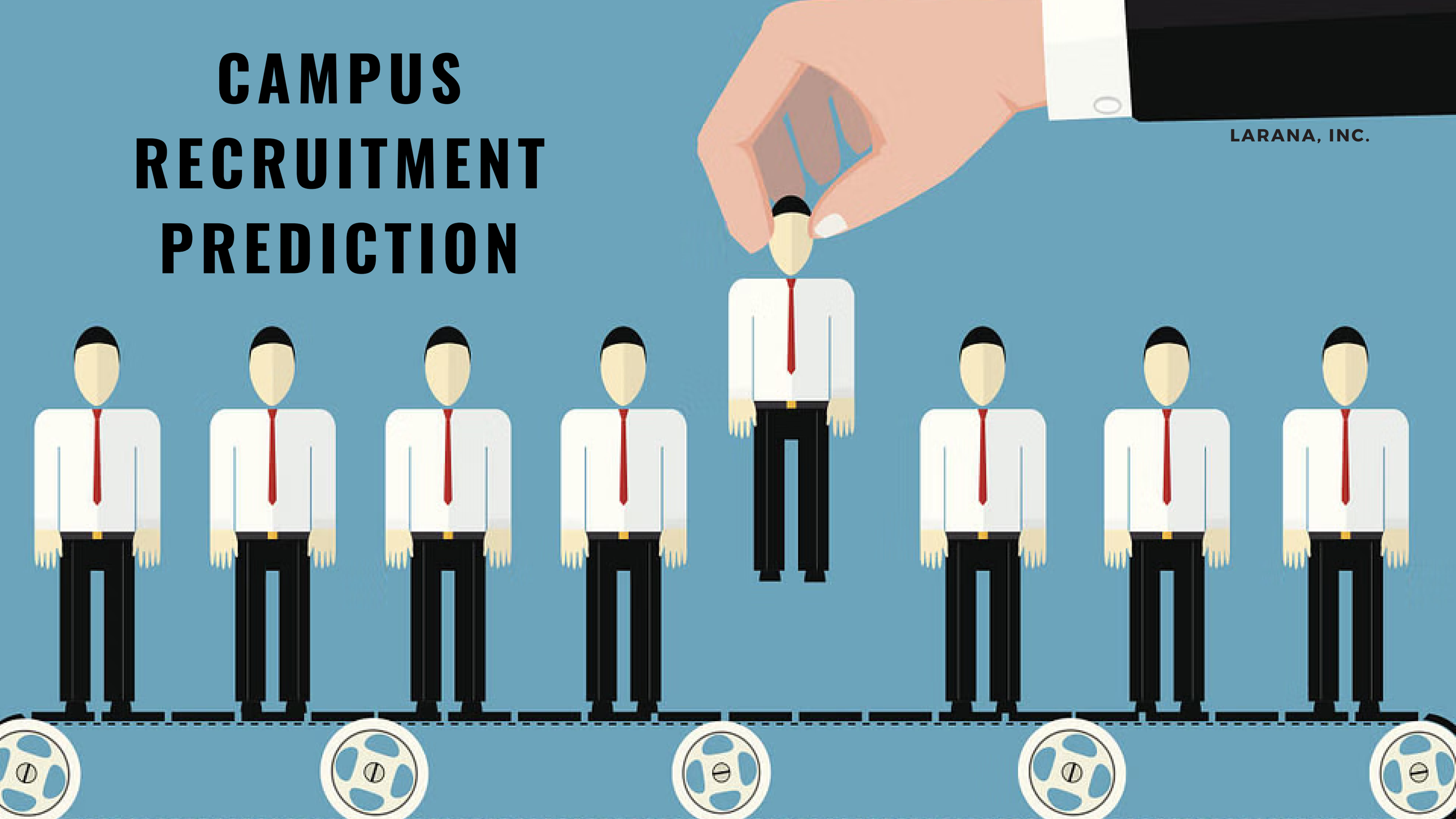
CAMPUS RECRUITMENT PREDICTION



CONTENT

INTRODUCTION

Placemnts play a very vital role in one's college tenure and is a deciding factor for the upcoming life he or she is to take.Placements are dependent on a variety of factors and here in this notebook we are going to make a machine learning model to predict the possibility of a final year student getting placed

PROBLEM STATMENT

- The Placement of students is one of the most important objective of an educational institution. Reputation and yearly admissions of an institution invariably depend on the placements it provides it students with. That is why all the institutions, arduously, strive to strengthen their placement department so as to improve their institution on a whole. Any assistance in this particular area will have a positive impact on an institution’s ability to place its students. This will always be helpful to both the students, as well as the institution.
- The main goal is to predict whether the student will be recruited in campus placements or not based on the available factors in the dataset

DATA DESCRIPTION

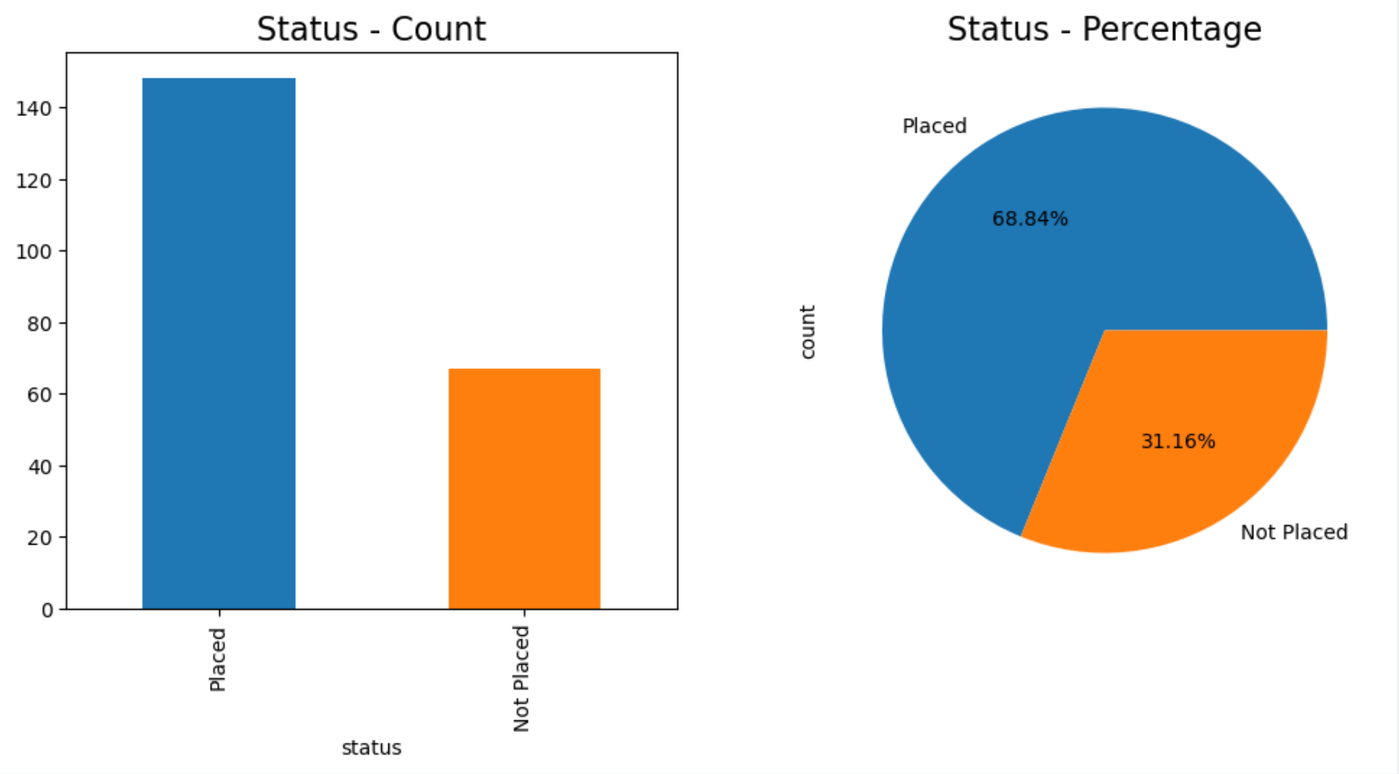
- sl_no: Serial number
- gender: Gender- Male='M',Female='F'
- ssc_p: Secondary Education percentage- 10th Grade.
- ssc_b: Board of Education- Central/ Others.
- hsc_p: Higher Secondary Education percentage- 12th Grade.
- hsc_b: Board of Education- Central/ Others.
- hsc_s: Specialization in Higher Secondary Education.
- degree_p: Degree Percentage.
- degree_t: Under Graduation(Degree type)- Field of degree education.
- workex: Work Experience.
- etest_p: Employability test percentage (conducted by college).
- specialisation : Post Graduation(MBA)- Specialization.
- mba_p: MBA percentage.
- status: Status of placement- Placed/Not placed.
- salary: Salary offered by corporate to candidates per anum (in Rupees).

CATEGORICAL VARIABLES

* gender
* ssc_b
* hsc_b
* hsc_s
* degree_t
* workex
* specialisation
* status

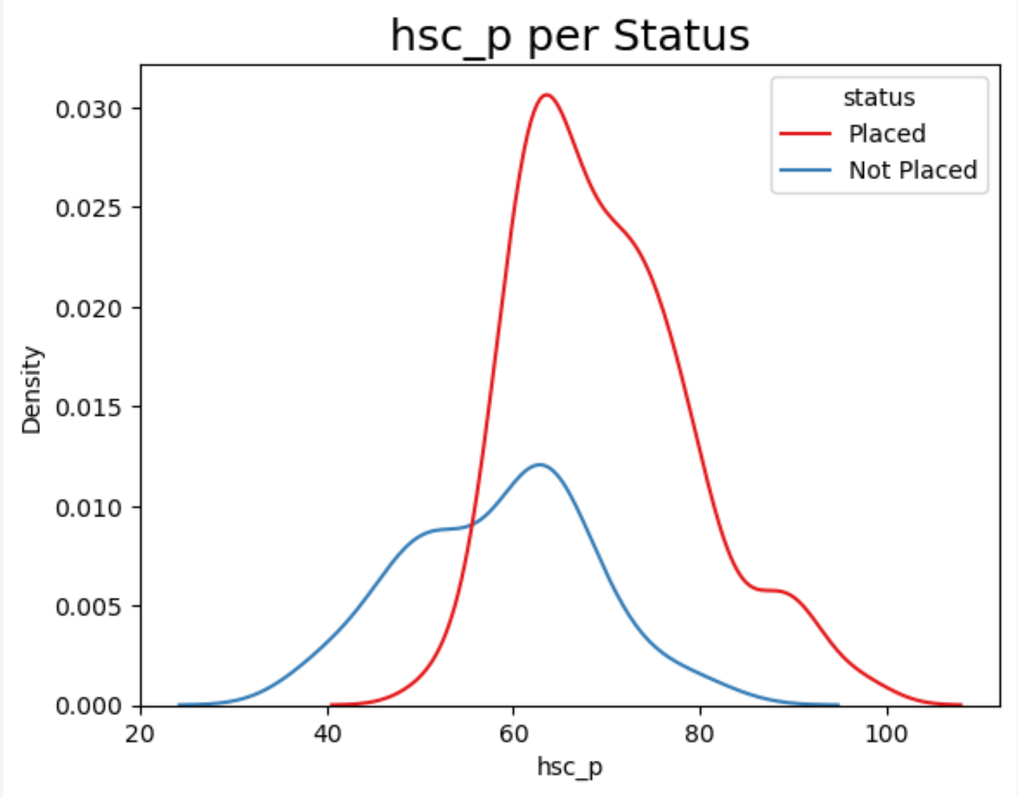
CONTINUOUS VARIABLES

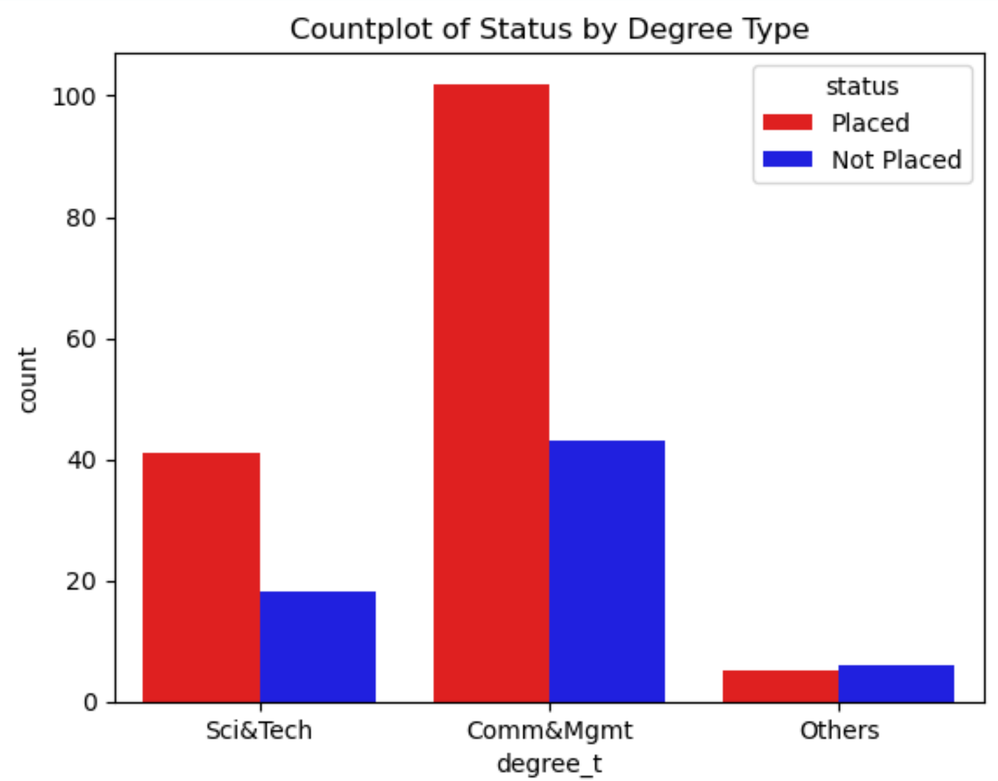
* ssc_p
* hsc_p
* degree_p
* etest_p
* mba_p
* salary



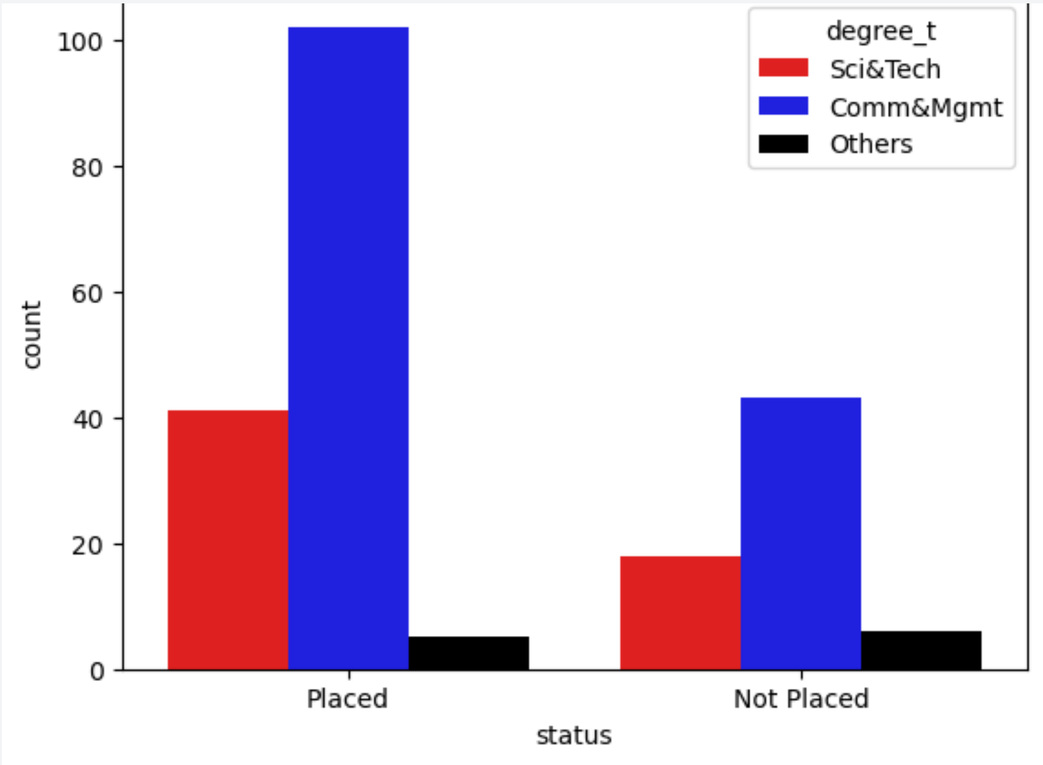
STATUS' IS THE TARGET COLUMN. WE CAN SEE THAT 68.8% OF THE DATA IS 'PLACED' AND 31.2% IS 'NOT PLACED

THE STUDENT WHO GOT HIGHER MARKS IN HIGHER SECONDARY EDUCATION PERCENTAGE- 12TH GRADE HAS A VERY HIGH CHANCE OF GETTING PLACED.¶

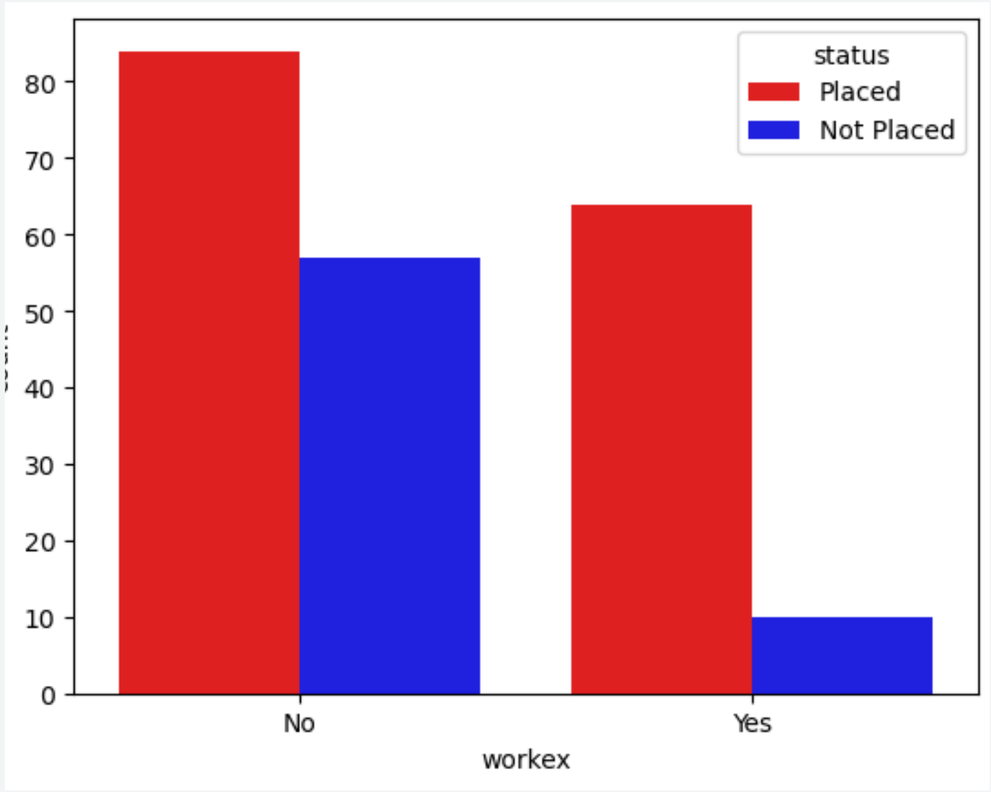




AS WE SEE IN THE GRAPH ITS VERY CLEAR THAT ,THE PEOPLE WHO HAD OPTED FOR COMMERCE AND MANAGEMENT IN DEGREE ARE THE ONES WHO GOT HIGHEST PLACEMENT

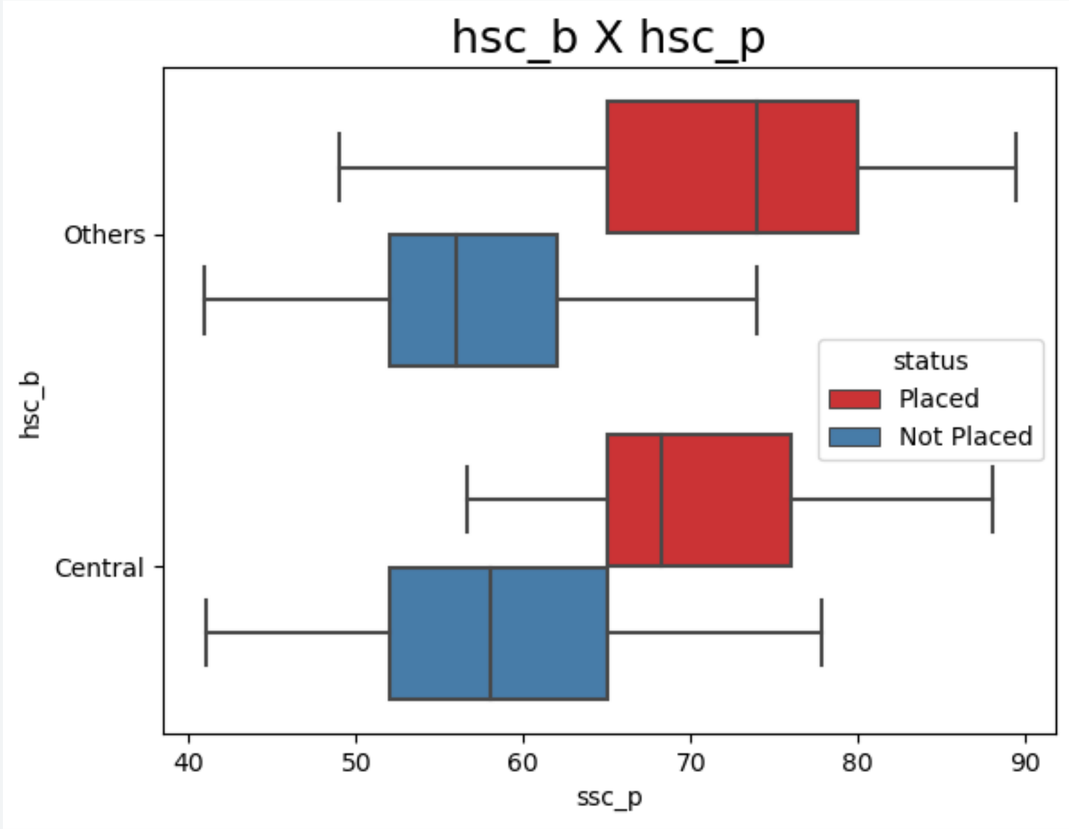


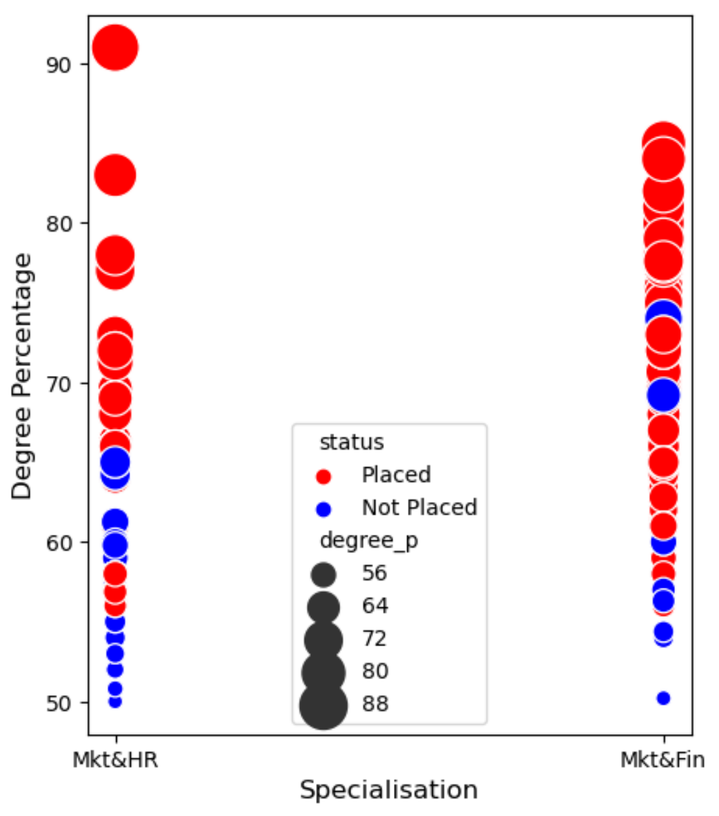
THE STUDENTS FROM 'COMM&MGMT' BACKGROUND HAVE BEEN PLACED IN LARGE NUMBERS WHEN COMPARED TO SCIENCE AND TECH ,AND OTERS.



THERE IS NO MUCH DIFFERENCE IN TERMS OF PLACEMENTS BUT THE STUDENTS WHO HAD WORK EXPERIENCE HAS ONLY FEW CHANCE TO GET REJECTED.

WE CAN GRAPHICALLY CONFIRM THAT BOTH 'CENTRAL' AND 'OTHERS' HAVE HIGHER AVERAGES FOR 'PLACED' IN 'HSC_B'





* WE CAN GRAPHICALLY CONFIRM THAT THE STUDENTS GETTING HIGH MARKS IN 'MKT&FIN' AND 'MKT&HR' HAS A HIGH CHANCE OF GET PLACED.
* ALSO THE STUDENTS WHO WERE OPTED FOR MARKETING AND FINANCE HAS BEEN PLACED AT GOOD NUMBERS WHEN COMPARE TO OTHER DEPARTMENT

INSIGHTS FROM THE EDA

1. 'WORKEX' PLAYS THE IMPORTANT ROLE IN PREDICTION. HIGHER THE VALUE OF 'WORKEX' MORE IS THE PROBABILITY OF GETTING PLACED.
2. 'SSC_P' IS ALSO IMPORTANT IN PREDICTION. HIGHER THE VALUE OF 'SSC_P' MORE IS THE PROBABILITY OF GETTING PLACED.
3. SIMILAR TREND FOR 'DEGREE_P' & 'HSC_P' I.E. HIGHER VALUES FOR THESE, HIGHER ARE THE CHANCES OF GETTING PLACED.
4. THERE ARE OTHER SMALLER TAKEAWAYS FROM THE MODEL HOWEVER THESE ARE CONSIDERED NOT SO IMPORTANT.

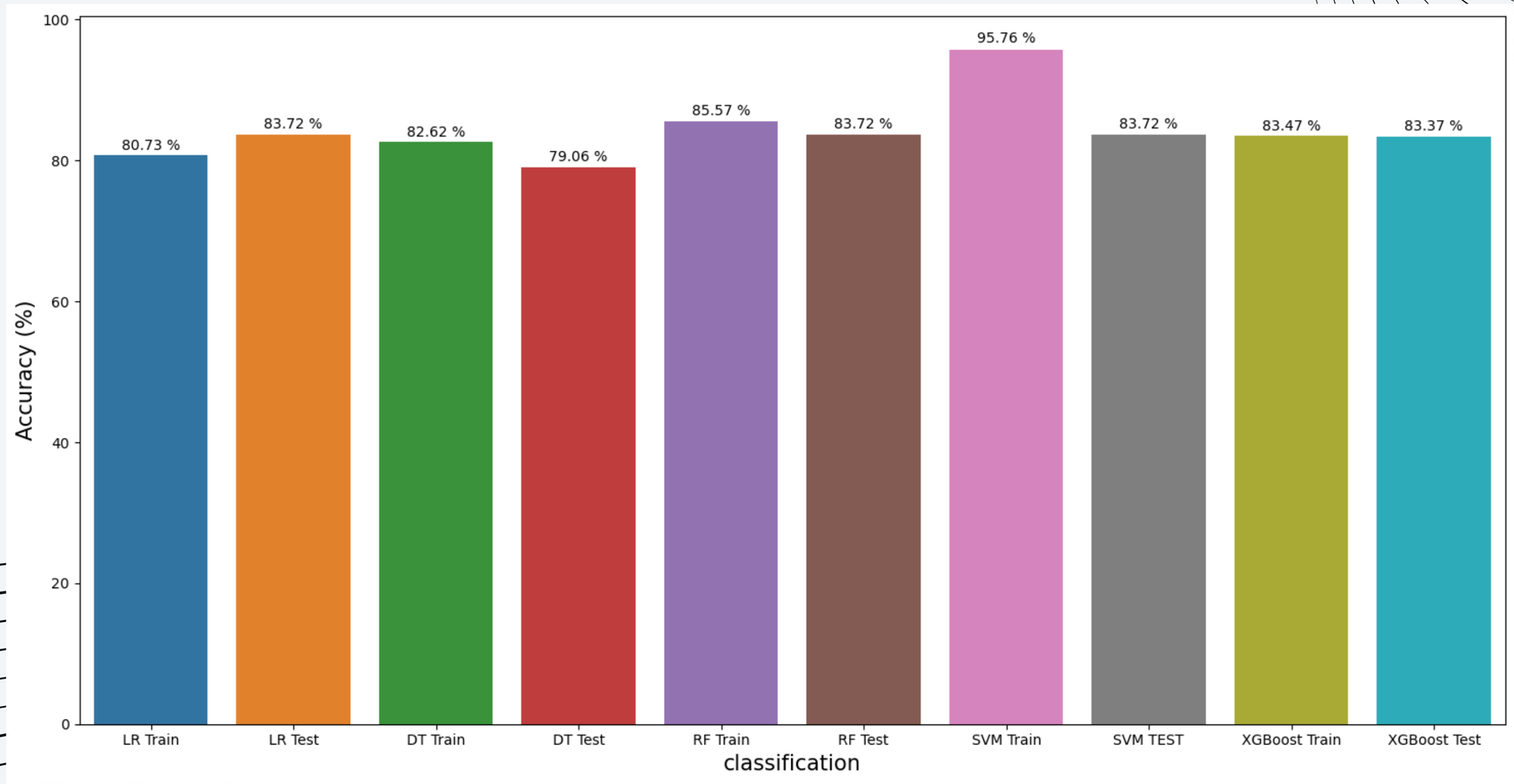
LARANA, INC.

CLASSIFICATION PROBLEM

I FOUND THERE WERE NULL VALUES PRESENT IN DATASET AND I HAVE TREATED IT , THERE WERE NO DUPLICATE VALUE PRESENT .

BEFORE MOVING TO MODEL BULIDING I FOUND THAT MY TARGET VARIABLE WAS IMBALANCED SO USING SMOTE I HAVE BALANCED MY TARGET COLUMN.

LATER ON I HAVE DONE LABEL ENCODING TO CONVERT MY CATAGORICAL COLUMN TO NUMERICAL COLUMNS.AND I HAVE USED STANDERD SCALER TO SCALE THE DATA.


100
80
60
40
20
0
Accuracy (%)
80.73 %
83.72 %
82.62 %
79.06 %
85.57 %
83.72 %
95.76 %
83.72 %
83.47 %
83.37 %
LR Train
LR Test
DT Train
DT Test
RF Train
RF Test
SVM Train
SVM TEST
XGBoost Train
XGBoost Test
classification

SUMMARY

THE LOGISTIC REGRESSION MODEL ACHIEVED AN ACCURACY OF 80.73% ON THE TRAINING DATA AND 83.72% ON THE TEST DATA WHICH LEADS TO UNDER FIITING.

THE DECISION TREE MODEL YIELDED HIGH ACCURACY LEVELS WITH 82.62% ON THE TRAINING DATA AND 79.06% ON THE TEST DATA.

THE RANDOM FOREST MODEL OBTAINED RESPECTABLE ACCURACIES OF 85.57% ON THE TRAINING DATA AND 83.72% ON THE TEST DATA.

THE SVM MODEL GOT 95.76% ACCURACY IN TRAINING AND 83.72% IN TESTING ,WHICH LEADS TO OVERFITTING.

FINALLY IN XGBOOST I GOT A BETTER ACCURACY OF 83.47 IN TRAINING AND 83.37 IN TESTING , AS XGBOOST IS OPTIMIZED FOR SPEED AND PERFORMANCE AND EVEN HERE I GOT BEST ACCURACY IN XGBOOST .

THANK'S FOR WATCHING