CAMPUS RECRUITMENT ANALYSIS

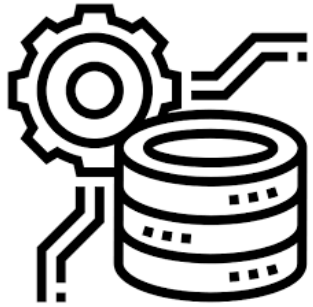
PYTHON PROJECT

Nupur Agnihotri| Neelam Arya| Sunita Barik | Jingxin Yao | Zihan Huang

Methodology

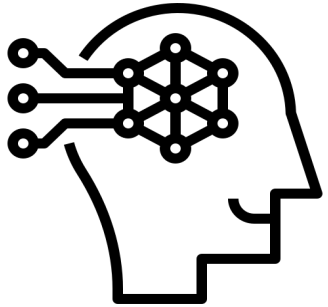
Research Questions

Data Description



Data Preprocessing

Descriptive Analysis



Machine Learning

Recommendation

Research Questions

- How to increase the rate of placement in the college?
- How to improve a student’s academic standing to boost their confidence for placements?

Dataset Description

The dataset is from a campus recruitment drive in an MBA college, in India. It has the credentials of all students from primary education to bachelor's and then MBA and their previous work experience.

The inputs are all the columns related to a candidate's credentials and the output column is 'status'.

Variable	Description	Datatype
sl_no	Index of records / Serial Number	Numerical
gender	Gender- Male='M', Female='F'	Categorical
ssc_p	Secondary Education / Matriculation percentage- 10th Grade	Numerical
ssc_b	Board of Education- Central/ Others	Categorical
hsc_p	Higher Secondary Education percentage- 12th Grade	Numerical
hsc_b	Board of Education- Central/ Others	Categorical
hsc_s	Specialization in Higher Secondary Education	Categorical
degree_p	Degree Percentage	Numerical
degree_t	Under Graduation(Degree type)- Field of degree education	Categorical
workex	Work Experience	Categorical
etest_p	Employability test percentage (conducted by college)	Numerical
specialisation	Post-Graduation(MBA)- Specialization	Categorical
mba_p	MBA percentage	Numerical
status	Status of job placement- Placed/Not placed	Categorical
salary	Salary offered by companies	Numerical

Descriptive Analysis
And
Data Visualization

Descriptive Analysis

Number of columns and rows: 215 rows and 15 columns.

Missing values: Only the 'salary' column has missing values, the percentage of that is 31.26%.

	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	0	0	0	0	0	0	0	0	0	0	0	0	0	67

Modes for all variables: It shows all the frequently occurring values in each variable.

gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
M	62.0	Central	63.0	Others	Commerce	65.0	Comm&Mgmt	No	60.0	Mkt&Fin	56.7	Placed	300000.0

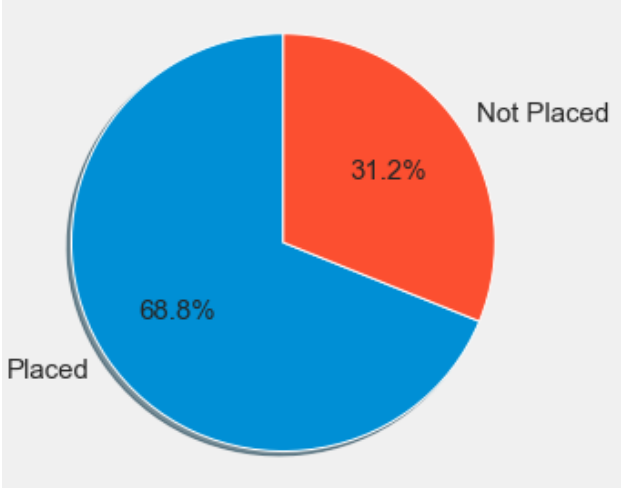
Descriptive Analysis Continued…

Descriptive statistics for all the numerical variables in the dataset:

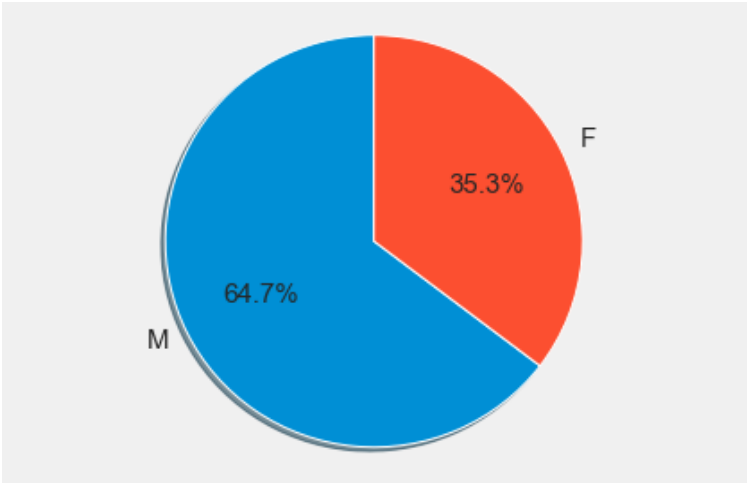
	count	mean	std	min	25%	50%	75%	max
ssc_p	215.0	67.303395	10.827205	40.89	60.600	67.0	75.700	89.40
hsc_p	215.0	66.333163	10.897509	37.00	60.900	65.0	73.000	97.70
degree_p	215.0	66.370186	7.358743	50.00	61.000	66.0	72.000	91.00
etest_p	215.0	72.100558	13.275956	50.00	60.000	71.0	83.500	98.00
mba_p	215.0	62.278186	5.833385	51.21	57.945	62.0	66.255	77.89
salary	148.0	288655.405405	93457.452420	200000.00	240000.000	265000.0	300000.000	940000.00

Distribution Based on Categorical Variables

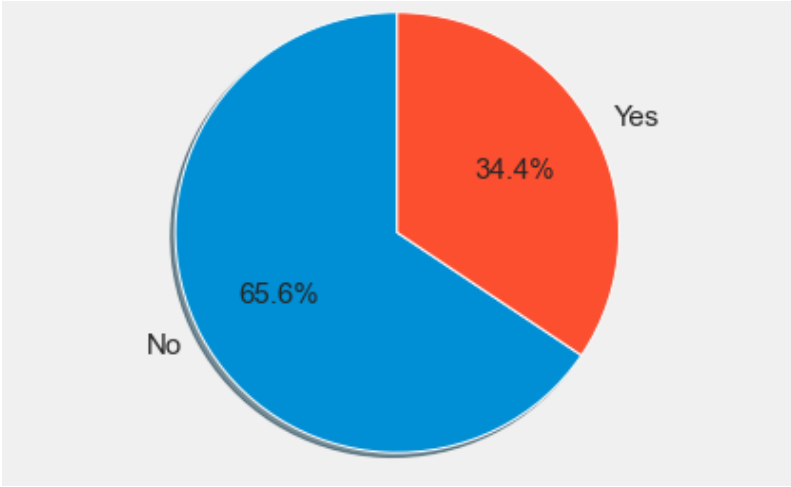
Status of placement



Gender



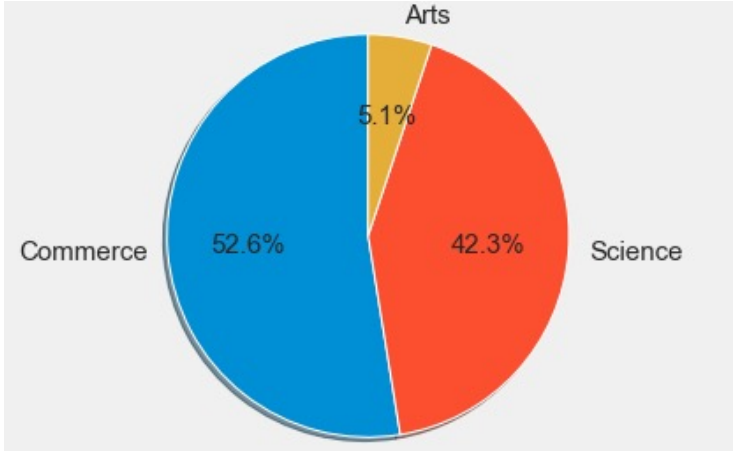
Work Experience



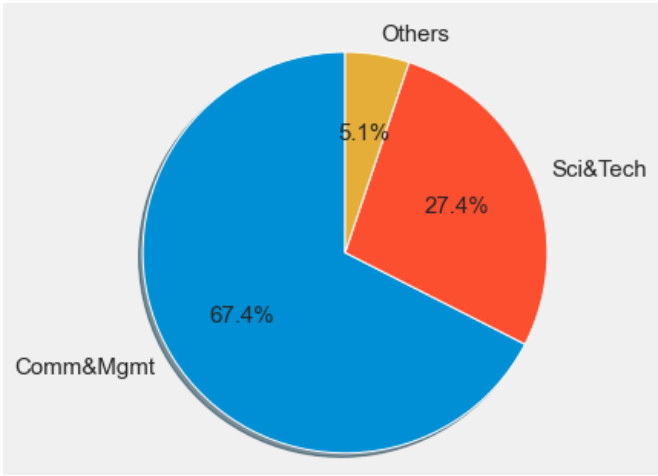
In these cases, pie charts give us an overall distribution of all students' data based on following categorical variables: status of placement, gender and work experience.

Distribution Based on Categorical Variables

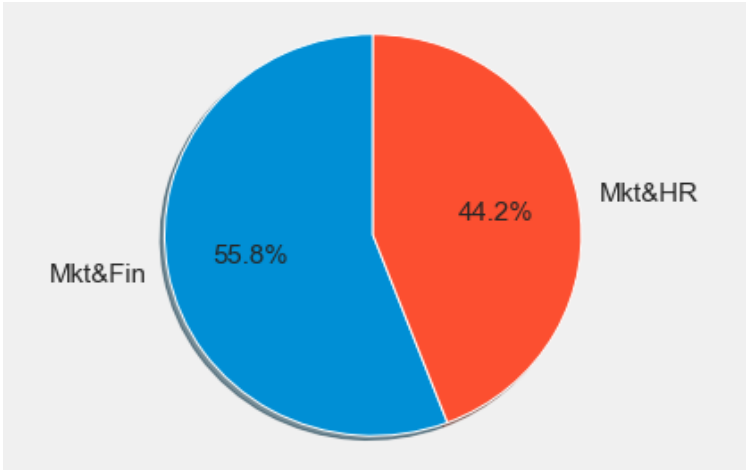
Higher Secondary Education



Under Graduation



Post Graduation



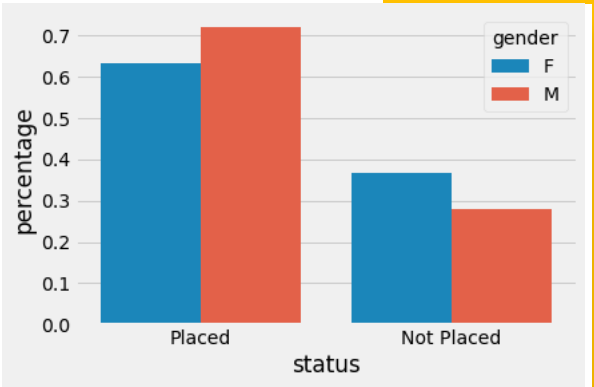
In these cases, pie charts give us an overall distribution of all students' data based on following categorical variables: higher secondary education stream, undergraduate major and post-graduation specialization.

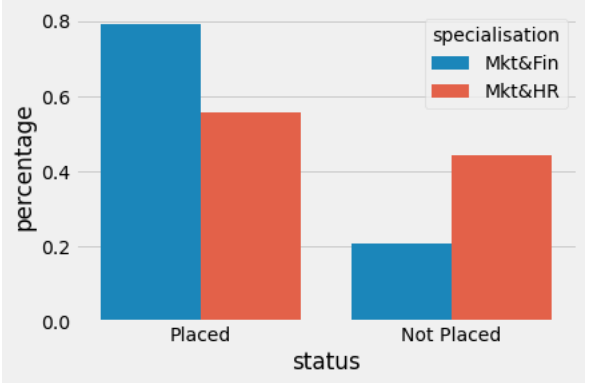
Impact of Gender, Specialization & Work-Experience on Placement

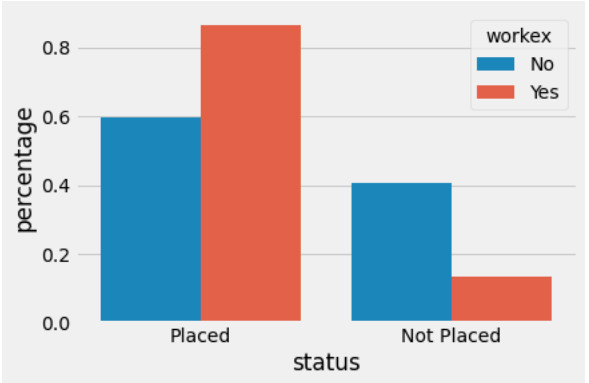
Majority of the male students got placed, whereas the percentage of placed females is lesser: 'Gender' could be a significant variable

Students with work ex are preferred by the recruiters, and the difference is visually significant between these categories so this may be a significant variable

Marketing & Finance MBAs are preferred over marketing & HR MBAs by the recruiters, so specialization could be significant variable

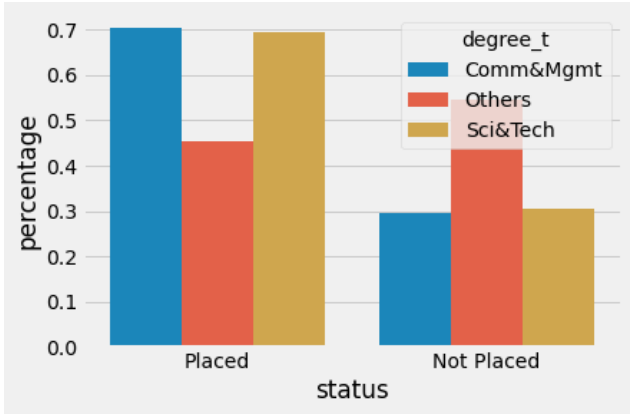


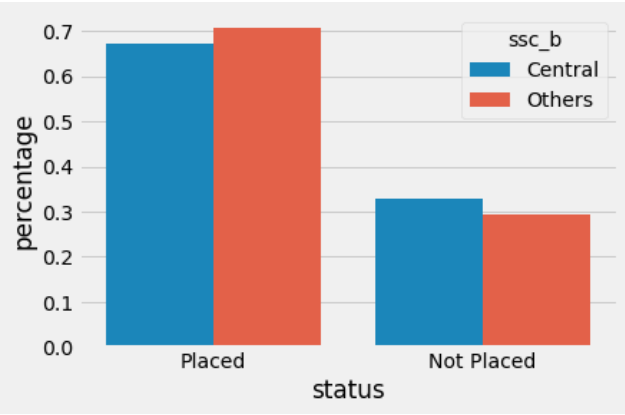


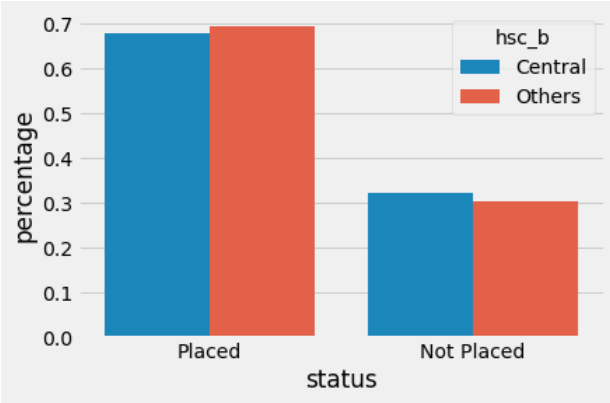


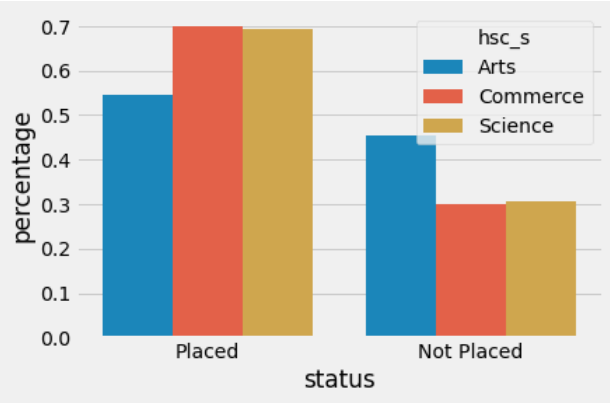
Impact of Levels of Education on Placement

It shows that there isn't much difference of percentage within these levels of education, as the percentage of students from any of these categories are similarly placed and unplaced. So, these may not be a useful variables for classification.









Salary Distribution

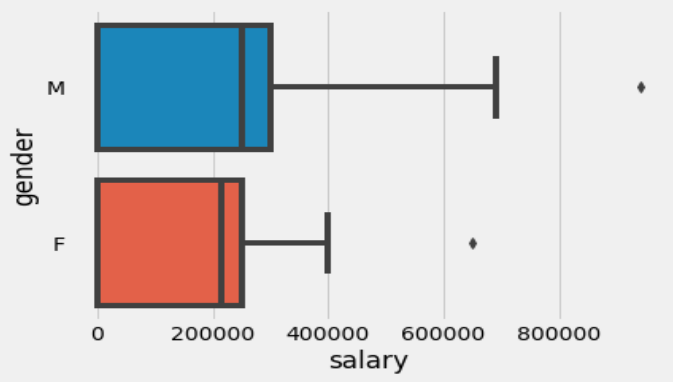
Male students are offered better salaries than females

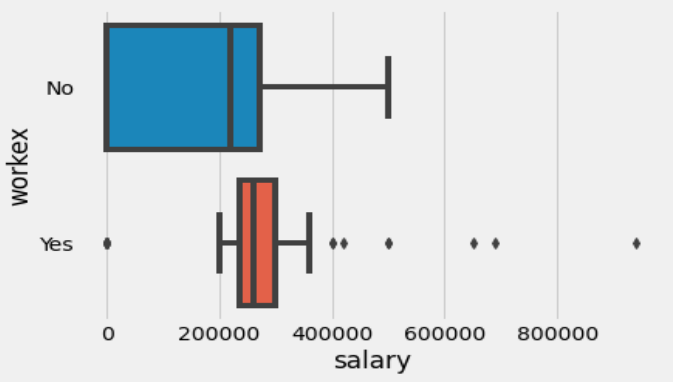
Students with work ex are offered better salaries

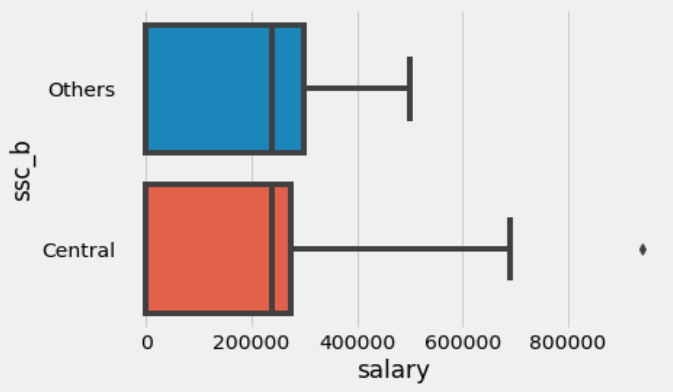
Board of education and high school streams have no impact on placement and salaries

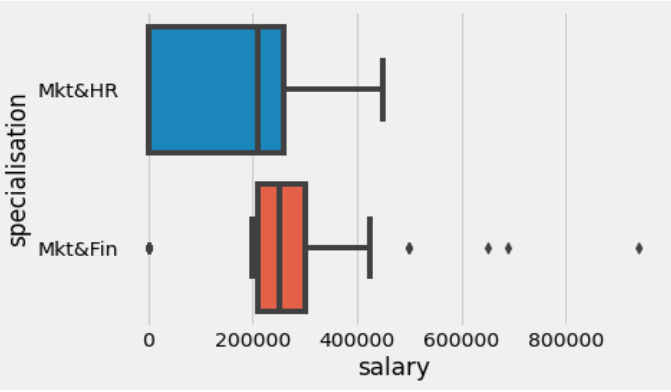
Marketing & Finance students are offered better salaries

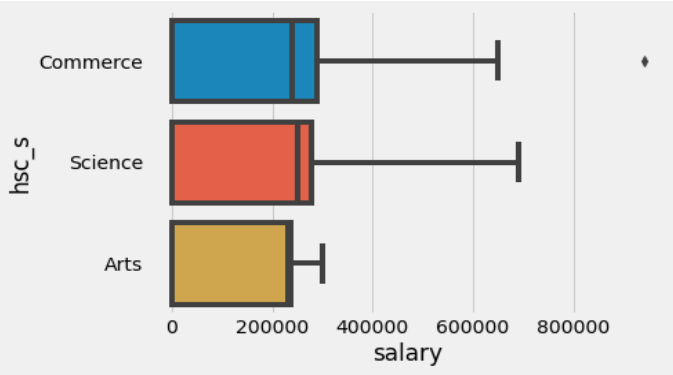
Science and commerce degree students are offered better salaries

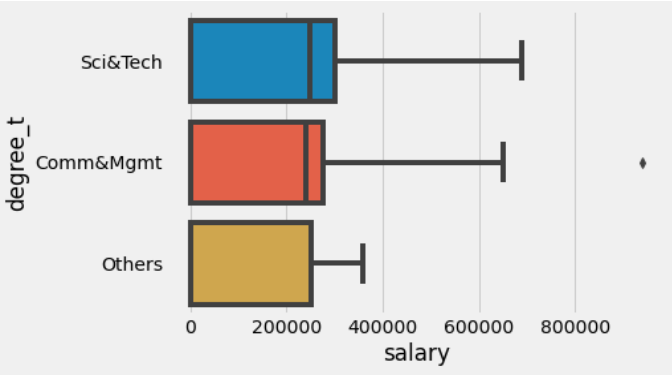






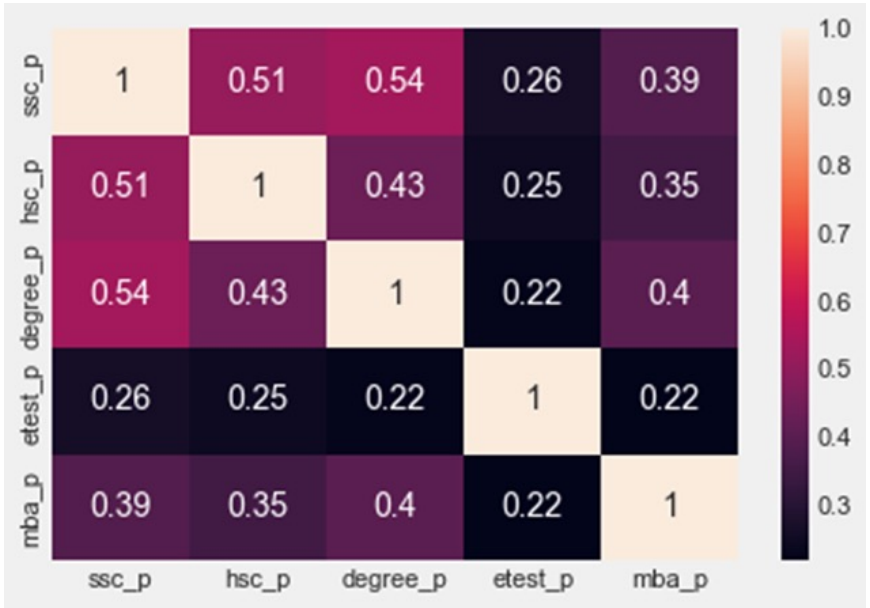


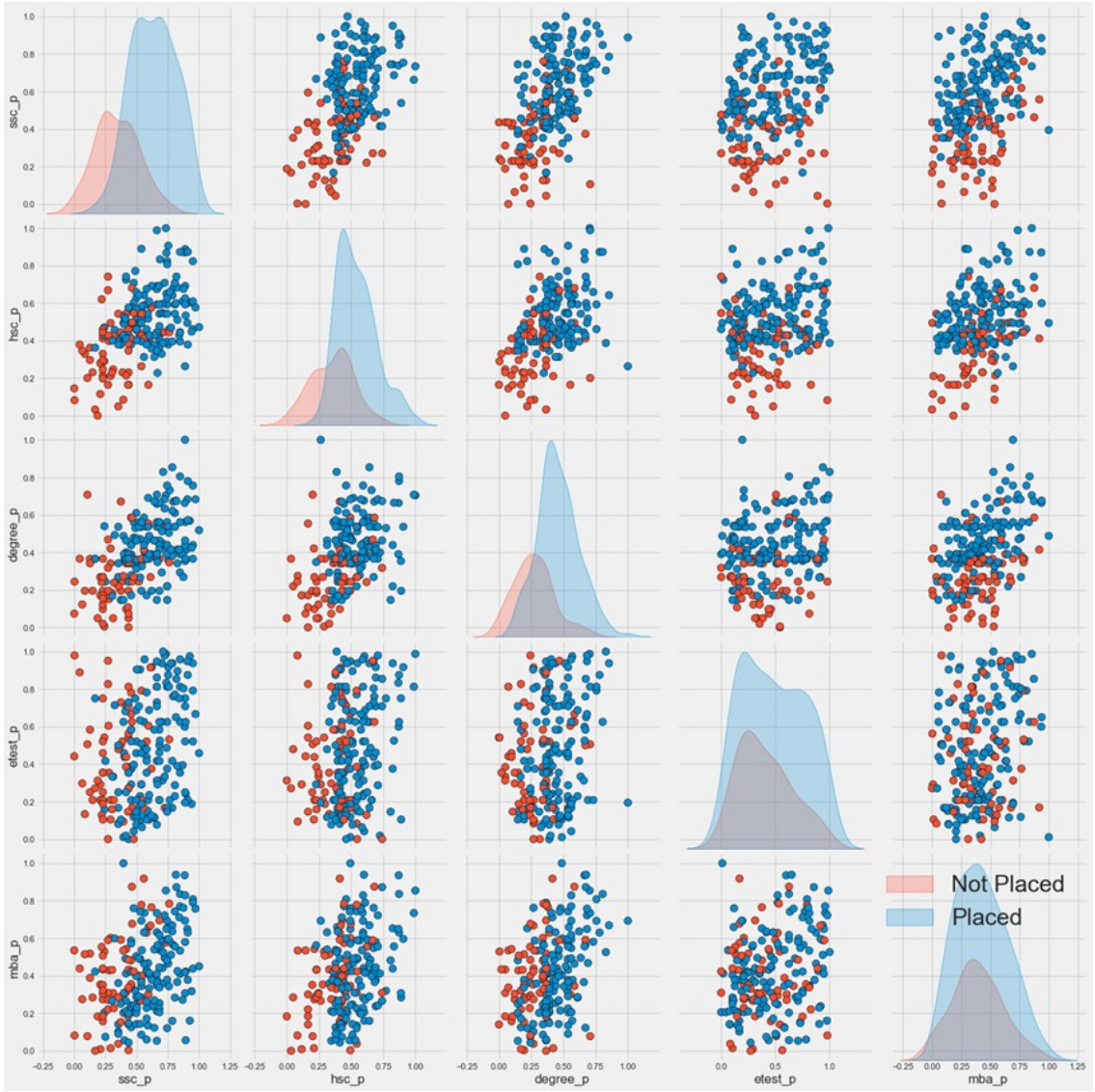




Correlation B/W Variables

As the correlation matrix graph depicts, it is possible that percentile of ssc has a strong association with the percentile of degree and that of hsc. We could make a hypothesis that they might influence a student's placement status significantly.





Data Preprocessing
And
Model Fitting

Data Preprocessing

1. Identified numerical and categorical variables
2. Dropped zero variance numerical variables
3. Dropped high cardinality categorical variables (I.e.,- sl_no)
4. Missing Data Analysis: Students not-placed had missing salaries, so we imputed ‘zero’ for salaries
5. Label Encoded categorical variables (Gender, ssc_b, hsc_s, Specialization, Work-Ex, Degree_t, Status)
6. Standardized all the numerical variables to avoid misclassifications and high collinearity

Fitting models

	accuracy	sensitivity	specificity
Logistic Regression	0.80	0.63	0.88
Random Forest	0.56	0.19	0.73
K-NN	0.69	0.07	0.97
Decision Tree	0.56	0.30	0.68

1. Defined independent variable dataset as ‘X’ and dependent variable as ‘y’
2. Split the dataset into 60% of training set and 40% of testing set
3. Fit the training dataset on following classifier models:
   - Random Forest Classifier
   - KNN Classifier
   - Logistic Regression
   - Decision Tree Classifier
4. Before fitting logistic regression, we checked the model summary to remove insignificant variables
5. We finally achieved the best accuracy of 80% with logistic regression
6. We plotted the ROC curve to assess the model as final measure

Recommendation

- Setup career development or Co-op opportunities for inexperienced students
- Invite more companies for arts students
- For students with lower SSC and HSC percentages, university can provide some fundamental academic courses to get them prepared for more advanced courses

Limitation of the Dataset

- Smaller Dataset: Larger the number of records, the better the model would be
- Two specializations merged in one: Break down the two specializations, Mkt&Fin and Mkt&HR, into Marketing, Finance, and HR separately, to understand their independent contributions
- Years of Experience: Provide years of experience for every student


Thank You
ISABELLE
FORBRICH
KELLY
ENGLISCH