CAMPUS RECRUITMENT PROJECT

By Darshan D S

Supervisor, Dr. Vinod Kumar Murti

1. Introduction

1.1. Abstract

There are so many factors affect the placements of a student for a job. In this project I am trying to understand some of those factors which recruiters consider important. I have also created a machine learning algorithm which classifies the students who have a high chance of getting placed from those who won't.

1.2. Data

The data set consists of Placement details of students in XYZ campus. It includes secondary and higher secondary percentage and specialization. It also includes degree percentage, degree specialization, MBA percentage, MBA specialization, work experience if any and the salary offered to those who got placed. The target variable is status, which contains two classes – 'Placed' and 'Not Placed'.

2. Software

The programming language used for this project is Python. Python is a general-purpose programming language with lots of libraries which are very useful for data analysis and machine learning projects. The python version used for this project is Python 3.8.3.

The major libraries used for this project are:

2. NumPy
3. Pandas
4. Matplotlib
5. Seaborn
6. SciPy
7. Scikit-learn

The Integrated Development Environment (IDE) used for this project is Jupyter notebook.

3. Data Description

The Data set contains 13 variables with 7 categorical (all of nominal level) and 6 continuous. The data size is small with only 215 entries. The description of each feature is:

1. gender: The gender of the student.
    a. ‘M’ Male
    b. ‘F’ Female
2. ssc_p: Secondary Education percentage – 10^{th} grade.
3. ssc_b: Board of Education.
    a. ‘Central’ Central board
    b. ‘Others’ Other boards
4. hsc_p: Higher Secondary Education percentage – 12^{th} grade.
5. hsc_b: Board of Education.
    a. ‘Central’ Central board
    b. ‘Others’ Other boards
6. hsc_s: Specialization in Higher Secondary Education.
    a. ‘Commerce’ Commerce branch
    b. ‘Science’ Science branch
    c. ‘Arts’ Arts branch
7. degree_p: Degree Percentage.
8. degree_t: Under Graduation (Degree type) – Field of degree education.
    a. ‘Comm&Mgmt’ Commerce and Management
    b. ‘Sci&Tech’ Science and Technology
    c. ‘Other’ Other branches
9. workex: Work Experience.
10. etest_p: Employability test percentage (conducted by college)
11. specialisation: Post Graduation (MBA) – Specialization.
    a. ‘Mkt&Fin’ Marketing and Finance
    b. ‘Mkt&HR’ Marketing and Human Resource
12. mba_p: MBA percentage.

13. status: Status of placement.
    a. ‘Placed’ Student got placement
    b. ‘Not placed’ Student doesn’t got placement
14. salary: Salary offered by corporate to candidates.

4. Understanding the data

4.1. Missing Values

In this dataset, there is only one column which contains missing values, which is the ‘salary’ column. There are 67 missing values in salary column. If a student does not get placed, then he/she won’t have a salary package, hence all 67 students without any values for salary are those who are not placed.

4.2. Response Variable

The response variable or dependent variable of this data is ‘status’. It is the status of the placement of a student, i.e., if the student got the placement or not. If the student is placed, then it is denoted as ‘Placed’ and if the student didn’t get the placements, then it is denoted by ‘Not Placed’.

148 students out of 215 got placements and 67 were not placed.

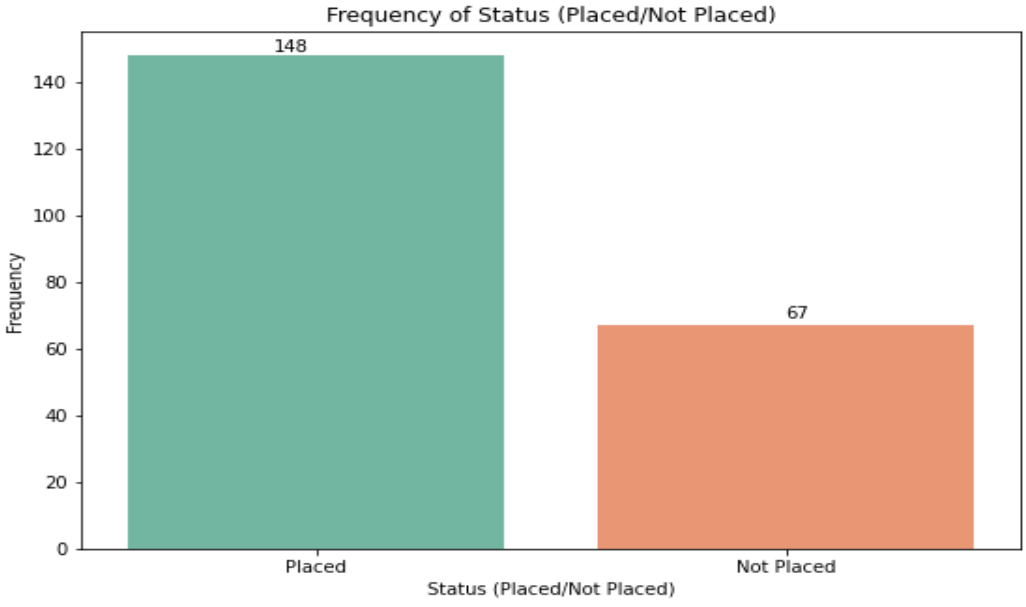

Figure 1: Count plot of Status

4.3. Findings from the data

- Does the 10th and 12th marks determine the chances of a student getting placed?

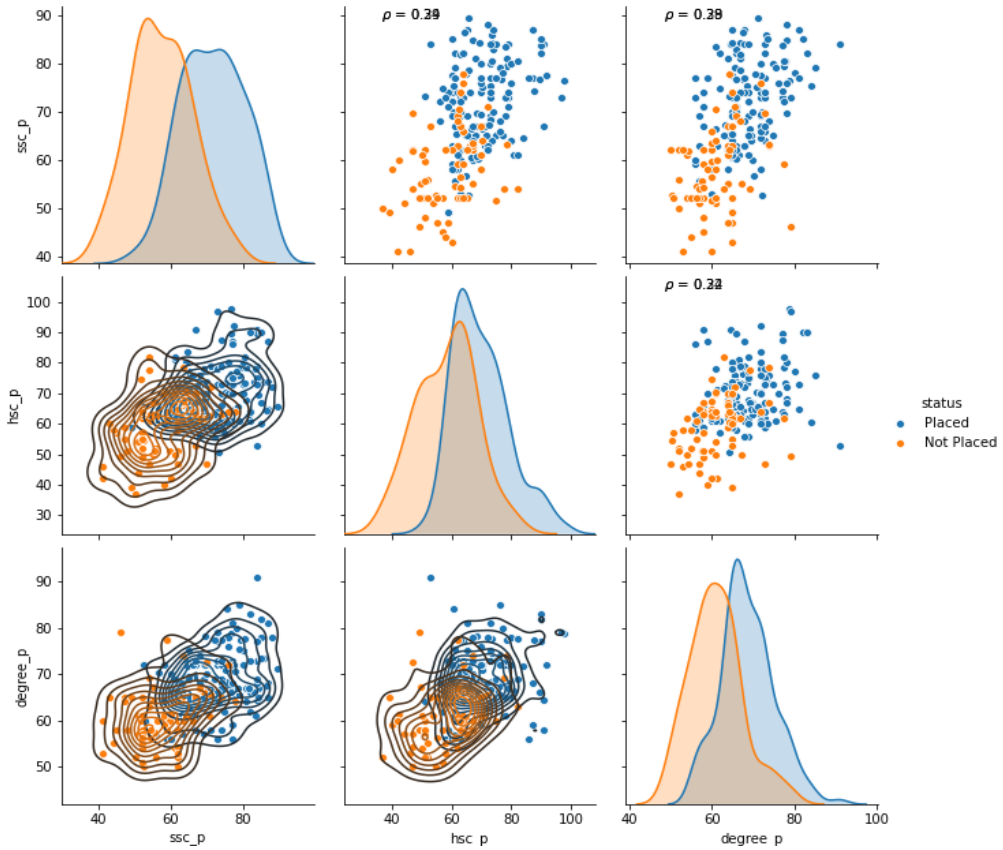

Figure 2: Pair plot of marks and status

Figure 2 shows the pair plot of 10th, 12th and degree marks which are categorized by status. The orange dots represent Not Placed and the blue dot represent placement. If the 10th, 12th and degree marks are low, then we can see more orange dots, that means these students were not placed, and those students who performed well in these exams were successful in securing the placements. So, the marks scored by a student in 10th, 12th and degree is a significant factor in securing the placements. But we can also see outliers in both the categories.

❖ Does having a work experience improve the odds of getting placed?

Status	Not Placed	Placed	Total
Work experience			
No	40.42	59.57	100
Yes	13.51	86.48	100

Table 1: Work experience vs Status

If the student doesn't have a work experience, then he/she has only 59.5% chance of getting placed, but if he/she have work experience then the odds of getting placed increases to 86.4%.

❖ Does the marks from Employability test conducted by the college impacted the odds of getting placed?

Figure 3: Box plot of employment score

From the boxplot we can see that, students who got placements have a median mark of 72%, but those who were not placed got a

median score of 67%. So, the employment score has an impact on placements.

- ❖ Is there a particular branch in MBA which offers good placement results?

Status	Not Placed	Placed	Total
Specialization in MBA			
Market & Finance	20.83	79.17	100
Market & HR	44.21	55.79	100

Table 2: Specialization vs Status

If the specialization is Marketing and Finance, then the student has a 79.17% chance of getting placed, whereas there is only 55.79% chance for students who specialized in Marketing and HR.

4.4. P-value

For all the variables Hypothesis testing is done and p-value is noted. If the p-value is below alpha (which is taken as 0.05) then we reject the null hypothesis and the variable is a good predictor of status of placements. If the p-value is above alpha then we fail to reject null hypothesis and the variable is not a good predictor of status of placements.

Feature name	P-value	Accept/Reject
gender	0.239802609	Reject
ssc_p	4.12E-23	Accept
ssc_b	0.689772943	Reject
hsc_p	1.85E-14	Accept
hsc_b	0.922283705	Reject
hsc_s	0.572711851	Reject
degree_p	8.81E-14	Accept
degree_t	0.226610762	Reject
workex	9.91E-05	Accept
etest_p	0.061720381	Reject
specialisation	0.000420184	Accept
mba_p	0.261445002	Reject

Table 3: P-value

Based on p-value, I have concluded that some features have no impact on the status of placement of students, they are:

1. Gender
2. 10^{th} board of education
3. 12^{th} board of education
4. 12^{th} specialization
5. Degree trade
6. Employability test percentage
7. MBA test percentage
8. Salary

4.5. Conclusion

After the exploratory data analysis, all the unnecessary variables are removed from the dataset and only the necessary variables are used for model building. From 13 independent variables I selected only 5 variables for my final model.

The selected variables are:

1. 10^{th} percentage
2. 12^{th} percentage
3. Degree percentage
4. Work experience
5. MBA specialization

5. Model Analysis

5.1. Logistic Regression Model

The whole dataset was split into training set and testing set, and both of them were scaled using Standard Scaler function provided by Scikit-learn library and encoded using One Hot Encoder function provided by Scikit-learn library.

A Logistic regression model was trained on the default parameters and got the following results.

Evaluation metric	Value
Accuracy	75.38 %
Accuracy of 0's	72.72 %
Accuracy of 1's	75.92 %
Precision	75.92 %
Recall	93.18 %
ROC area	89.28 %

Table 4: Logistic regression evaluation metrics

5.2. Random Forest Classifier

A Random Forest Classifier was trained using the training set, with number of estimators as 300, criterion as Gini and max depth as None.

The following results were observed:

Evaluation metric	Value
Accuracy	80.00 %
Accuracy of 0's	72.72 %
Accuracy of 1's	82.97 %
Precision	82.97 %
Recall	88.63 %
ROC area	90.09 %

Table 5: Random Forest Classifier evaluation metrics

5.3. K – Nearest Neighbors

A k-nearest neighbor algorithm with number of neighbors as 5, weights as uniform, algorithm as auto, leaf size as 30, metrics as minkowski, metric parameter as None and number of jobs as None was trained to get the following results:

Evaluation metric	Value
Accuracy	75.38 %
Accuracy of 0's	72.72 %
Accuracy of 1's	75.92 %
Precision	75.92 %
Recall	93.18 %
ROC area	79.87 %

Table 6: K – Nearest Neighbor model evaluation metrics

5.4. Support Vector Machines

A Support Vector Machine classifier with C=1.0, kernel='rbf', degree=3, gamma='scale', shrinking=True, probability=True, decision function shape='ovr' was trained to get the following results:

Evaluation metric	Value
Accuracy	75.38 %
Accuracy of 0's	72.72 %
Accuracy of 1's	75.92 %
Precision	75.92 %
Recall	93.18 %
ROC area	83.00 %

Table 7: Support Vector Machines evaluation metrics

5.5. Model comparison

Model	Accuracy	Accuracy 0	Accuracy 1	Precision	Recall	ROC area
Logistic Regression	0.753846	0.727273	0.759259	0.759259	0.931818	0.892857
Random Forest	0.8	0.722222	0.829787	0.829787	0.886364	0.900974
KNN	0.753846	0.727273	0.759259	0.759259	0.931818	0.798701
SVM	0.753846	0.727273	0.759259	0.759259	0.931818	0.830087

6. Conclusion

Based on the model comparison overall accuracy is higher for Random Forest Classifier which is 80% accuracy. All the other model having the overall accuracy at 75.38%. But the accuracy of classifying the negative class (accuracy of 0's) are almost equal for all the models which is 72.72% accuracy.

Overall, we can select Random Forest Classifier to classify this data, as it performs well on most evaluation metrics.

8. Reference

1. https://www.kaggle.com/benroshan/factors-affecting-campus-placement
2. https://www.geeksforgeeks.org/countplot-using-seaborn-in-python/#:~:text=countplot()%20method%20is%20used,each%20categorical%20bin%20using%20bars.&text=Parameters%20%3A%20This%20method%20is%20accepting,for%20plotting%20long%2Dform%20data.
3. https://towardsdatascience.com/visualizing-data-with-pair-plots-in-python-f228cf529166
4. https://www.datacamp.com/community/tutorials/pandas-to-csv?utm_source=adwords_ppc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=332602034358&utm_targetid=aud-299261629574:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=1007779&gclid=CjwKCAjwkN6EBhBNEiwADVfyazLgCCdlDA8GBMOyS5aOcdDjPVN28ULMXZIjXjmmq9HKyrXk_YL0KxoCrg4QAvD_BwE
5. https://www.python-graph-gallery.com/35-control-order-of-boxplot