Prediction of placement of candidates using KNN, Logistic Regression and Random Forest model

Antarjita Mandal, Sai Shruthi S, Yogadisha S

***Abstract*—Job Placements provide an opportunity for candidates to find suitable work according to their educational qualification and past experience. The competition among the candidates for a job placement is extremely high . Thus it becomes necessary to analyse the various factors that influence the job placement of candidates. The motive behind this paper is to analyse the placement dataset of an educational institution and understand the academic and employability factors that affect the job placements. Further on, a methodology is proposed to predict the job placement of various candidates using KNN, Logistic Regression and Random Forest models.The accuracy of the models is found out and compared to determine the best fit for the problem statement in hand. This proposed solution contributes towards identifying whether a particular candidate will secure a job placement or not.**

Keywords—placements, factors, Campus Placements, regression models, random forest, logistic regression, KNN, exploratory data analysis.

I. Introduction

Most of the candidates preparing for job placements need to understand the various areas they need to work on in order to get the job they desire.Thus analysing the factors that influence placements is crucial for the candidates, For the educational institutions as well, it gives an idea on the areas they need to train their students to help them get a good offer.This paper offers a detailed analysis of the placement dataset of 250 odd MBA students of a college in Andhra Pradesh. The factors influencing the job placements are highlighted. It also proposes a solution to predict the chances of a candidate getting placed through 3 models. The accuracy of the models are then compared to determine the most appropriate one for the specified scenario.

II. Related Works

The key problem in hand is to predict the campus placement of students. Various models have been developed to solve this problem.

We have referred to three papers that describe the techniques that can be used to approach this problem.

A. Literature Paper 1

P Manvitha and Neelam Swaroopa in their paper **Campus Placement Prediction using Unsupervised Machine Learning Techniques** [*2*] talk about how supervised Machine Learning techniques like Random Forest and ID3 algorithms can be used to predict campus placement of students.A systematic approach that include data gathering, pre-processing ,processing and interpretation is followed. During data pre-processing , the missing values are cleaned and only the relevant attributes are retained.Feature scaling is done to standardize the range of independent variables. Then Random Forest and ID3 algorithms are applied on the given dataset, the accuracy of the Random Forest algorithm was found to be better than ID3 . Thus a conclusion was drawn that Random Forest Algorithm is a better technique to predict the campus placement of students.No assumptions were made and no limitations were reported in this paper.

B. Literature Paper 2

D. Satish Kumar, Zailan Bin Siri, D.S. Rao, S. Anusha in their paper **Predicting Student's Campus Placement Probability using Binary Logistic Regression** [3] refers to how logistic regression can be used predicting the placement status.Assumptions are made taking only 6 hypothetical statements that affect the student's campus placements.The predictors considered are: CGPA in UG and PG, Specialisation in UG and PG, Soft Skill Score and Gender. Also the collection of data is from the state of Andhra Pradesh of about 250 odd MBA students.

Choosing the right sample hypothesis space which includes the factors affecting the placements needs to be carefully chosen. According to this sample hypothesis space we can analyse our data and build our model. Model diagnostic measures that need to be carried out with at least 70% accuracy which is considered optimal. Plotting ROC curve and Cox & Snell's R2 / Nagelkerke's R2 provides the pseudo R2 measures for the model. After doing the model diagnostic we can draw the conclusion about the and check for the significance of the model.

Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew in their paper **Student Placement**

C. Literature Paper 3

Prediction Using Machine Learning [4] focuses on machine learning techniques to predict placement status of the student provided through text input.The placement prediction is done by machine learning using Naïve Bayes and K-nearest neighbor (KNN) algorithm. The algorithm considers the parameters such as USN, Tenth and PUC/Diploma results, CGPA, Technical and Aptitude Skills

. Here they use two different machine learning classification algorithms, namely Naive Bayes Classifier and KNearest Neighbors [KNN] algorithm. These algorithms independently predict the results and we then compare the efficiency of the algorithms, which is based on the dataset. This model helps the position cell at intervals a corporation to spot the potential students and concentrate to and improve their technical and social skills. Many of the previous research papers concentrate on a less number of parameters such as CGPA and Arrears for placement status prediction which leads to less accurate results, but proposed

work contains many educational parameters to predict placement status which will be more accurate

III. Problem Statement

The problem statement is analysis of placement dataset to gain insights about the academic and employability factors that affect placements and collection of data is from the state of Andhra Pradesh of about 250 odd MBA students.

Choosing the right sample hypothesis space which includes the factors affecting the placements needs to be carefully chosen. According to this sample hypothesis space we can analyse our data and build our model. Model diagnostic measures that need to be carried out with at least 70% accuracy which is considered optimal. Plotting ROC curve and Cox & Snell's R2 / Nagelkerke's R2 provides the pseudo R2 measures for the model. After doing the model diagnostic we can draw the conclusion about the same and check for the significance of the model.

Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew in their paper **Student Placement** building various prediction models for the same . The performance of each model is found out to determine the one that best fits the dataset.

IV. Solution Design

Any dataset cannot be processed without performing data cleaning and pre- processing like Normalization,choosing the training and testing dataset etc. Thus they are the initial steps to be followed while approaching our problem. Performing various Exploratory Data Analysis to find out the correlation between attributes and identify the attributes that have an influence on the placement of a given candidate becomes the next crucial step that helps in gaining initial insights about the data set. This is going to be followed by the model building that involves building various models like Random Forest, KNN ,Logistic Regression separately predicting the same on the training data. The accuracy of the individual models is noted.The model with better accuracy, precision and ^{2} value will be considered suitable for this problem statement.

A. *Pre-processing*

Data preprocessing plays a major role in the process of model building .If the data is not pre-processed properly, it can lead to drastic errors in the final output of the model.

As a part of data preprocessing , we have performed the following :

1. Data Cleaning : Under data cleaning , the missing values in the dataset were handled . It was found that only the salary attribute had missing values which were replaced by 0. Apart from this, all the other columns were clean.
2. Normalization : Data normalization basically refers to scaling the values of numeric columns in the dataset to a common scale that is between 0 and 1. Our dataset needed data normalization as the range of the salary attribute and marks was different . Thus it was necessary to bring the numeric values to a common range.
3. Dropping columns : Dropping the unnecessary columns is extremely important and can drastically affect the performance of the model built.

For example : Our dataset had a salary attribute. This field is not applicable for candidates who have not been placed. As a part of the data cleaning step, the missing values in salary were replaced with 0 . Incase the salary attribute was not dropped, during the training stage , the model would assume that 0 salary implies that the candidate was not placed and thus this could lead to very high accuracy .But in reality this does not make sense as we are trying to predict the probability of a candidate being placed and here the salary does not come into picture. Thus dropping this column becomes necessary .

4. Removing the outliers in the dataset: Most of the columns did not have any outliers except for column hsc_p. Plotting box plots helped in identifying the regarding.

B. *Exploratory Data Analysis*

Exploratory Data Analysis was performed to gain certain initial insights from the data. From the graph plotted , it can be observed that males have a higher chance of getting placed though it is only one of the factors.

In the early stages , it looks like people with work experience have a higher chance of being placed against freshers. In this particular dataset, Comm Mgmt has a higher frequency of getting placed , though this could be due to the data collection method but as far as the initial Exploratory Data Analysis is concerned , this is the trend observed. Visualising the distribution of the salary attribute, the skewness and kurtosis value is found out . The kurtosis value is greater than 3. Thus it is leptokurtic. A scatter plot is plotted between salary and percentage , and the correlation coefficient value is found to be 0.4083 indicating a positive relation. A box plot is plotted to understand the variation in salary for male and female. From the plot, it is evident there are hardly any outliers .The mode of the specialization attribute is found out and it can be seen that Marketing and Finance is the most common specialization. There is a positive correlation between salary and secondary school percentage. The scatterplot between MBA percentage and salary show no patterns. 10.There are no outliers in other degree types but there are outliers present in comm & mgmt and in sci&tech.

C. Model Building

The three models used to predict the placements of students. Predicting variable (aka. status) is a binary variable which tells whether the student gets placed (ie..
1) or not (ie.. 0). We divide the complete dataset into training and testing dataset. The split factor we used is 0.67*(number of rows) for the training dataset and the rest of the testing dataset.
The three models are:-

1. **Random Forest :** After the general pre-processing step, as a part of the model building , the categorical variables , in specific the target variable, were converted to factor data type. The random forest was then constructed with ntree=301 and mtry=4. This implies that the number of decision trees constructed was 301 and the number of features selected at each stage was 4. Now the importance of the features was found out , it was observed that hsc_s did not have any significant importance and hence was dropped. Dropping this column proved to be useful as the accuracy increased there after. The model thus built was used to predict the target variable for the test dataset.
2. **Logistic regression** : After the general pre-processing of data Logistic regression model requires creating the dummy variables for the categorical columns like splitting the hsc_c column into Commerce, Arts, CommMgmt, Science, Science & Art. We then test the variables of their overall significance on the model using ANOVA model. We see that gender, ssc_p, hsc_p, degree_p, workex ,mba_p are statistically significant at 0.05 (ie.. 95% confidence interval). Using only these variable models to create the model we then predict the values of the testing variable(ie.. status) on the testing set.
3. **KNN**: As there are some missing values at the salary column, corresponding to the students that weren't placed.Since the goal is predicting the student's placement the columns *sl_no* and *salary* provide no significant importance hence it was dropped.First we transform all our categorical variables (apart from our target variable: **status**) into numerical by creating dummy boolean columns.Afterwards normalize the quantitative variables to express them in the same range of values.

TESTING

Random Forest : Initially, ntree value was set to 1
. This implies that only one decision tree was constructed. It was found that using only one decision tree, the accuracy for the test data stood at 73.91%.. Now when a random forest was built using 301 decision trees(that is increasing the value of ntree), the accuracy for the test data stood at 88.41%. Thus varying the value of the number of decision trees in the random forest had an impact on the final accuracy of the model.
Logistic Regression : Trying for various thresholds ie.. threshold=0.5 has an accuracy of
85.29. Further increasing the threshold to 0.6 we see no significant increase in the threshold. Also when the threshold is decreased to 0.4 we get an accuracy of 86.76% .We also notice this cutoff threshold when plotting the ROC curve.
KNN : Accuracy of Knn Test Prediction for k = 1 . The model is doing a decent job predicting students to be placed (which is of main focus) but doing a bad job on identifying students that weren't placed. This model has accuracy of ~65%. Accuracy of Knn Test Prediction for k = 2:The model accuracy is the same(~65%), but it is doing a worse job since it decreases the number predicted "Placed" students, meaning that is increasing the "False Positives". Our goal is to correctly predict the students that will be placed and maxime the precision.When we graphically see which values of k give us the best classification, we plot "Accuracy vs k number of Neighbors" which will help us understand our testing procedure.From the above mentioned visualization we understand that the for k=6, we will get the highest accuracy of ~80% for the given dataset when tested with our test data.

VI. RESULTS

1. Random Forest : The final accuracy obtained for the test data set using the Random Forest model was **88.41%.**
Accuracy : 0.8841
95% CI : (0.7843, 0.9486)
No Information Rate : 0.7246
P-Value [Acc > NIR] : 0.001191
Kappa : 0.6779
Mcnemar's Test P-Value : 0.077100 Sensitivity : 0.6316
Specificity : 0.9800
Pos Pred Value : 0.9231 Neg Pred Value : 0.8750
Prevalence : 0.2754 Detection Rate : 0.1739
Detection Prevalence : 0.1884 Balanced Accuracy : 0.8058

Confusion Matrix

	0	1
0	12	1
1	7	49

Table 1.

2. Logistic Regression : Plotting the ROC curve we see that the threshold of 0.4 becomes the most appropriate. The chances of overfitting the data is also less.Other Test include:-
McFadden's R squared test:-0.6045018 (df=7) Null deviance: 168.617 on 138 degrees of freedom.
Residual deviance: 66.688 on 132 degrees of freedom.
AIC: 80.688
Accuracy: 86.76%
95% CI : (0.7636, 0.9377)
P-Value [Acc > NIR] : 0.0006336
Kappa : 0.6681
Mcnemar's Test P-Value : 0.1824224 Sensitivity : 0.6667
Specificity : 0.9574
Pos Pred Value : 0.8750 Neg Pred Value : 0.8654 Prevalence : 0.3088 Detection Rate : 0.2059
Detection Prevalence : 0.2353 Balanced Accuracy : 0.8121

Confusion Matrix:-

	0	1
0	14	2
1	7	45

Table 2.

3. KNN:

To reach the maximum accuracy with this model the k value is 6. The model is doing a great job with a high accuracy and only not predicting very few "Placed" student.We could achieve higher precision of 100% with k = 14 but we could be suffering from overfitting our data. Thus the final accuracy obtained is **~80%**

knn=6	0	1
0	10	11
1	1	41

Table 3.

VII. Concluding Remarks

This paper talked about a detailed analysis of the placement dataset inorder to understand the factors affecting the same.The factors that were found to have significant impact in the model building were : gender,ssc_p, ssc_b,hsc_p,hsc_b,degree_p,degree_t,work ex, etest_p,specialization,mba_p.

This paper also proposed a methodology to predict the placements of candidates. KNN, Random Forest and Logistic Regression models were developed as a part of the solution. The Random Forest was found to have an accuracy of 88.41% when compared to the other two models. Thus it can be concluded that the Random Forest model is the best fit for the current scenario.

References

[1] Barrow, S., & Mosley, R. (2011). *The Employer Brand: Bringing the Best of Brand Management to People at Work*. John Wiley & Sons.

[2] Predicting Student's Campus Placement Probability using Binary Logistic Regression. (2019). In *International Journal of Innovative Technology and Exploring Engineering* (Vol. 8, Issue 9, pp. 2633–2635). https://doi.org/10.35940/ijitee.i8984.078919

[3] International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 9 (2019) pp. 2188-2191 © Research India Publications. http://www.ripublication.com 2188 Campus Placement Prediction Using Supervised Machine Learning

[4] International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072 © 2019, IRJET | Impact Factor value: 7.211 | ISO 9001:2008 Certified Journal | Page 4577 STUDENT PLACEMENT PREDICTION USING MACHINE LEARNING

APPENDIX VISUALIZATIONS :

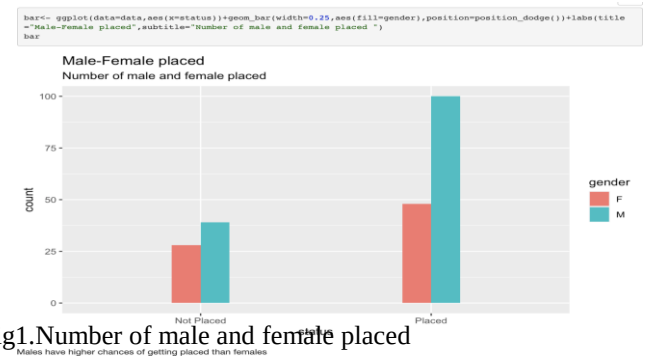

Fig1.Number of male and female placed

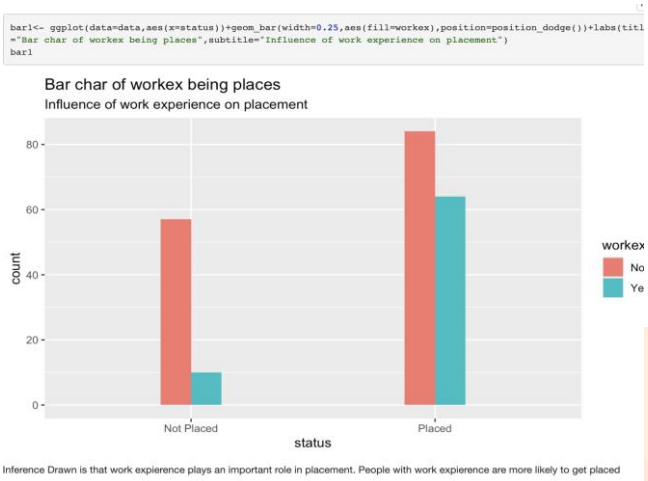

Fig2.Influence of work experience on placement

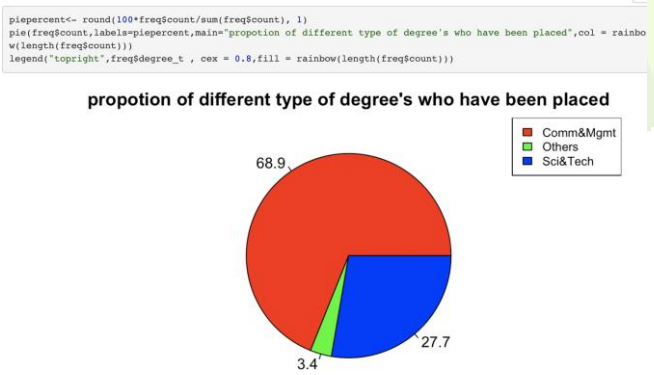

Fig3. Proportion of different type of degrees who have been placed

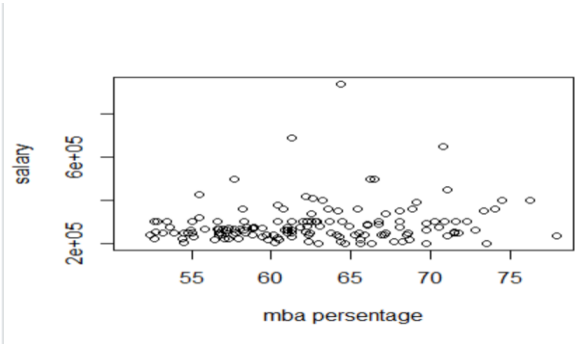

Fig4 .Scatter plot for salary and MBA percentage

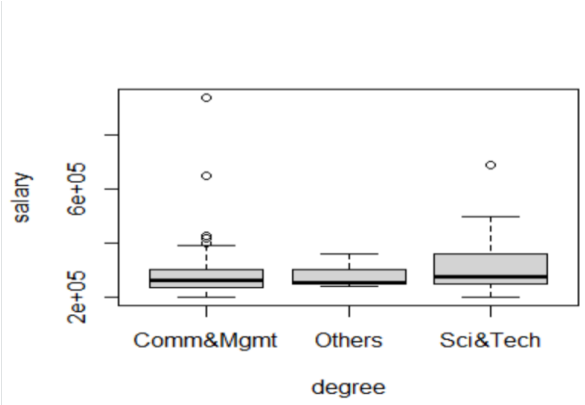

Fig5. Box Plot for Salary and Degree Type

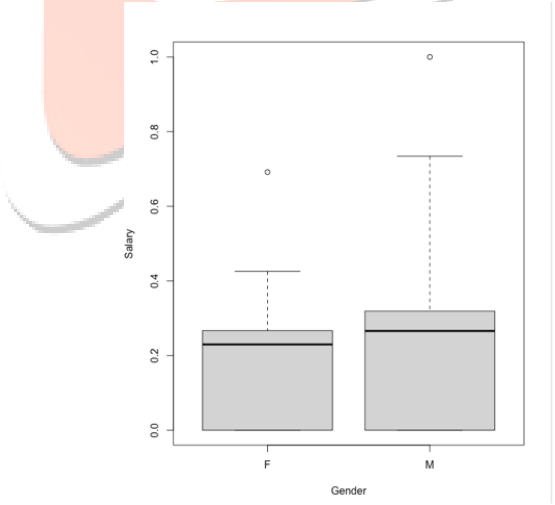

Fig6. Variation in the salary for male and female

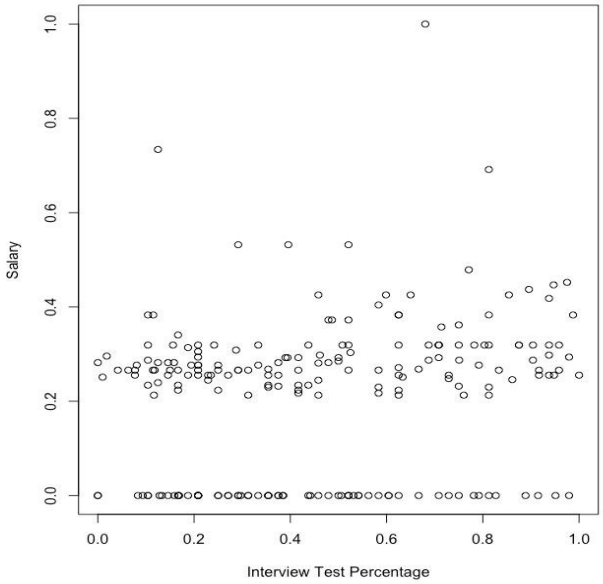

Fig7. Scatterplot between Interview test percentage versus the salary offered for each student