Project 3

Team 13

Nicolas Renaud

Introduction:

The data set used in this project was ‘Campus Recruitment’ found on the Kaggle website posted by Ben Roshan. The data set contain data about students undergrad and graduate experience at Jain University Bangalore as well is if they were placed in a job after graduation and their salary at that job. Although this is not an American university, I found this data set interesting to see what features a recent graduate student possess that make them appeal to be hired. The rest of this paper is broken down into four additional sections:

1. Data
   1. Feature Definitions
   2. Data Cleaning
   3. Data Exploration (graphs and summary statistics)
2. Statistical Analysis
   1. Hypothesis Testing (is difference in average salary between groups significant)
   2. Linear Regression (predicting the salary feature)
3. Machine Learning
   1. Classification (decision tree classification of status feature)
4. Conclusion
   1. Summary of the Results

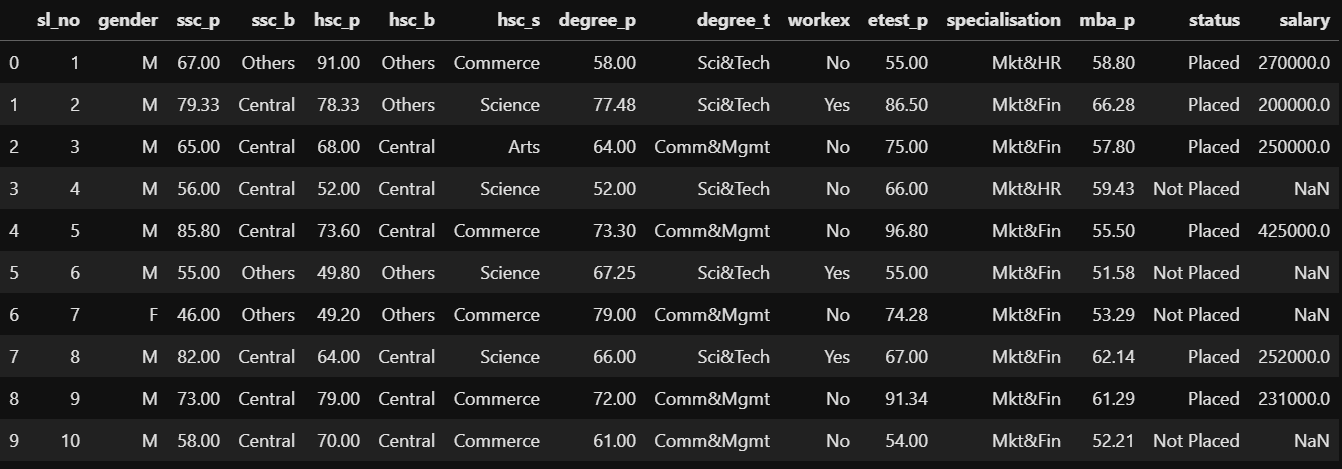
Data:

The ‘Campus Recruitment’ data set contains the following variables:

|  |  |
| --- | --- |
| Variable | Definition |
| sl\_no | A serial number used as a unique index of the student |
| gender | The gender of the student (M = male, F=female) |
| scp\_p | Secondary education percentage (10th grade) |
| ssc\_b | Board of secondary education (Central or Others) |
| hsc\_p | Higher secondary education percentage (12th grade) |
| hsc\_b | Board of higher secondary education (Central or Others) |
| hsc\_s | Specialization in higher secondary education (Commerce, Science, or Arts) |
| degree\_p | Undergraduate degree percentage |
| degree\_t | Undergraduate degree field (Comm&Mgmt, Sci&Tech, or Others) |
| workex | Work experience (Yes or No) |
| etest\_p | Employability test percentage |
| specialisation | MBA specialization (Mkt&Fin or Mkt&HR) |
| mba\_p | MBA percentage |
| status | Job placement status (Placed or Not Placed) |
| salary | Salary offered to candidate (in Indian Rupees) |

Before the data cleaning process, the dataset contained 215 records and 14 features some were numerical, and some were categorical.

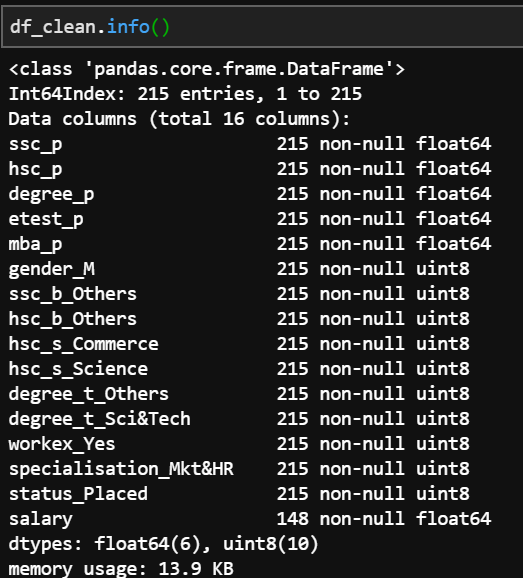
Cleaning:

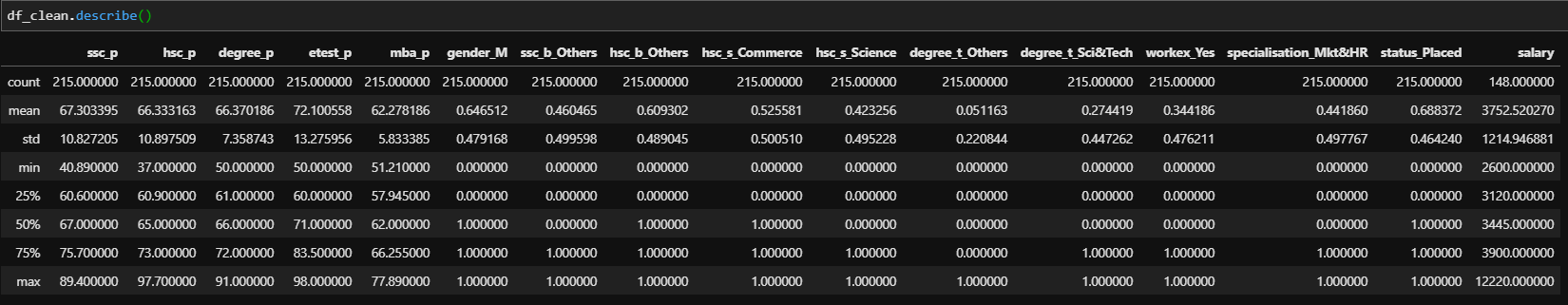


After reading in the dataset the first thing that was done as part of cleaning was to set sl\_no to the index because it is unique for every entry in the dataset. Next, the salary column which was initially in Indian Rupees was converted to US dollars, accomplished by multiplying the column in the data frame by 0.013 (at the time this project was conducted 1 Indian Rupee was equal to $0.013). The final step in the data cleaning process was to use one hot encoding to convert the categorical variables ('gender', 'ssc\_b', 'hsc\_b', 'hsc\_s', 'degree\_t', 'workex', 'specialisation', and 'status') into dummy variables and drop one of the dummy variables to avoid problems later on with model fitting. At the end of data cleaning there were two datasets created one with null values in the salary feature containing 148 records and 16 features (to be used on the linear regression model) and another which kept the null values containing 215 records and 16 features (to be used for the decision tree since the salary column is not included in the model).

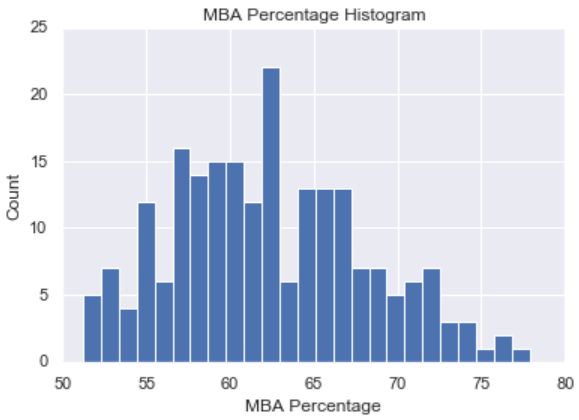
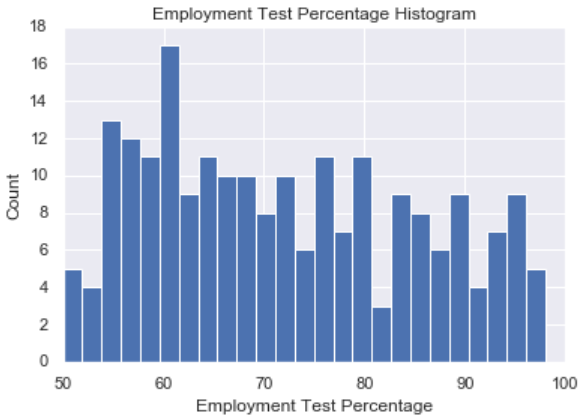
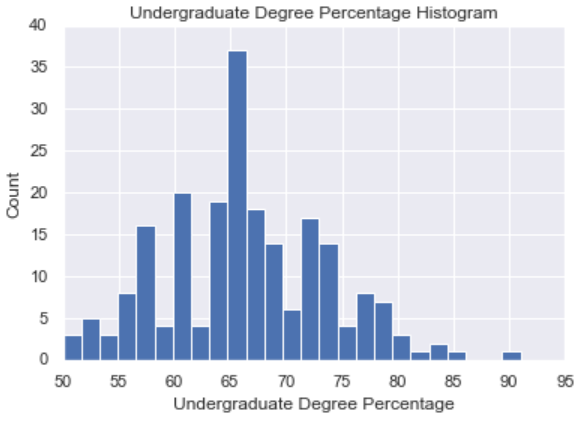
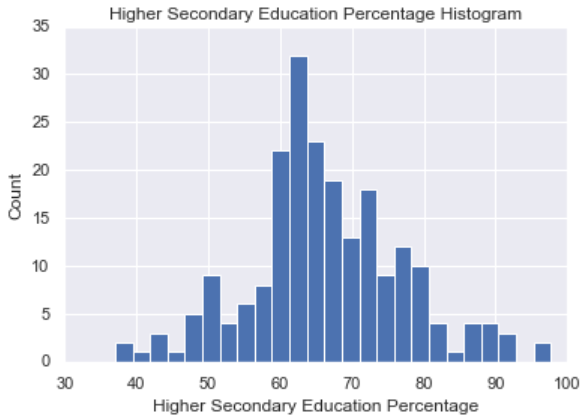
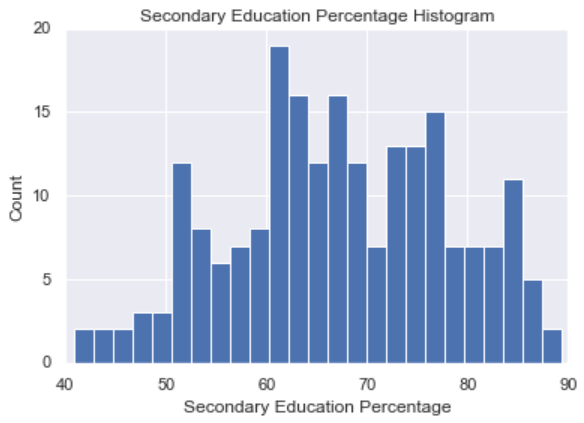
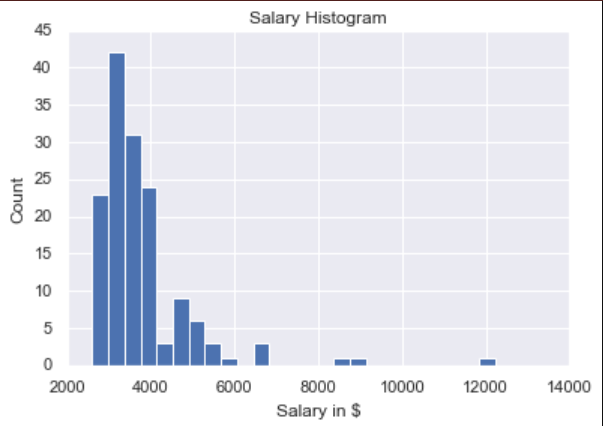
Exploration:

The final variables are described and the two tables below. The tables show that the only variable with any missing values still is salary as well as some summary statistics of each of the features (count, mean, standard deviation, minimum, 25, 50, 75 percentiles, and maximum).

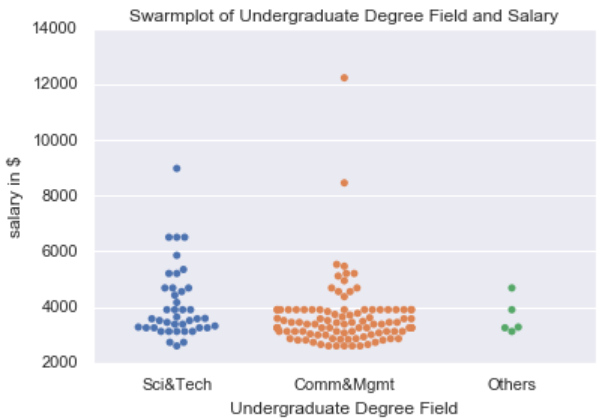
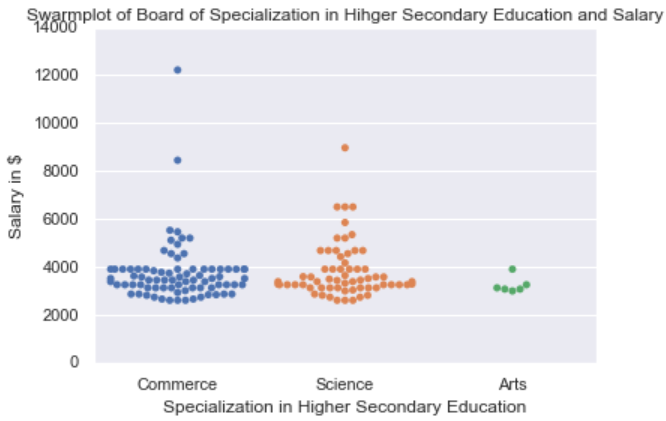
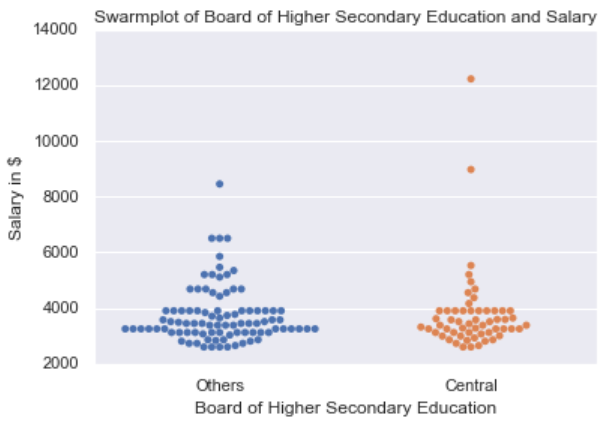
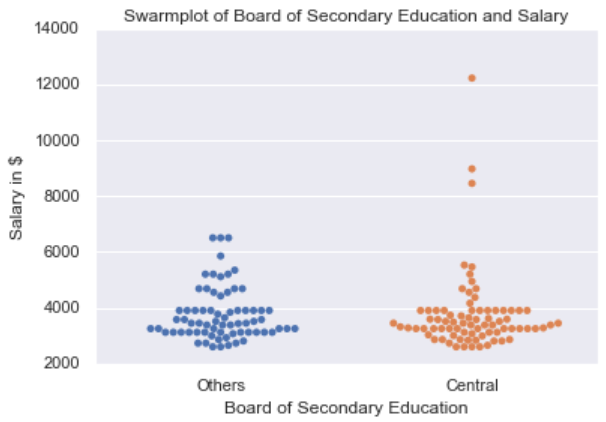
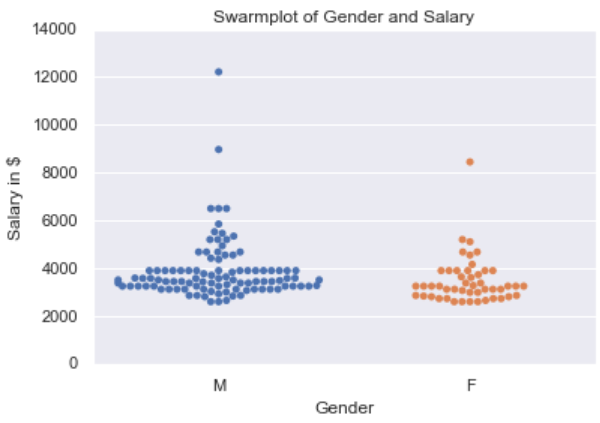


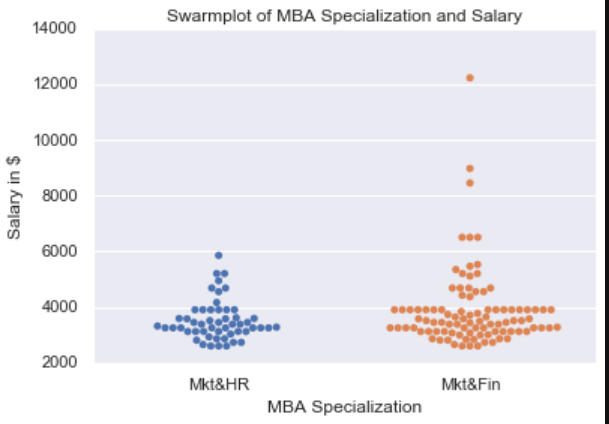


From looking at the summary statistics salary may be skewed to the left and there is at least one record that is pulling up the mean. Next I moved onto some visualization to explore the data and various features using histograms and swarm plots.



For the most part the numerical features appear to have a normal distribution apart from salary which seems skewed as well as employment test percentage which also seems to not be normally distributed. Next I move onto swarm plots to see if there are differences between various category groups and their salaries.





From this graphical exploration it would appear there may be a difference between male and female salaries as well as marketing and finance MBA specialization versus marketing and human resources specialization in salary. These as well as other differences that may be harder to see from a swarm plot will be explored in the next section on statistical analysis and hypothesis testing.

Statistical Analysis:

Hypothesis Testing:

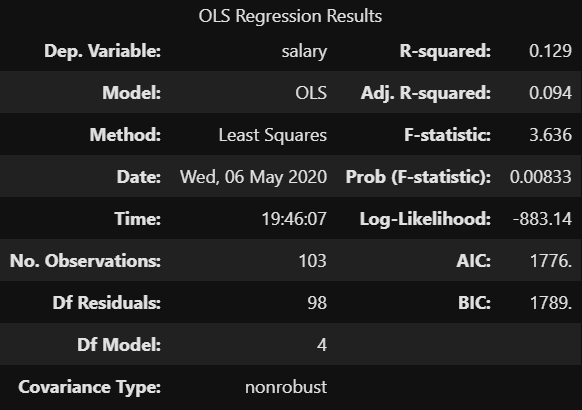
S. refers to Secondary Education, H.S. refers to Higher Secondary, U.G. refers to undergraduate.

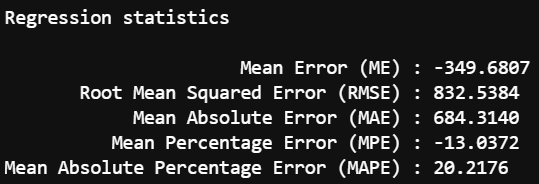
\* indicate significant at 10% level, \*\* indicates significant at the 5% level.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature 1 | Mean 1 | Feature 2 | Mean 2 | P-Value |
| Male: | **3885.83** | Female: | **3474.79** | 0.053\* |
| S. Board Other: | **3745.96** | S. Board Central: | **3746.16** | 0.956 |
| H.S. Board Other: | **3745.28** | H.S. Board Other: | **3764.07** | 0.927 |
| H.S. Specialization Commerce: | **3736.43** | H.S. Specialization Science: | **3822.20** | 0.682 |
| H.S. Specialization Commerce: | **3736.43** | H.S. Specialization Arts: | **3232.66** | 0.349 |
| H.S. Specialization Science: | **3822.20** | H.S. Specialization Arts: | **3232.66** | 0.219 |
| U.G Field Comm&Mgmt: | **3622.15** | U.G. Field Others: | **3645.20** | 0.965 |
| U.G Field Comm&Mgmt: | **3622.15** | U.G Field Sci&Tech: | **4089.92** | 0.039\*\* |
| U.G Field Sci&Tech: | **4089.92** | U.G. Field Others: | **3645.20** | 0.458 |
| Work Experience: | **3942.45** | No Work Experience: | **3607.80** | 0.097\* |
| MBA Specialization Mkt&Fin: | **3885.08** | MBA Specialization Mkt&HR: | **3514.90** | 0.075\* |

After conducting the hypothesis testing there was only one difference between groups that was significant at the 5% level and three other that were significant at the 10% level. Those who attended their undergraduate degree in the field of communications and management had an average salary lower than the average salary of those who attended their undergraduate degree in the field of science and technology, this was significant at a 5% level. Additionally men had an average salary higher than women, those with work experience had a higher salary than those without work experience, and MBA graduates specializing in marketing and finance had a higher average salary than MBA graduates specializing in marketing and human resources, these were each significant at the 10% level.

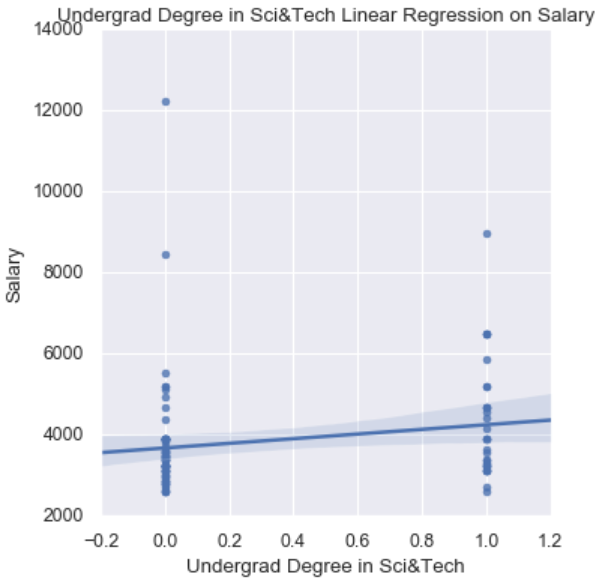
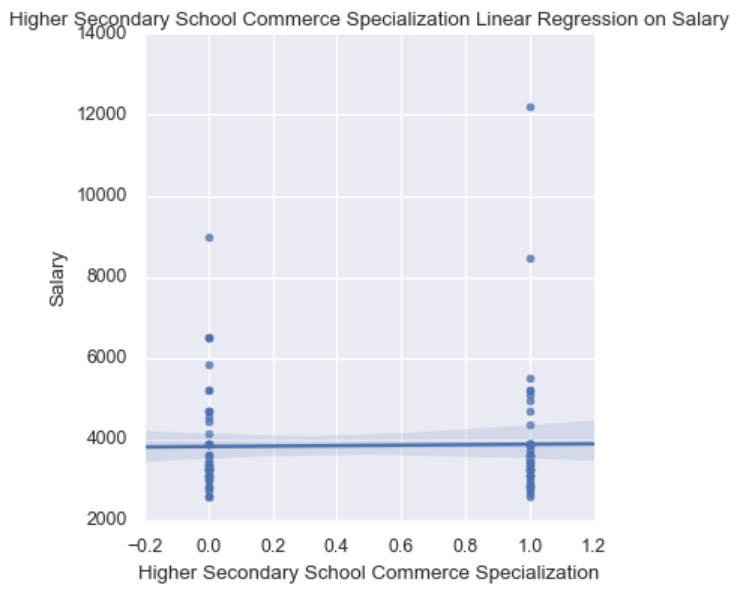
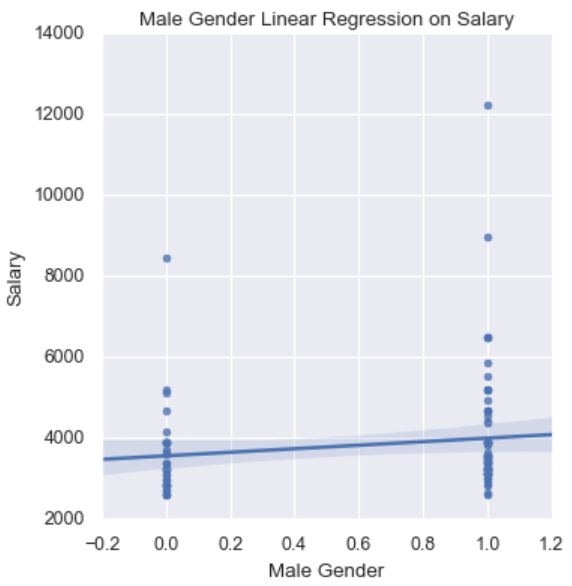
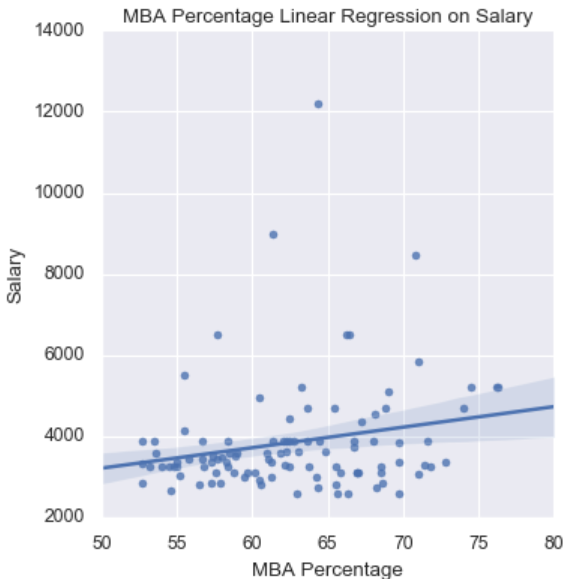
Linear Regression:





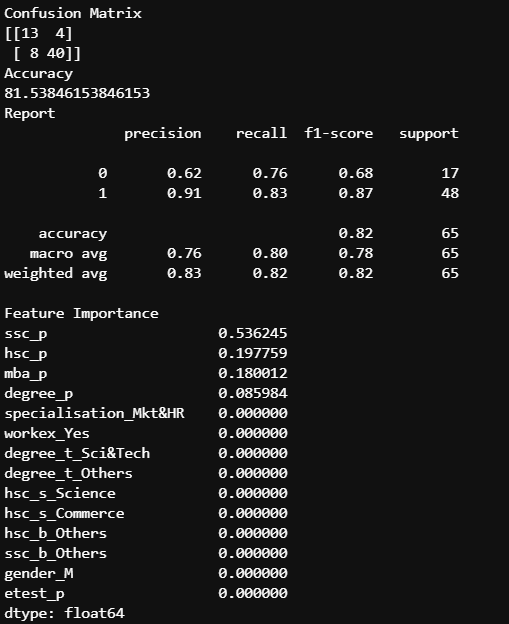
The above tables show the statistics and metrics related to the linear regression used to predict salary with the variables in this data set. The data was split into training and testing sets using a 70% split for training data and the random state of 717. This regression was found using a backward elimination process for feature selection and contains the variables for MBA percentage, male, H.S. specialization commerce, and U.G. degree in science and technology. The final formula for this regression is shown below as well as graphs showing how each individual variable predicts salary by itself:

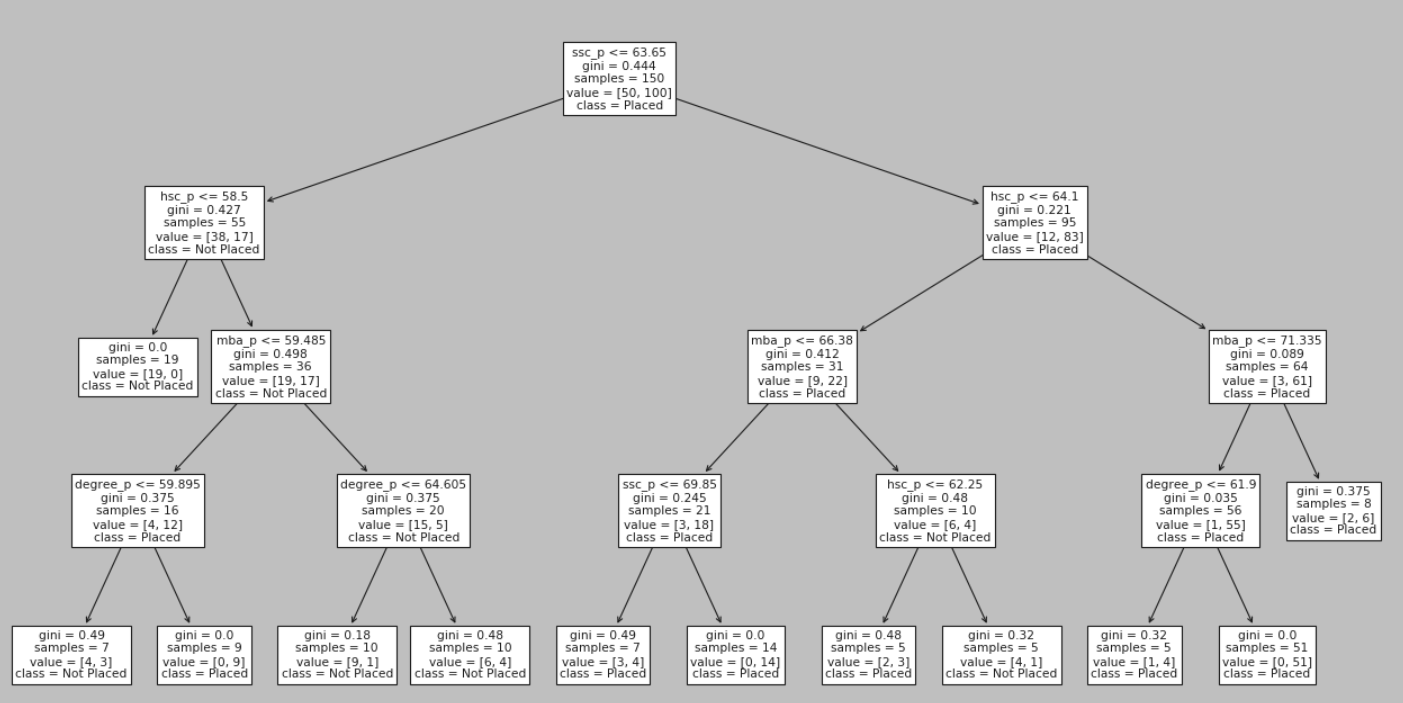
Salary = -429.8484 + 55.3856(MBA Percentage) + 488.6161(Male) + 519.8344(H.S. Commerce) + 776.3383(U.G. Science and Technology)



Machine Learning:

The machine learning task done using this data was classification using a decision tree to predict whether the student was placed in a job or not after graduation. First the data was split using the same method as the linear regression, a 70% training 30% testing split with a random state of 717, however the variable salary was not included because there was no salary for those not placed including that would allow the model to perfectly predict whether the student was placed in a job or not. Then the decision tree was fit using the training data, the testing data was used to make predictions and calculate evaluation metrics like accuracy, precision recall, a confusion matrix, the most important features, and f1 score. Finally, the decision tree was plotted. Below are the evaluation metrics and the plot of the final decision tree.





Conclusion:

In conclusion the above linear regression containing the variables MBA percentage, male, higher secondary commerce, and undergraduate science and technology can account for 9.4% of the variation in salary. While the decision tree had an overall accuracy of 81.5% and found secondary education percentage, higher secondary education percentage, MBA percentage, and undergraduate degree percentage to be important features. The decision tree fared well with precision, recall, and f1 score for job placement of 91%, 83%, and 87% respectively. However, the decision tree performed worse with no job placement obtaining a precision, recall, and f1 score of 62%, 76%, and 68% respectively.

Neither model produced is perfected and should be improved in future work with additional variables and records to increase performance. For example, a variable containing where the student attended undergraduate school, how many years of work experience, or their age may improve one or both models’ overall performance. One striking thing that this data does show is a discrepancy between male and female salaries, more work should be done on this to see if it is the result of some other variable that could be correlated to gender like degree field, specialization, or work experience or if it is alternatively due to gender discrimination.