is better fitted in this case. The overall AIC of the model is 103.73. Lower the AIC, better the performance of the model.

##   
## Call:  
## glm(formula = status ~ ., family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.25479 -0.20232 0.06789 0.32917 2.09894   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -12.4426324927 6.2053264187 -2.005 0.04495 \*   
## genderM 1.2956995948 0.8133711833 1.593 0.11116   
## ssc\_p 0.2426073444 0.0558554015 4.343 0.000014 \*\*\*  
## ssc\_bOthers -0.0488944507 0.8239702086 -0.059 0.95268   
## hsc\_p 0.0769152331 0.0429496855 1.791 0.07332 .   
## hsc\_bOthers 0.7889205063 0.8909301048 0.886 0.37589   
## hsc\_sCommerce -1.6046384028 1.9179184811 -0.837 0.40279   
## hsc\_sScience -0.8581130301 1.9621913687 -0.437 0.66188   
## degree\_p 0.1382905836 0.0585639513 2.361 0.01821 \*   
## degree\_tOthers -1.4248515190 1.8160964572 -0.785 0.43271   
## degree\_tSci&Tech -1.4873256837 0.9294079952 -1.600 0.10953   
## workexYes 1.8076810407 0.8672220233 2.084 0.03712 \*   
## etest\_p -0.0124277310 0.0290518735 -0.428 0.66881   
## specialisationMkt&HR -0.9325942689 0.7009650706 -1.330 0.18337   
## mba\_p -0.2432059106 0.0765803029 -3.176 0.00149 \*\*   
## salary 0.0000008972 0.0000041898 0.214 0.83044   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 190.326 on 150 degrees of freedom  
## Residual deviance: 71.735 on 135 degrees of freedom  
## AIC: 103.73  
##   
## Number of Fisher Scoring iterations: 7

The logistic regression model is tested on the test set and the result thus obtained is attached below. It can be seen that the accuracy of the model on the test set is 87% which is good in this case. The confusion matrix tells us that 13 + 43 = 56 observations are correctly classified whereas 5 + 3 = 8 observations are miss classified in this case. The confidence interval tells us that we are 95% confident that the accuracy of our model lies between 76 and 94% approximately. The kappa statistic is 68% which is good and could be improved further. Higher value of kappa statistics means a model with better performance. The specificity and sensitivity of the model is also good which is 93% and 72% respectively.

## Confusion Matrix and Statistics  
##   
##   
## predicted.classes Not Placed Placed  
## Not Placed 13 3  
## Placed 5 43  
##   
## Accuracy : 0.875   
## 95% CI : (0.7685, 0.9445)  
## No Information Rate : 0.7188   
## P-Value [Acc > NIR] : 0.002425   
##   
## Kappa : 0.68   
##   
## Mcnemar's Test P-Value : 0.723674   
##   
## Sensitivity : 0.7222   
## Specificity : 0.9348   
## Pos Pred Value : 0.8125   
## Neg Pred Value : 0.8958   
## Prevalence : 0.2812   
## Detection Rate : 0.2031   
## Detection Prevalence : 0.2500   
## Balanced Accuracy : 0.8285   
##   
## 'Positive' Class : Not Placed   
##

## Conclusions:

The logistic regression model trained gave us an accuracy of 87% on the test set which is excellent in this case. We found out that the the SSC, HSC, Degree and MBA percentage affects the decision of being hired or not. The average test score of those who are being hired is greater than those who are not hired. The most demanding specialization in this data set is Marketing & Finance. The summary of logistic regression also tells us that male people have higher chances of being hired. More and more insights could be generated using the available.

## References:

Douglas, P. K., Harris, S., Yuille, A., & Cohen, M. S. (2011). Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. Neuroimage, 56(2), 544-553.

Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., & Sarker, K. U. (2018, August). Student academic performance prediction by using decision tree algorithm. In 2018 4th international conference on computer and information sciences (ICCOINS) (pp. 1-5). IEEE.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2019). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomedical Signal Processing and Control, 52, 456-462.