CAMPUS PLACEMENT DATASET

Presented by

TOLULOPE AYODEJI ALE

22/05/2022

## TABLE OF CONTENTS

TABLE OF CONTENTS ii

CHAPTER 1: INTRODUCTION 1

1.1 Objectives 1

1.2 Dataset Description 2

CHAPTER 2: EXPLORATORY DATA ANALYSIS (EDA) 7

2.1 Correlation Plot 7

2.2 Score Distribution 11

2.3 Gender Difeerences on Performance Score using T-test. 14

2.4 Academic Performance Impact on Placement using T-test 21

CHAPTER 3: PREPROCCESSING 29

3.1 Splitting the data 29

3.3 Subset Selection 30

1. Forward selection 30

2. Backward selection 31

3. Best subset selection 32

CHAPTER 4: PREDICTIONS 34

4.1. Resampling 34

4.2. Logistic Regression 34

4.3. Linear Discriminant Analysis 37

4.4. Generalize Linear Model 38

4.5. Support Vector Machine: Linear 39

4.6. Quadratic Linear Discriminant 40

4.7. K-Nearest Neighbors 42

4.8. Random Forest (Ranger) 43

4.9. Gradient Boosting 45

CHAPTER 5: EVALUATION 47

5.1. Model Accuracy, Kappa, and Confusion Matrix 48

5.2. ROC Curve 50

CHAPTER 6: CONCLUSION 51

## CHAPTER 1: INTRODUCTION

CAMPUS RECRUITMENT PLACEMENT DATASET

A major quest for any student is to get a good job placement after graduation, but the criteria which an employer uses to determines if a graduate is a good fit for employment may not be the same with what makes a good student. Education can be subdivided into lower degree and upper degree. The quality of education in the lower degree classes determines the qualify of the foundation a student is getting for future classes. Also, some professional degree programs such an MBA could make for a distinguishable factor during placement. The dataset for this study was obtained from Kaggle

## Objectives

We will be doing some analysis on a CAMPUS RECRUITMENT dataset obtained from Kaggle repository. The dataset consists of 14 attributes, 1 target attribute, 215 instances and 2 classes. We will be analyzing the dataset to predict attributes that can influence if a student will be placed or not. We will do exploratory data analysis and visualization using R.

**Research Questions**

1. Which variable influenced a candidate in getting placed?
2. Does previous work experience matter for one to get placed?
3. Does high percentage in employment test matters for one to get placed?
4. Which Degree type and MBA specialization is in much demand?

We will attempt to answer the research questions using some statistical learning techniques by determining the significance -p value of the attributes.

Also, We will divide the dataset into a train and test set to build a model that could predict if a student will be placed based on the attributes.

* 1. **Dataset Description**

The dataset has 14 attributes, 1 target attribute, 215 instances and 2 classes as shown in the table below. 148 students were placed and 67 were not placed.

|  |  |
| --- | --- |
| **Input Attributes** | **Values** |
| gender | Gender (1=male; 0=female) |
| ssc\_p | Secondary education percentage |
| ssc\_b | Board of education (central/others) |
| hsc\_p | Higher secondary education percentage |
| hsc\_b | Board of education (central/others) |
| hsc\_s | Specialization in higher secondary education |
| degree\_p | University Degree percentage |
| degree-t | University Degree type (Undergraduate degree field) |
| workex | Work experience |
| etest\_p | Employability test percentage |
| specialization | Post-graduation (MBA) specialization |
| mba\_p | MBA percentage |
| status | Status of placement (Placed/Not placed) |
| salary | Salary offered by corporate to candidates |

For analysis on this dataset, We will be loading the following R libraries as seen in the code snippet.



Using the summary function, we printed out the statistics of each attribute on the dataset.

* The dataset has two classes: numeric and character.
* 215 instances and 14 attributes and the serial number (sl\_no).
* The salary attribute has 67 NAs
* 8 attributes has a character data type and 6 has a numeric data type

A screenshot of a computer screen

Description automatically generated with medium confidence

The below code printed out the count of Nas in the dataset. The Nas are on the salary attribute, it indicate students who are not receiving salary. Hence, they are replaced with the value “0”.

The 67 students matched the number of students that are yet to be placed, because only students with placement received salaries.

Background pattern

Description automatically generated

This is a multi-class dataset comprising of numbers and character datatypes with a dimension of 215 observations and 15 attributes.

Text

Description automatically generated

Visualizing the top 6 rows using the Head function

Graphical user interface, text

Description automatically generated

Visualizing the bottom 6 rows using the Tail function

Graphical user interface, text

Description automatically generated

* 1. **Data Cleaning**

Firstly, we will be renaming the column headers using the below line of code

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

## CHAPTER 2: EXPLORATORY DATA ANALYSIS (EDA)

With EDA we hope to make some inference on the dataset and explore answering some of the research questions such aswhich factor influenced a candidate in getting placed, does previous work experience matter for one to get placed, does high percentage in employment test matters for one to get placed and which degree type and MBA specialization is in much demand.

## Correlation Plot

Correlation plot shows the relationship between the numeric variables

A picture containing text

Description automatically generated

To visualize the correlation heatmap, we installed and loaded the ggcorrplot package.

Text

Description automatically generated

Chart, histogram

Description automatically generated

From the correlation heatmap, we can make two inferences:

1. It can be observed that the Secondary Education Percentage, Higher Secondary Education Percentage and Degree Percentage has medium correlation. This could suggest that students who performed well in secondary education will likely perform well in further education.
2. The correlation between Employability Test Percentage and other education percentage score is low. This could suggest that the employability test was more practical, the implication of that indicates employers are looking for more than a higher academic grade.
   1. **Score Distribution**

Visualizing the distribution of all the scores across the education levels should highlight if the trend observed in the correlation plot persist.

1. MBA percentage score

Text

Description automatically generated

Chart, histogram

Description automatically generated

1. Employee Test Percentage Score

Text

Description automatically generated

Chart, histogram

Description automatically generated

## 

## University Degree Percentage Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Higher Secondary Education Percentage Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Secondary Education Percentage Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## It was observed that the score distributions got narrower as the students move from Secondary Education to MBA.

## The score became more concentrated between the range 60-70 as the students’ progress in education.

## We can also observe that the Employee Test Score has a different distribution, which support our earlier hypothesis on the correlation plot. The distributions were almost equal.

* 1. **Gender Differences on Performance Score Using T-test**

## We will explore the differences between the genders on the performance score at each academic level.

## T-tests in R is one of the most common tests in statistics. So, we will be using it to determine whether the means of the two genders are equal to each other. The assumption for the test is that both genders are sampled from normal distributions with equal variances. The null hypothesis is that the two means are equal, and the alternative is that they are not equal. It is known that under the null hypothesis, we can calculate a t-statistic that will follow a t-distribution with ****n1 + n2 – 2 degrees of freedom****. The p-significance level less than <= 0.05 will help determine the significance level of a gender against the other.

## Load the library ‘rstatix’.

## University Degree Percentage Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Text Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DegreePercent | F | M | 76 | 139 | 2.431405 | 132.0024 | 0.0164 | \* |

## The p-value score reject the null hypothesis. Female students have a significantly higher score than their male counterpart. With the p-value score, it provide a statistical support that female students with a much higher score have a higher potential of being placed than their male counterpart.

## MBA Percentage Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Text Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MBApercent | F | M | 76 | 139 | 4.725214 | 166.8781 | 4.9e-06 | \*\*\*\* |

## Female again score significantly higher than their male counterpart. With a p-value significantly lower could mean they have a higher probability of being placed. The null hypothesis is again rejected by this result.

## Employee Percentage Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Text Description automatically generated

## Graphical user interface, text, application Description automatically generated

## We don’t have enough statistical significance to reject the null hypothesis.

## Higher Secondary Education Percentage

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Text Description automatically generated

## Graphical user interface Description automatically generated with medium confidence

## Secondary Education Percentage

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Text, logo Description automatically generated

## Graphical user interface Description automatically generated with medium confidence

## From the statistical analysis of gender on performance of each student across the academic levels, it could be observed that:

## Females have a significantly higher score than males at the University Degree and MBA levels.

## There was no significant difference in gender performance at Secondary, Higher Secondary, and Employability Test.

## The result was supported by the p-value for each academic level.

## Academic Performance Impact of Placement using T-test

## It is expected that academic performance should have the most influence on a student getting placed. The earlier visualization on score distribution showed that the performance of student on the employability test was average, from this analysis we should have a clear understanding which level of education has influence on student placement.

## Secondary Education

## Text Description automatically generated with medium confidence

## Chart, histogram Description automatically generated

## Text Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SecEducationPercent | Not Placed | Placed | 67 | 148 | -11.33316 | 132.0192 | 0 | \*\*\*\* |

## Higher Secondary Education

## Text Description automatically generated

## Histogram Description automatically generated

## Text, logo Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| HigherSecEduPercent | Not Placed | Placed | 67 | 148 | -8.043664 | 120.8047 | 0 | \*\*\*\* |

## University Degree Score

## Text Description automatically generated with medium confidence

## Chart, histogram Description automatically generated

## Text, logo Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DegreePercent | Not Placed | Placed | 67 | 148 | -8.054153 | 130.335 | 0 | \*\*\*\* |

## MBA Score

## Text Description automatically generated with medium confidence

## Chart, histogram Description automatically generated

## Text, logo Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| MBApercent | Not Placed | Placed | 67 | 148 | -1.139201 | 131.2069 | 0.257 | ns |

## Employability Test Score

## Text Description automatically generated

## Chart, histogram Description automatically generated

## Text, logo Description automatically generated

| **.y.** | **group1** | **group2** | **n1** | **n2** | **statistic** | **df** | **p** | **p.signif** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| EmpTestPercent | Not Placed | Placed | 67 | 148 | -1.980112 | 145.392 | 0.0496 | \* |

## Looking at the p-value score and analyzing the p-significance for each education level, it could be observed that:

## The score differences between those who received an offer and those that do not were highly significant at Secondary, Higher Secondary, and University Education Level.

## The Employability Test has a significant score

## But the MBA score is not significant.

## From the above statistical analysis, we could assume that lower-level education has more influence than higher-level education on the placement outcome. Hence, we plot a scatter plot of lower-level education and higher-level education visualize the score distribution across the percentile.

## Text Description automatically generated

## Chart, scatter chart Description automatically generated

## The following was observed:

## Almost all the students that did well in the lower education were placed; all the students in the top 25th percentile received placement regardless of their higher education performance.

## Majority of the student in the bottom 25th percentile across the lower education did not receive a placement.

## The hypothesis that performances in both Secondary and Higher Secondary Education having more influence on a student getting placed is supported by this analysis.

## CHAPTER 3: PREPROCCESSING

## Splitting the data

## Now that the structure of the data is clear, we will proceed to split the data into training and testing set. We used the stratified split approach because we want even distribution of the variable Status in both the training and testing set.

## The variable Salary will be dropped because it is not needed in our analysis, it can only be referenced after a placement offer has been received.

## The caret library package was loaded.

## Text Description automatically generated

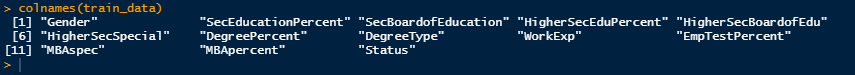
We partitioned the data in train and test set, with training set having 70% and testing set 30%.

Train\_data table is the training set and Test\_data table is the testing set.

Text

Description automatically generated

Now we have 13 variables, the 12 predictors and the response variable. Subset selection method will be applied on the 12 variables to choose that with the best model output.

****

* 1. **Scaling the data**

The data was scaled using the below code.

The three options c(“center”, “scale”, “nzv”) does scale and center. Method = “center” subtracts the mean of the predictor’s data from the predictor values while method = “scale” divides by the standard deviation. And “nzv” basically excludes variables that have near zero variance, meaning they are almost constant and most likely not useful for prediction.

Text

Description automatically generated

## Subset Selection

## To select the best model for further analysis, we will be implementing 3 subset selection approach to select the best variable.

## Forward selection

## Table Description automatically generated with low confidence

## 

## Backward selection

## Table Description automatically generated

## 

## Best subset selection

## Table Description automatically generated

## 

## For this dataset, the best one-variable through seven-variable models are identical for forward selection, backward selection, and best subset.

## EVALUATING SUBSET SELECTION MODELS

## RSS and Adjusted R-squared

## However, plotting the RSS and Adjusted R-squared for the forward selection models shows that the best model uses 11 to 13 variables.

## Text Description automatically generated

## Chart Description automatically generated

## We will printed out the RSS and Adjusted R-Squared values for the three models for comparison.

## RSS values

## Graphical user interface, text, email Description automatically generated

## Adjusted R-Squared values

## Graphical user interface, text Description automatically generated

## The above output shows that the three models predicted similar models. The least RSS is in 13 variable model and the highest adjusted r-squared is in 11 variable model. Hence, we will be using the 12 predictors for my prediction.

## CHAPTER 4: PREDICTIONS

## Resampling

## We will perform a 10-fold Cross Validation, use each of the 10 parts as a testing set for the and train on the remaining 9 parts.

## 

## Logistic Regression

## We fit a logistic regression model to predict the Status variable using all the 12 predictors. We will use the function glm to fit the model and specify family = binomial so that R can can fit a logistic regression.

## Text Description automatically generated

## The smallest p-value is SecEducationPercent with a positive coefficient. It is an indication that students that usually get placed have a high score in secondary education.

## To check the accuracy of our model on the training set.

## Text Description automatically generated

## Our accuracy on the training set is 0.8792, kappa is 0.7044.

## We will now apply the model to make prediction using the test set, we then obtain the probability between being Placed and Not Placed, add the logistic regression prediction to the test\_data and output the confusion matrix for the prediction.

## Logistic Regression accuracy on the test\_data set is 0.8281 and kappa is 0.6054

## Text Description automatically generated

## Linear Discriminant Analysis

## Graphical user interface, text Description automatically generated

## Accuracy on training set is 0.8940, Kappa is 0.7444

## We will now apply the model to make prediction using the test set, we then obtain the probability between being Placed and Not Placed, add the LDA prediction to the test\_data and output the confusion matrix for the prediction.

## Text Description automatically generated

## LDA model accuracy on the test set is 0.8594 and kappa is 0.6771. It is higher than the logistic regression accuracy.

## Generalize Linear Model

## Text Description automatically generated with low confidence

## The training accuracy was 0.8761, kappa was 0.7138 and testing accuracy was 0.8281, kappa was 0.6054.

## Text Description automatically generated

## Support Vector Machine: Linear

## We will try one more linear algorithm to see if our result will improve.

## Text Description automatically generated with medium confidence

## Training accuracy is 0.8883 and Kappa is 0.7293

## Testing accuracy is 0.8594 and Kappa is 0.6771

## The Linear SVM model and LDA model both have the same accuracy on the testing set.

## Text Description automatically generated

## To improve the accuracy on the testing data, we shall use more flexible models.

## Quadratic Linear Discriminant

## The result of the model on the training set shows that QDA performed poorly when compared to the linear models.

## Text Description automatically generated

## Let’s see how the model does on the testing set. We could say that the QDA did not capture the true representation of our data.

## Text Description automatically generated

## K-Nearest Neighbors

## KNN printed the accuracy of the model on the training set for different values of K at 5,7,9,11,13 since we set tunelength equals 5. The average will give us the result on our training set.

## Text Description automatically generated

## Text Description automatically generated

## The accuracy on the testing set is 0.8281 and Kappa is 0.544

## Random Forest (Ranger)

## Grow a random forest on the training data. For each observation of interest (test data), the weights of all training observations are computed by counting the number of trees in which both observations are in the same terminal node. For each test observation, grow a weighted random forest on the training data, using the weights obtained in step 2. Predict the outcome of the test observation as usual. In total, n+1 random forests are grown, where n is the number observations in the test dataset.

## Graphical user interface Description automatically generated with medium confidence

The Random Forest model was evaluated using 5 different number of predictors ‘mtry = (2,5,8,11,14) until m=p which is Bagging. The highest accuracy on the training set when all the predictors were considered m=8 meaning that the Random Forest approach will produce a better predictor than Bagging.

The accuracy on the test set is 0.8281 and Kappa is 0.5707.

For an ensemble model, it performed worse than the linear models of LDA and SVM.

## Text Description automatically generated

## Gradient Boosting

## Graphical user interface Description automatically generated

## A picture containing text Description automatically generated

## Gradient boosting is an ensemble learning method, the accuracy on the test set is 0.8281 kappa is 0.6054.

## The plot of the summary of the result show that the top 4 variables are Secondary education, Higher secondary education, University degree percentage and MBA percentage.

## Chart Description automatically generated with medium confidence

## EVALUATION

## We evaluated our models by comparing the accuracy and the kappa. Accuracy is the percentage of accurate predictions out of all samples and kappa is the accuracy that would be generated by chance. We also considered the amount of false positive and false negative in the model.

## Model Accuracy, Kappa, and Confusion Matrix

## Text Description automatically generated with low confidence

## Chart, box and whisker chart Description automatically generated

The accuracy of the model indicates that most of the models have similar performance. But to determine which models’ prediction error is costlier than the other, we will look at the confusion matrix and identify the amount of false negative and false positive in each prediction. False Negative happens when the model incorrectly predicts that someone would not be placed, while False is when the model incorrectly predicts that someone would be placed.

* The model with the least False Negative is KNN having a prediction accuracy of 82.8% (10 false positive and 1 false negative). With KNN model a student, the likely hood that a student will miss the opportunity of getting placed due to model error is low compared to other models. This model favors the students.
* The model with the least False Positive is LDA. It has a prediction accuracy of 85.9% (4 false positive and 5 false negative). This model maximizes accuracy, but the output could be costly on students.

Using the tidy function to compare the two models.Graphical user interface, text, application

Description automatically generated

Looking at the p-value, we can assume that there is no statistical reasoning to assume that one model is better than the other.

* 1. ROC Curve

Lastly, let us compare the ROC curve of the two models. From the ROC curve, we can conclude that the LDA is a better model.

Chart, line chart

Description automatically generated

1. **CONCLUSION**

From the result of our statistical analysis, we could conclude the following.

* Secondary education percent is the most important variable with influence on the placement of a student. Students that did well in secondary education got placed.
* The topmost important features are secondary education percent, higher secondary education percent, university degree percent and MBA percentage score.
* Employability test score does not correlate with the academic score. That is the employability test score is a practical test.
* As the education level increases, female students got higher score than the male students.
* The LDA model favors the students because if has the least false negative model errors and the KNN favors the recruiters because if has the least false positive errors.

1. **REFRENCES**

* <https://www.kaggle.com/benroshan/factors-affecting-campus-placement>
* 10.5281/zenodo.5109070 ISBN:978-93-5426-386-6@2021 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam. Proceedings of the National Conference on Emerging Computer Applications (NCECA) -2021 Vol.3, Issue.1
* <https://cran.r-project.org/>
* https://www.analyticsvidhya.com/blog/2021/03/introduction-to-k-fold-cross-validation-in-r/