Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions

Conference Paper · October 2020

DOI: 10.1109/ICCCNT49239.2020.9225441

CITATIONS
19

READS
876

2 authors:

Jumana Nagaria
Amity University Dubai
3 PUBLICATIONS **33** CITATIONS

SEE PROFILE

Senthil Velan S.
CMR Institute of Technology
34 PUBLICATIONS **177** CITATIONS

SEE PROFILE

Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions

Jumana Nagaria
Department of Computer Science and Engineering
Amity University Dubai
Dubai, UAE
jumanaN2@amitydubai.ae

Senthil Velan S
Department of Computer Science and Engineering
Amity University Dubai
Dubai, UAE
svelan@amityuniversity.ae

***Abstract*—In Exploratory Data Analysis (EDA) the given large data is visually analyzed to extract the embedded deep. Application of the technique has a wide range and aids in the informed decision making abilities of the managers. In an educational institution, the success of its imbibing model is usually measured using the career opportunities of the graduates. Hence, the placement data has an important relevance for the future plan and growth. Quite a good amount of information can be gained by all the stakeholder by carefully looking at this information. In this context, the technique of EDA can be used to visually analyze the placement of students in a higher educational institution. In this paper the data about the placement of student is visually analyzed to generate inferences using mathematical models. Based on the study it was found that student with MBA specialization in *Mkt&Fin* are highly placed, a vast majority of the students have Commerce and Management degrees. The score on the employability test don't seem to have a major impact on the placement of students.**

Keywords—Exploratory Data Analysis, R Language, Random Forest Algorithm, Decision Tree Algorithm, Campus placement

I. INTRODUCTION

Rich and high volume data is the modern fuel that possesses inherent characteristics for driving today's intelligent decision-making abilities of smart businesses and services. When comparing with the energy sector, unprocessed raw data is equivalent to crude oil. The fuel that powers the internal combustion engines is the intelligent information that is processed from the raw data. Similar to the extraction of different products using fractional distillation of crude oil, the removal of intelligent information at different levels will improve the decisions of varying levels across the business unit.

Exploratory data analysis (EDA) is a process by which the given data set is analyzed to interpolate useful information. The process commonly depicts the data in a visual form enabling betting understanding and to adept informed decision making of the business entities.

Campus placement is considered to be an important parameter for measuring the success of the learning model developed by each higher educational institution. It provides measurable and decipherable confidence of the learning methodology imbibed by a differently practiced model. It also includes information for future stakeholders (prospective students) to make decisions during their career plans. Hence, the placement data obtained from the higher educational institution will help in understanding the success of study for the enrolled and prospective students.

The placement data obtained about the higher educational institution can be used to analyze using relevant models statistically. Visualization of the placement data and its factors helps in understanding the success and grade of the institutions. In this scenario, EDA could be a relevant methodology for its application to the placement data. Later, using the visualized entities, an informed set of inferences can be generated for a better understanding of the data. This can also play a crucial role in further doing analytics on the placement data.

The rest of the paper is organized as follows: The next section explains the steps of a generic EDA process and further on the application of this process to the placement data. Section 3 explains about the visualization of the placement datasets using the different kinds of mathematical visualization models. Based on the obtained visualization models, a set of inferences are explained in Section 4. Section 5 elucidates the prediction of data using random forest and decision tree algorithms. Finally, Section 6 provides a brief conclusion of the work done in this research paper.

A. Applying EDA For Placement Data

1) Process Flow of EDA

EDA is the initial step of deciphering data by first showing the visual representation using different tools available in a data processing tool. It helps in summarizing your findings and display the data with graphics and can help in interpretation and finding the underlying patterns or trends[1]. Humans, by nature, are attracted to visuals, and it is one of the fastest methods to recognize a result rather than reading the text. Visualizations can simply explain any complex idea.

The most crucial step is to study the data thoroughly- identify if it is structured data or unstructured data, observe the number of columns and rows of structured data, read the structure of the data, find out the important columns that are useful in the dataset, find out the number of null values or missing data and decide either to replace it or remove it from the dataset. The data is first cleaned and trimmed using inbuilt functions of the tool such as *filter*, *select*, *group by* and others.

Fig 1 displays the generic process of applying EDA to a different group of datasets. The significant subdivisions are the categorical data and numerical- continuous or discrete data. Followed along with its type of plots such as histogram can be plotted through continuous, whereas a box plot can be plotted between the continuous and categorical data.

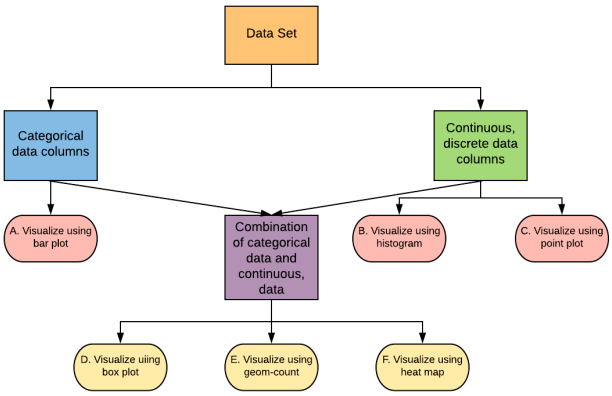

Fig. 1. A Generic Process Flow of EDA

2) Application to Placement Data

The dataset is about campus recruitment shows the influencing factors of academic and employability that helps in the placement of a student. The dataset was downloaded from Kaggle[2]. The Campus Recruitment data includes information regarding gender, Board of education, secondary schools, higher secondary schools percentage, and specialization. Along with the degree type, work experience, degree percentage, the salary of the placed students. The dimension of the dataset is 215 records and 15 columns.

```
  sl_no gender ssc_p    ssc_b hsc_p    hsc_b     hsc_s degree_p  degree_t workex etest_p
1     1      M 67.00   Others 91.00   Others Commerce     58.00   Sci&Tech     No    55.0
2     2      M 79.33 Central 78.33   Others   Science     77.48   Sci&Tech    Yes    86.5
3     3      M 65.00 Central 68.00 Central      Arts     64.00 Comm&Mgmt     No    75.0
4     4      M 56.00 Central 52.00 Central   Science     52.00   Sci&Tech     No    66.0
5     5      M 85.80 Central 73.60 Central Commerce     73.30 Comm&Mgmt     No    96.8
6     6      M 55.00  Others 49.80  Others   Science     67.25   Sci&Tech    Yes    55.0
  specialisation mba_p      status salary
1         Mkt&HR 58.80      Placed 270000
2        Mkt&Fin 66.28      Placed 200000
3        Mkt&Fin 57.80      Placed 250000
4         Mkt&HR 59.43 Not Placed     NA
5        Mkt&Fin 55.50      Placed 425000
6        Mkt&Fin 51.58 Not Placed     NA
```

Fig. 2. Insight into the data set

Now to look at the structure of the dataset to get an idea about the campus data, we can use *str()* or *summary()* command in R. Table I shows the structural information about the dataset.

The second most essential step is the data preprocessing part where the *NA* or *Null values* are handled[3]. Data preprocessing allows us to deal with inconsistent values such as the address in place of phone number and removable of duplicate values and also the extra spaces by applying *str_trim()* method on the character data.

TABLE I. STRUCTURE OF THE CAMPUS PLACEMENT DATA

S. No.	Column		
	Column Name	***Data type***	***Description***
1	sl_no	Integer	Serial Number
2.	Gender	String	Male='M',Female='F'
3	ssc_p	Decimal	Secondary Education percentage- 10th Grade
4	ssc_b	String	Board of Education- Central/ Others
5	hsc_p	Decimal	Higher Secondary Education percentage- 12th Grade
6	hsc_b	String	Board of Education- Central/ Others
7	hsc_s	String	Specialization in Higher Secondary Education
8	degree_p	Decimal	Degree Percentage
9	degree_t	String	Under Graduation(Degree type)- Field of degree education
10	Workex	Boolean	Work Experience
11	etest_p	Decimal	Employability test percentage (conducted by the college)
12	Specialization	String	Post Graduation(MBA)- Specialization
13	mba_p	Decimal	MBA percentage
14	Status	String	Status of placement- Placed/Not placed
15	Salary	Integer	Salary offered by corporate to candidates

One of the ways in which we perform data Preprocessing was by check if there were any NA values in the campus dataset:

```
> #to check if there is any null value
> table(is.na(data))

FALSE  TRUE
 3158    67
> View(data)
> table(is.na(data$salary))

FALSE  TRUE
  148    67
```

Fig. 3. *NA* values in tabular form

So we concluded that the salary column has all 67 null values that mean out of 148 placed students; we don't know the salary of 67 students who have not been placed.

EDA is comparable to the storytelling of the analyzed data. The data was analyzed using R. R is open-source and can import many formats of data[4]. R can handle classes, graphics, high-level statistic functions, loops, matrices, hash tables, expressions, and many more. EDA is a state of mind where every idea can be investigated.

There are four types of EDA[5]:

1. Univariate non-graphical which is a tabulation of the categorical data. The below figure represents the frequency of the student who is placed and not placed.

```
> table(data$status)

Not Placed     Placed
        67        148
```

Fig. 4. Tabulation of the status column

2. Univariate graphical using histograms, and box plots
3. Multivariate non-graphical represents the correlation between data, Cross-tabulation. The below figure represents the placement of the student for gender.

```
   Not Placed Placed
F          28     48
M          39    100
```

Fig. 5. Tabulation of the status column with gender

```
> table(data$gender,data$hsc_s,data$workex)
, ,  = No

   Arts Commerce Science
F    4        28      22
M    2        49      36

, ,  = Yes

   Arts Commerce Science
F    2        12       8
M    3        24      25
```

Fig. 6. Cross-tabulation of gender,wokex,hsc_p

The above figure shows the cross-tabulation between gender, work experience, and the stream of 12th grade.

4. Multivariate graphical are univariate graphs by different categories using a scatterplot, side by side box plots.

II. VISUALIZED DATASETS

EDA is applied to the dataset. The plots and helps us to gain insight into the dataset[6]. The package 'ggplot2' was used to create graphics by providing the data and mapping variables to aesthetics[7][8].

A. Bar Plot

A bar plot provides a comparison between cumulative tools across several groups.

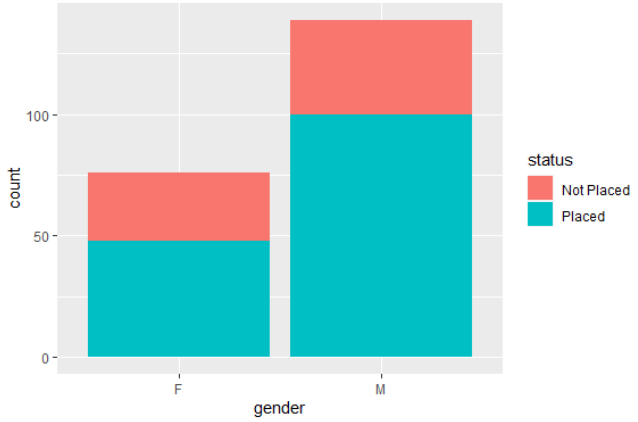

Fig. 7. Bar Graph of Gender column filled with status

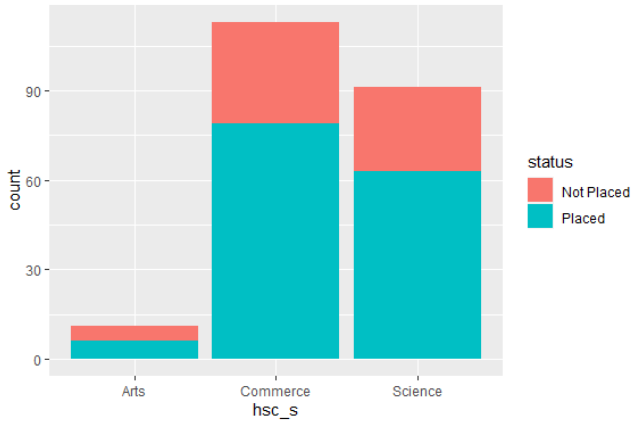

Fig. 8. Bar Graph of hsc_s column filled with status

B. Histogram Plot

A histogram plot breaks the data into bins and shows its frequency distribution of these bins.

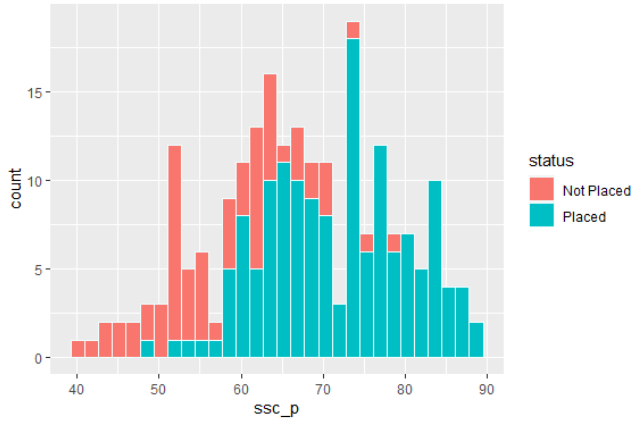

Fig. 9. Histogram of scc_p column filled with status

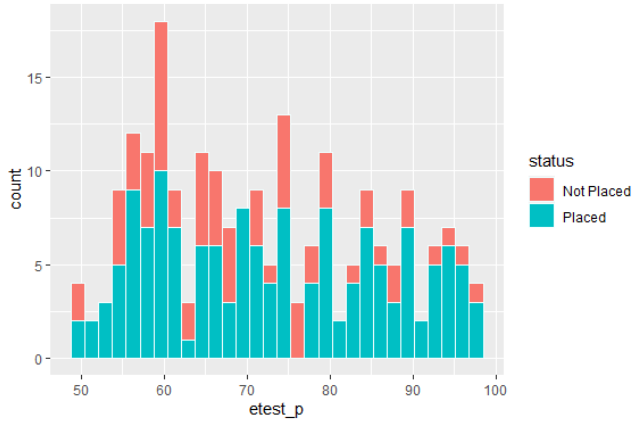

Fig. 10. Histogram of estest_p column filled with status

C. Point Plot

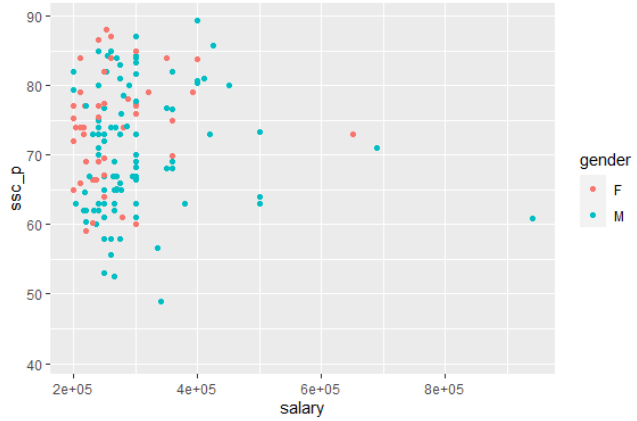

Fig. 11. Scatter Plot between ssc_p & salary split by gender

The Point plot is also known as Scatter Plot. *Geom_point()* function helps in understanding how one variable changes with respect to other variables. It is used for two continuous values.

D. Box Plot

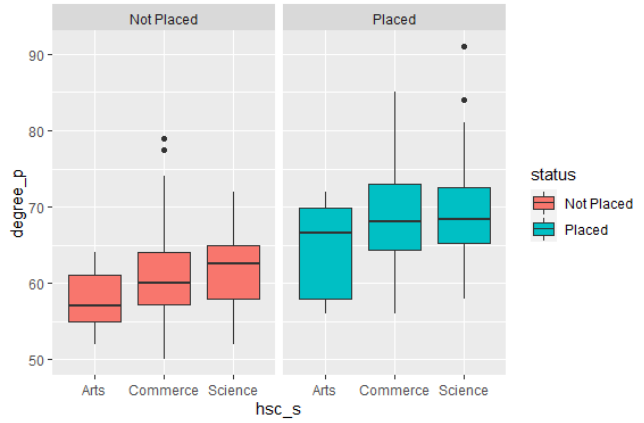

Fig. 12. Box Plot between degree_p & hsc_s separated by status

Geom_box() can be used to visualize the spread of data and derive inference accordingly[9]. A box plot shows five most important numbers- the minimum, the 25th percentile, the median, the 75th percentile, and the maximum[10].

E. Geom_count

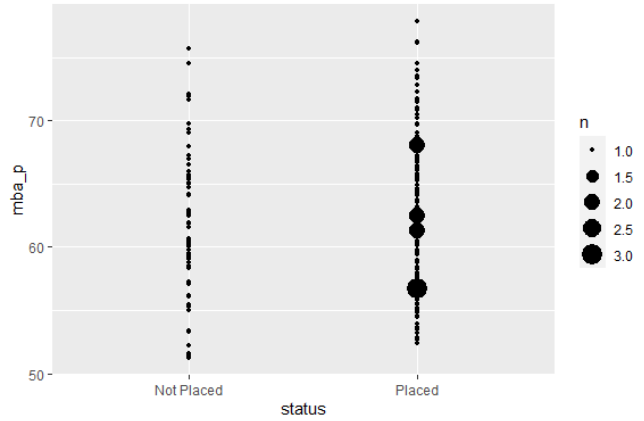

Fig. 13. Geom_count between status and mba_p

Geom_count() is useful in scenarios of discrete data or overpotting. It counts the number of observations at each feature and maps the count to point area[11]. Figure n is the number of representative of the status column with respect to the MBA percentage.

F. Geom_tile

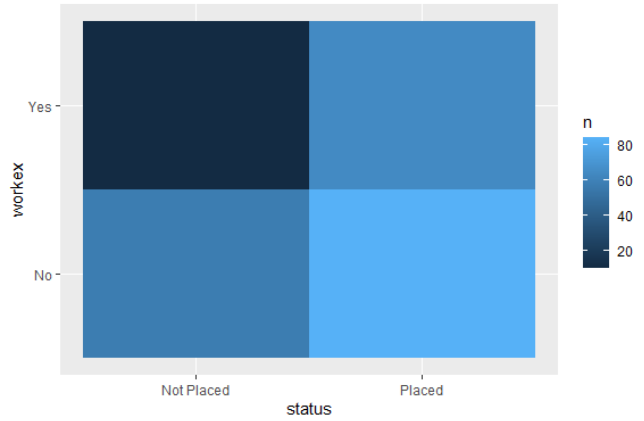

Fig. 14. Heat map between workex and status

To make a 2D heatmap in *ggplot2*, we use *geom_tile()*[12]. It takes three variables: x, y represents the position on X, Y-axis, respectively, and fill indicates the numeric value of the column that will be translated into color.

Before we had data but with the help of advanced technology and tools, now we can easily analyze the campus placement data and find suitable opportunity for students according to their requirements and present market.

For students in their final year who will step forward into the industry with no experience in addition to what surplus factors determines their campus placement motivated me to write this paper.

III. Inferences From the Dataset

The inference from the bar plot is that the majority of the students have no work experience, whereas only a smaller percentage of them have work experience. The majority of students have placements, and a lower number of them do not have placements. The number of male students is about twice as much as the number of female students. In secondary school, most of the students studied under the Central board while a slightly smaller number of them studied under other boards. In high school, most of the students picked Commerce as their stream followed by Science streams, and very few of them chose Arts stream. A vast majority of the students have Commerce and Management degrees while fewer of them have Science and Technology degrees, and an even smaller percentage have other types of degrees. Most of the students pursuing an MBA currently are specializing in Marketing and Finance, whereas a lower smaller portion is specializing in Marketing and HR.

A more significant number of people who don't have work experience have placements while only a small amount of people who have work experience have placements. More male students have placements as compared to female students. Most of the students who took Commerce, Science, and Arts streams in high school have placements, the Commerce stream has a higher number of students who didn't get placements compared to the other two streams. While the majority of students with Commerce and Management, Science and Technology and other degrees have placements; the number of students who don't have placements is highest in Commerce and Management and lowest in other degrees. There is a greater number of students who didn't get placements in the Marketing and HR specialization as compared to the Marketing and Finance specialization; a greater number of students with Marketing and Finance specializations have placements as compared to those with Marketing and HR specialization. A slightly larger number of students from the Central board have placements as compared to students from other boards.

The inference from the histogram is that more than 70% of students scored over 60% in secondary school. About 90% of students scored between 50-80% in high school. Over 90% of students scored between 60-80% on their degrees. Most of the students scored between 60-90% on the employability test. Most of the students pursuing their MBA degree scored between 60-70%. A vast majority of the students have salaries within the range of 200k to 400k.

Most of the students who scored higher percentages (above 60%) in secondary school and high school, as well as on their degrees, have placements. The majority of the

students who scored above 70% on the employability test have placements. The majority of the students who scored over 60% on their MBA degrees have placements.

A majority of the students whose salary is between 200k to 400k are male. Most of the students whose salaries are between 200k to 400k have Commerce and Management degrees whereas a lesser number have Science and Technology degrees and very few have other degrees. Most of the students whose salaries lie between 200k to 400k have no work experience. A greater number of students specializing in Marketing and Finance have salaries in the range of 200k to 400k as compared to students specializing in Marketing and HR.

The inference from the point plot is that most of the students who scored between 60-90% in secondary school have salaries between 200k to 400k. Most of the students who scored between 55-80 % on their degrees have salaries between 200k to 400k. A huge majority of the students who scored over 50% on the employability test have a salary range between 200k to 400k. Most students who scored between 55-75% on their MBA degrees have salaries between 200k to 400k.

Most of the students who scored over 60 % in secondary school and on their degrees; and whose salaries lie between 200k to 400k are male. Most of the students who were assessed on the employability scale are male. A majority of the students who scored between 55-70% on their MBA degrees are male.

Inference from the box plot is that those students who are placed have a maximum of 80 above percentage, and those who are not placed have a minimum of 50 percent in their degree.

The Student with commerce background has a higher degree percentage than compared to Arts stream maximum percentage being 72. Most of the students in Arts, Commerce, and science stream with low degree percent are not placed.

Geom_count helps us conclude that more number of men a with *comm&Mgmt* degree are placed. Similarly, we can prove through *geom_count()* that students with a good degree percentage have a high chance of being placed. Students with any score in their employability test conducted by their college can get placed. These scores have less influence on placement.

Through the heat map, it's easy to find the correlation between fields such as we noticed very few students who pursue Arts in their *hsc* are highly not placed. Those students who did their specialization in *Mkt&Fin* are highly placed.

IV. Predicting the Data Using Random Forest and Decision Tree Algorithm

The decision tree algorithm is one of the most popular algorithms used in data mining and machine learning. It is a type of supervised learning algorithm and works for both continuous and categorical variables[13]. The decision to do strategic splits affects a tress's accuracy. The best split is the one that separates two different labels into two sets[14].

Whereas Random forest is a flexible machine learning method for performing both classification and regression problems[15]. It creates a powerful model by combing a group of weal models. It creates multiple decision trees and then merges the result yielding a much better-generalized model[16].

A. Performing the Random Forest Algorithm

The campus recruitment data were first cleaned by deleting two columns, which are salary due to null values and Sl_no, which is not useful in the model. Then the data is split into the ratio of 70:30. The number of rows in the training set is 151, and the test set has 64 rows. After making the random forest model on the train data set with the target column as status respect to other columns, we calculated the Gini index for all independent variables using importance function. The Gini index is calculated by subtracting the sum of the squared probabilities of each class from one:

```
> importance(mod_forest)
               MeanDecreaseGini
gender                1.0785529
ssc_p                 9.8769470
ssc_b                 0.7865621
hsc_p                 7.2603710
hsc_b                 0.3943178
hsc_s                 1.0247127
degree_p              7.8150669
degree_t              0.8078646
workex                2.3886184
etest_p               2.5436559
specialisation        1.6211979
mba_p                 4.2538585
```

Fig. 15. Mapping independent column & respective GINI Index

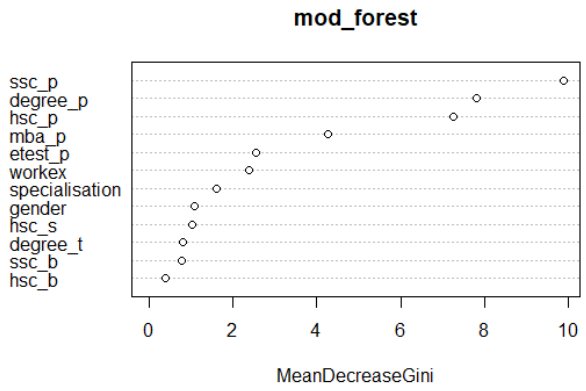

Fig. 16. The graph between the MeanGINI value and the columns

The Gini index helps us to determine which all columns will result in the best split. The column with the higher GINI value will give the best split.

We can predict the values of the test set using the trained model, and below is the peak of the predicted six values:

```
     1      12      13      16      17      18
Placed  Placed Not Placed  Placed  Placed  Placed
Levels: Not Placed Placed
```

Fig. 17. Prediction of the status column using Random forest model

A confusion matrix helps us to describe the performance of our model. It allows us to measure the accuracy of our model[17].

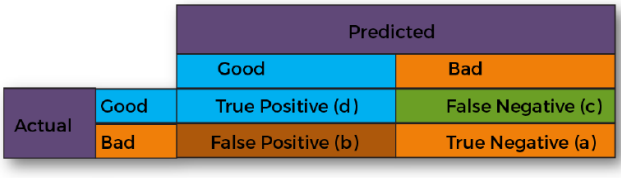

Fig. 18. Confusion matrix

```
                Not Placed Placed
Not Placed              11      9
Placed                   0     44

               Accuracy : 0.8594
                 95% CI : (0.7498, 0.9336)
    No Information Rate : 0.8281
    P-Value [Acc > NIR] : 0.319755

                  Kappa : 0.6269

 Mcnemar's Test P-Value : 0.007661

            Sensitivity : 1.0000
            Specificity : 0.8302
         Pos Pred Value : 0.5500
         Neg Pred Value : 1.0000
             Prevalence : 0.1719
         Detection Rate : 0.1719
   Detection Prevalence : 0.3125
      Balanced Accuracy : 0.9151

       'Positive' Class : Not Placed
```

Fig. 19. Confusion matrix and another measurement of our model

Our random forest model has an accuracy of 0.85, which depicts the usefulness of the model for the inferences.

B. *Using Decision Tree Algorithm*

Similarly, when we Performing Decision Tree Algorithm on the same train and test dataset. Our model determines the best split and is plotted as below:

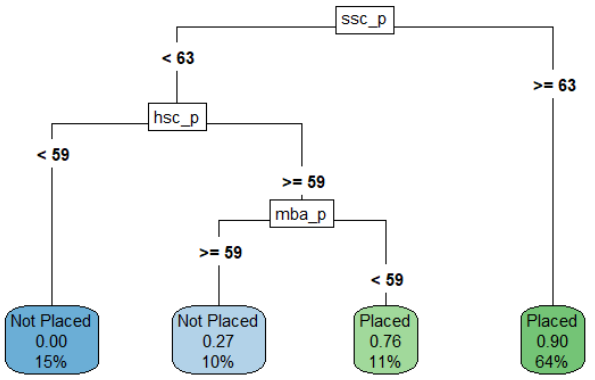

Fig. 20. A decision tree created by the decision tree model

The predicted value with decision tree on test dataset are:

```
     6      8     10     14     16     17
Not Placed Placed Placed Placed Placed Placed
Levels: Not Placed Placed
```

Fig. 21. Prediction of the status column using the Decision tree model

The confusion matrix:

```
                result
                 Not Placed Placed
Not Placed                8     12
Placed                    2     42
```

Fig. 22. Confusion Matrix for Decision tree Model

The accuracy of the decision tree model:

```
               Accuracy : 0.7812
                 95% CI : (0.6603, 0.8749)
    No Information Rate : 0.8438
    P-Value [Acc > NIR] : 0.93397

                  Kappa : 0.4105

 Mcnemar's Test P-Value : 0.01616

            Sensitivity : 0.8000
            Specificity : 0.7778
         Pos Pred Value : 0.4000
         Neg Pred Value : 0.9545
             Prevalence : 0.1562
         Detection Rate : 0.1250
   Detection Prevalence : 0.3125
      Balanced Accuracy : 0.7889

       'Positive' Class : Not Placed
```

Fig. 23. Detailed accuracy and another parameter regarding our model

Our decision tree model has an accuracy of 0.78, which means it is a good model but not compared to the Random Forest model.

V. Conclusion

In this work, the main reason for choosing R over Python is due to the statistical and analytical abilities in R. The *ggplot2* is a powerful R package used in EDA to envisage the graphics grammar, data, or statistics. We can conclude through our study that EDA is a detailed examination that helps in discovering the structure of the data. It is essential for all domain as it reveals trends, patterns, and those relations which are not quickly evident. EDA is the best way to detect outliers but if not performed properly can misdirect us.

We observe that the Random Forest yields an increase in classification performance and results in higher accuracy than the Decision Tree. Therefore, for such classification problems, we recommend using Random Forest. The splitting of the dataset and training the dataset are major factors key to achieve a good model. However, there can be certain limitation using decision tree algorithm & random forest algorithm as overfitting can occur easily. Trees can be unstable as a slight change in data can lead to completely different tree.

VI. References

[1] S. M. Thaung et al., “Exploratory Data Analysis Based on Remote Health Care Monitoring System by Using IoT,” Communications, vol. 8, no. 1, pp. 1–8, 2020.

[2] “Campus Recruitment | Kaggle.” [Online]. Available: https://www.kaggle.com/benroshan/factors-affecting-campus-placement/kernels. [Accessed: 30-Apr-2020].

[3] “RStatTutorial_Basic.” [Online]. Available: http://cis.csuohio.edu/~sschung/CIS660/RStatTutorial_BasicLab1. [Accessed: 18-Jun-2020].

[4] J. Tuimala and A. Kallio, “R, Programming Language,” in Encyclopedia of Systems Biology, New York, NY: Springer New York, 2013, pp. 1809–1811.

[5] Cox, Victoria. "Exploratory data analysis." Translating Statistics to Make Decisions. Apress, Berkeley, CA, pp. 47-74, 2017.

[6] X. Qin, Y. Luo, N. Tang, and G. Li, “Making data visualization more efficient and effective: a survey,” VLDB J., vol. 29, no. 1, pp. 93–117, 2020.

[7] H. Wickham, “ggplot2 by Hadley Wickham,” Media, vol. 35, no. July, p. 211, 2009.

[8] K. Ito and D. Murphy, “Tutorial: Application of ggplot2 to pharmacometric graphics,” CPT Pharmacometrics Syst. Pharmacol., vol. 2, no. 10, p. e79, Oct. 2013.

[9] J. S. Kendrick, D. F. Williamson, P. D, R. A. Parker, and J. S. Kendrick, “The box plot : a simple visual method to interpret data . Ann Intern Med 110 : 916 The Box Plot : A Simple Visual Method to Interpret Data,” Acad. Clin., vol. 10, no. July 1989, pp. 916–921, 1989.

[10] C. Thirumalai, M. Vignesh, and R. Balaji, “Data analysis using box and whisker plot for lung cancer,” 2017 Innov. Power Adv. Comput. Technol. i-PACT 2017, vol. 2017-January, no. March, pp. 1–6, 2017.

[11] “Count overlapping points — geom_count • ggplot2.” [Online]. Available: https://ggplot2.tidyverse.org/reference/geom_count.html. [Accessed: 30-Apr-2020].

[12] “geom_tile | ggplot2 | Plotly.” [Online]. Available: https://plotly.com/ggplot2/geom_tile/. [Accessed: 30-Apr-2020].

[13] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” Int. J. Comput. Sci. Issues, vol. 9, no. 5, pp. 272–278, 2012.

[14] D. Berrar and W. Dubitzky, “Decision Tree,” in Encyclopedia of Systems Biology, Springer New York, 2013, pp. 551–555.

[15] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[16] C. Vens, “Random Forest,” in Encyclopedia of Systems Biology, Springer New York, 2013, pp. 1812–1813.

[17] “Confusion Matrix-based Feature Selection.” [Online]. Available: https://www.researchgate.net/publication/220833270_Confusion_Matrix-based_Feature_Selection. [Accessed: 30-Apr-2020].

[18] Senthil, Velan S. "Quantitative Assessment of Inheritance Hierarchies for Aspect Oriented Software Development using a proposed Aspect Inheritance Reusability Model." In 2019 International Conference on Automation, Computational and Technology Management (ICACTM), pp. 573-576. IEEE, 2019.