Campus Placement Data Analysis Using Classification Techniques

E. Hari Haran[1], S.Ramesh[2], A.S. Omkumar[3], S.R.Arun[4], V. Saravanan[5], M.S. Sassirekha[6], Anbarasan Balakrishnan[7]

[1,2,3,4,5]*Department of Computer Application Thiagarajar College of Engineering (Affiliated to Anna University) Madurai, India*

[6]*Assistant Professor, Department of Computer Application Thiagarajar College of Engineering (Affiliated to Anna University) Madurai, India*

[7]*Director - Engineering, Capgemini Engineering*

Abstract **- The dataset revolves around the placement season of a Business School in India. Where it has various factors on candidates getting hired such as work experience, exam percentage etc. Finally it contains the status of recruitment and remuneration details. A Set of Classification analysis is performed to predict whether the student is placed or not.**

Index Terms **– Classification, Numpy, Pandas, Sklearn, Seaborn, Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours, Support Vector Machine, XGBoost, AdaBoost, Outliers, Label Encoding and One Hot Encoding**

I.INTRODUCTION

Campus recruitment is a strategy for sourcing, engaging and hiring young talent for internship and entry-level positions. It is typically a tactic for medium- to large-sized companies with high-volume recruiting needs but can range from small efforts (like working with university career centers to source potential candidates) to large-scale operations (like visiting a wide array of colleges and attending recruiting events throughout the spring and fall semester). Campus recruitment often involves working with university career services centers and attending career fairs to meet in person with college students and recent graduates. Primary goals of this analysis are:

- Do a exploratory analysis of the Recruitment dataset
- Do an visualization analysis of the Recruitment dataset
- To find whether a student got placed or not using classification models
- Check the accuracy of each classification model

II. METHODOLOGY

A. Examining the dataset:

There are 14 columns and 215 rows in the dataset and following are the inferences taken from the dataset.

- We have Gender and Educational qualification data
- We have 6 float and 8 object data types in our dataset
- We have all the educational performance(score) data
- We have the status of placement and salary details
- We can expect null values in salary as candidates who weren't placed would have no salary
- Status of placement is our target variable, the rest of them are independent variables except salary.

sl.no	Column	Non-Null-Count	Dtype
0	sl_no	215 non-null	int64
1	gender	215 non-null	object
2	ssc_p	215 non-null	float64
3	ssc_b	215 non-null	object
4	hsc_p	215 non-null	float64
5	hsc_b	215 non-null	object
6	hsc_s	215 non-null	object
7	degree_p	215 non-null	float64
8	degree_t	215 non-null	object
9	workex	215 non-null	object
10	etest_p	215 non-null	float64
11	specialisation	215 non-null	object
12	mba_p	215 non-null	float64
13	status	215 non-null	object
14	salary	148 non-null	float64

Table 1: Campus Placement Dataset Info

B. Data Cleaning

The noisy data, outliers and the null values in the dataset can produce the biased output while doing prediction. So we have to clean them by removing them or replacing them. Sometimes this could also provide valuable information while collecting the data but for good results most of the times we have to drop them. We used the Seaborn Library and Heatmap technique to visualize the null value and Matplotlib to visualize the Outliers and null values. The inference from the visualization:

- The field Salary has more null values.
- The field HSC_P (Higher secondary percentage) has more outliers The null values are replaced by 0. Now remove the records with outlier data

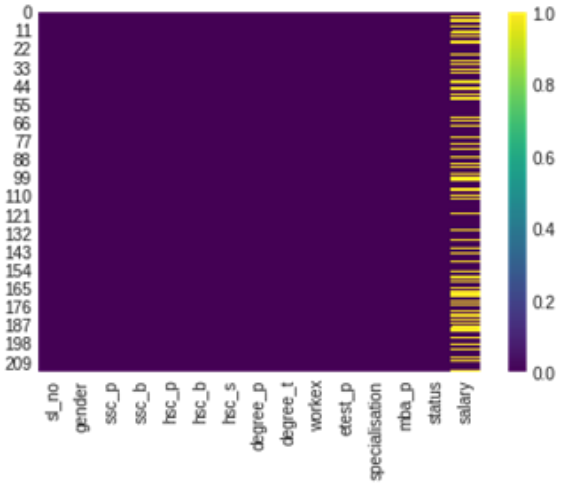

fig 1. Null value visualization

C. Visualization

The visualization is primarily to understand the data. We used the Matplotlib library to plot the data to understand what it is conveying. The inference from the visualization:

- We have more male candidates than female
- We have candidates who did commerce as their hsc course and as well as undergraduate
- Science background candidates are the second highest in both the cases
- Candidates from Marketing and Finance dual specialization are high
- Most of our candidates from our dataset don't have any work experience
- Most of our candidates from our dataset got placed in a company

D. Data Preprocessing:

We have to encode the categorical variables like the following into 1s and 0s:

- Gender
- Workex
- Specialisation
- Status

Because the machine learning models require the input to be numeric. Also we perform one hot encoding for the degree_t field. Because we have more than two classes there.

E. Training and Testing Split:

Before splitting the data for training and testing, we have to assign the response variable and predictor variable to Y and X respectively. Now we have to split the data in an 80:20 ratio. 80% of the data will be used for training the models and 20% of the data will be used for testing.

F. Performing Classification:

Prepare the model using the X_train and y_train (training data) using the following algorithms:

- Logistic regression
- Decision Tree Classifier
- Random Forest Classification
- Support Vector Machine
- K- Nearest Neighbour
- XGBoost

With the prepared model ,test that with the 20% (X_test) testing data and assign that to the y_pred variable Now test the performance of the model using Confusion matrix to get (Accuracy, Precision, f1-score and Recall)and ROC curve

III. INFERENCES

From the analysis report on Campus Recruitment dataset here are the conclusions

- Educational percentages are highly influential for a candidate to get placed
- Past work experience doesn't influence much on your masters final placements
- There are no gender discrimination while hiring, but higher packages were given to male
- Academic percentages have no relation towards salary packages

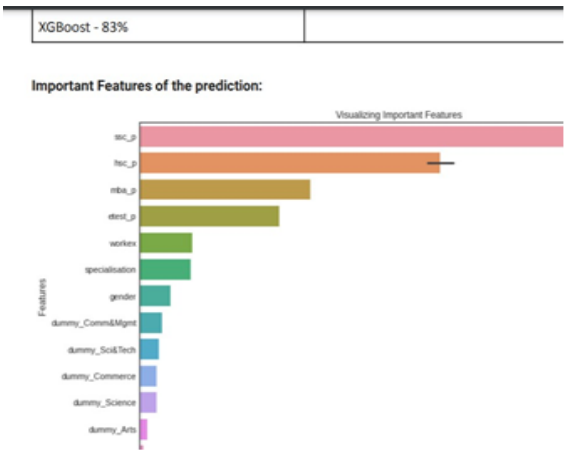

fig 2. Important features influencing the results

IV. ALGORITHMS ACCURACY RESULT

- Logistic regression - 81%
- Decision Tree Classifier - 73%
- Random Forest Classification - 83%
- Support Vector Machine - 83%
- K- Nearest Neighbour - 76%
- XGBoost - 83%
- ADABoost - 81%

V. CONCLUSION

- We have found the important features which could play a vital role in campus placement of a student and non influential features as well. We studied the report of all the algorithms from the author carefully and we performed the AdaBoost algorithm.
- We also generated a comparison table based on the comparison of analytics we performed with the author's analytics.

REFERENCES

[1] https://www.kaggle.com/benroshan/you-re-hired-analysis- on-campus-recruitm ent-data
[2] https://machinelearningmastery.com/types-of-classificatio n-in-machine-learnin g/
[3] https://seaborn.pydata.org/
[4] https://matplotlib.org/
[5] https://pandas.pydata.org/
[6] https://scikit-learn.org/
[7] https://numpy.org/