ABSTRACT

The purpose of this project is to conduct data analysis and interpret the results of Multiple Regression Model of Placement Data of a class of MBA students with specialisation in Marketing & Finance, and Marketing & HR. The objective is to find out the factors that have the most effect on Placement of students from a given set of factors and analyze them. For this problem, Python Programming Language is used on Jupyter Notebook as it is a good tool to process and model data. Two analyses are performed on the dataset: Multiple Linear Regression taking salary (continuous) as the dependent variable, and Binary Logistic Regression taking placement status (categorical) as the dependent variable. Feature Selection is used to get the top features that affect the dependent variable in the models. For both the models, the top features obtained are the secondary education percentage, the higher secondary education percentage, the under graduation degree percentage, and work experience. When all the assumptions for both the models are met, analysis is performed, where both the models give satisfactory results.

INTRODUCTION

In general recruitment terminology, placement refers to the successful allocation of a person to a job. People usually go for a masters degree after their under-graduation education to get better job opportunities and salary. The dataset that is used in this project consists of Placement data of MBA students in a campus. It includes factors like secondary, higher secondary and degree school percentage and specialisation. It also includes Work Experience and salary of the placed students.

Linear Regression is a predictive analytics technique that uses data to predict the output variable. The idea behind this is that if a linear regression model can be fitted to some observed data, this model can be used to predict any future values. We assume the equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...$$

where Y= dependent/ predicted variable,

β_0 = intercept,

X_i =independent/ predictor variables, and

β_i = coefficients of X_i ; for i=1,2,3,...

Logistic Regression is a supervised machine learning classification algorithm used to predict the probability of a categorical dependent variable. Categorical, here, means that the dependent variable only takes values from a set of values. For this model, the dependent variable, Placement status only takes the values 0 and 1.

0: when the student is not placed (do not have a job), and 1: when the student is placed (have a job). So, this is a Binary Logistic Regression model. When there are three or more categories for dependent variable, that model is known as the Multinomial Regression model. Binary Logistic Regression is one of the most widely used ways to fit models for binary categorical data. It can directly predict probabilities, and it preserves the marginal probabilities of the training data. The independent/ predictor variable can be either categorical, or continuous; or a mix of both variables in one model.

OBJECTIVE

The objective of the analysis is to perform regression on the placement dataset and evaluate the connection between various factors that affect the placement status and salary, determine the factors that give the best performance for a continuous and a categorical dependent variable, verify the assumptions for the Multiple Linear Regression model and predict the continuous dependent variable, verify the assumptions for the Binary Logistic Regression method and fit the model.

METHODOLOGY

For this analysis, data is obtained from Kaggle. The dataset consists of the following:

- Columns:
    - sl_no (serial Number), ssc_p (Secondary Education Percentage), hsc_p (Higher Secondary Education Percentage), degree_p (Degree Percentage), etest_p (Employability Test Percentage, conducted by college), mba_p (MBA percentage), and salary (Salary offered by corporates to candidates) as *numeric columns*;
    - gender (Gender- Male='M', Female='F'), ssc_b (Board of Secondary Education- Central/ Others), hsc_b (Board of Higher Secondary Education- Central/ Others), hsc_s (Specialisation in Higher Secondary Education), degree_t (Under Graduation Degree type- Field of degree education), specialisation (Post Graduation- MBA Specialisation), and status (Status of Placement- Placed/ Not Placed) as *string columns*;
    - workex (Work Experience) as *boolean column*.
- Rows: Values consist of 215 rows.

Data is analyzed by using Python Programming Language as it has a lot of packages for regression modelling and visualisation. Various tools such as pandas, numpy, LinearRegression, LogisticRegression, r2_score, MinMaxScaler, mean_absolute_error, train_test_split, seaborn are used to aid in that process. Jupyter Notebook is very convenient and easy to use as it saves the output along with the data and a particular line of code can be executed at a time, we do not need to run the whole algorithm code each time we add or edit a line of code. String columns are converted to numeric by assigning the values of 0, 1 to the columns containing two parameters and 0,1,2 to the columns containing three parameters so that analysis can be performed on all the variables given in the dataset and that Logistic Regression can be applied on the model.

STATISTICAL ANALYSIS

1. PREDICTION OF SALARY AND ASSUMPTIONS OF MULTIPLE LINEAR REGRESSION:

To predict the salary as a dependent variable, the data is cleaned (Pre- Processed). All the factors that contained values as strings are changed to numeric values by converting values to 0, 1 where the factors had two distinct values; and to 0, 1, 2 where the factors had three distinct values. The raw data also had data for students who did not get placed on- campus recruitment and so their salary column contained null values. Those values are changed to 0 for this analysis since their salary is 0 if they do not have a placement. Placement status column was also dropped as it would have introduced bias in the analysis. Y (Salary) is taken as the dependent variable and X_i (all the other factors) as independent variables.

Now, as this data contains a lot of factors, determining the number of features required for best performance becomes very important as if we work with all the factors given, over- fitting of data may happen. Over- fitting, as the name suggests, is the analytic quality of the distribution that happens when there are a lot of independent variables and the fitted variables generated correspond too closely or the model learns to train the data too well. When this happens, the model starts learning the noise in the dataset as well. Also, it becomes somewhat difficult to explain the model if there are a lot of independent variables. Feature Selection is done by two methods: Determining the Least Significant variable by R^2 Score, and Eliminating the Least Significant variable by p- value.

➢ **R^2 score** is the coefficient of determination. It is a statistical measure of how close the theoretical data is to the fitted data. It ranges between 0 and 1, where 0 indicates that the model does not at all explain the variability of the dependent variable that is predicted from the independent variable, and as the value of R^2 increases, the variability of the predicted dependent variable from the independent variable increases, taking its highest value at 1. It provides a measure of how well the observed outcomes are predicted by the model, based on the portions of total variation of outcomes explained by the model. So, for this analysis, the least significant variables will be determined which have the lowest values of R^2, and the remaining variables will be taken as the best performance factors for this model.

Sequential Backward Elimination is used to determine the number of factors for the best performance of the model for salary as the dependent variable. The graph of this technique plotted by importing Sequential Feature Selection from mlxtend shows that six factors/ independent variables will give the best performance in this model.

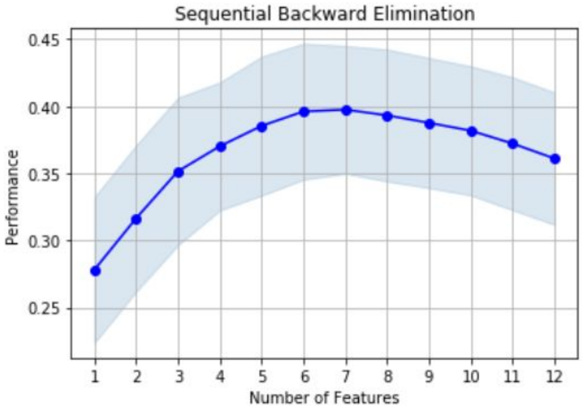


Figure 1.1

The six most significant features found on running the algorithm are:

a. ‘Experience?’: if the student has work experience
b. ‘Male?’: the gender of the student,
c. ‘mba_p’: the percentage scored in MBA degree by the student, and
d. ‘degree_p’: the percentage scored in under graduation degree by the student
e. ‘hsc_p’: the percentage scored in higher secondary education by the student
f. ‘ssc_p’: the percentage scored in secondary education by the student

➢ **P- value** (probability value) is the probability of getting the predicted values very close to the theoretical values when the null hypothesis is assumed to be correct. When the p- value is less than a previously determined significance value, then the null hypothesis is rejected, and if it is more than that significance value, then the null hypothesis is accepted. For this model, the null hypothesis,

$\mathbf{H_0}$ is taken that there is no significant difference between the samples of R^2 scores, and the alternate hypothesis,

$\mathbf{H_1}$ is that there is significant difference between the samples of R^2 scores, where the different R^2 scores are obtained by taking the R^2 value of the fitted OLS (ordinary least square) base model first, then find out the factor with the highest p-value while looking at the summary of it, and drop that factor from the set of independent variables. Again take the R^2 value of the fitted OLS model after the elimination of the factor, find the factor with the highest p- value, drop it, and continue with this process until the p- values of leftover variables is less than the previously determined significance value.

For this model, the significance value is taken to be 0.05. If the calculated p-value comes out to be less than 0.05, the null hypothesis is rejected, and we consider the alternative hypothesis to be correct. Here, a small p- value indicates that there is a rare chance of observing a relationship between the independent/ predictor and dependent/ predicted variables just due to chance.

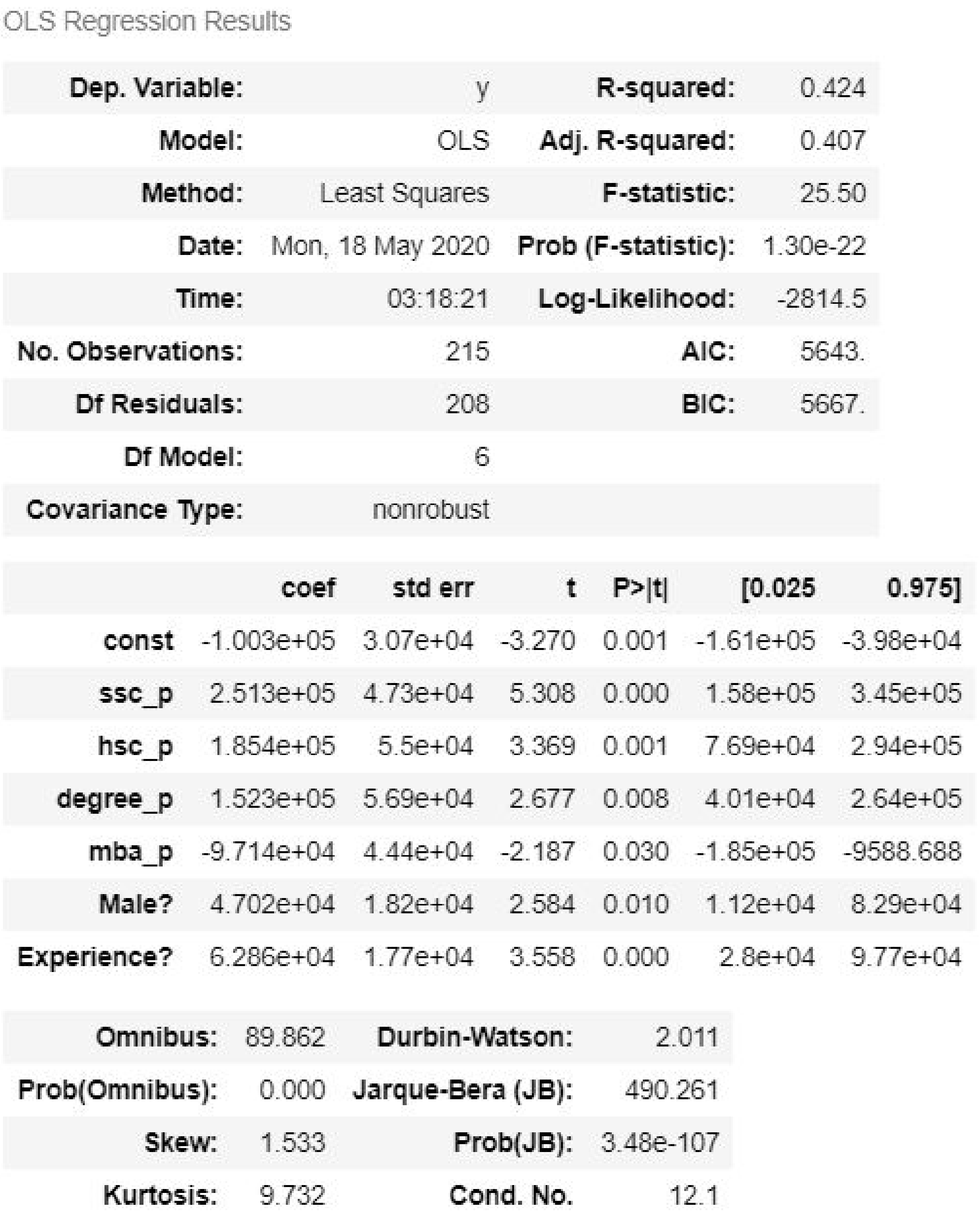
On the execution of the base model, the highest p- value (0.906) was found to be of the feature 'etest_p' so that feature is dropped from the model. Then 'hsc_s' is dropped with a p- value of 0.701 from the resulting model. Continuing this way, seven features are left: 'ssc_p' with a p- value of 0.000, 'hsc_p' with a p- value of 0.002, 'degree_p' with a p- value of 0.016, 'mba_p' with a p- value of 0.029, 'Male?' with a p- value of 0.019, 'SpecialisationInHR?' with a p- value of 0.087 and 'Experience?' with a p- value of 0.001. Since the 'SpecialisationInHR' feature still has a p- value more than 0.05, it is eliminated from the model.

Thus, the most significant two features found by elimination of the least significant variable by p- value method are:

a. 'Experience?': if the student has work experience
b. 'Male?': the gender of the student,
c. 'mba_p': the percentage scored in MBA degree by the student, and
d. 'degree_p': the percentage scored in under graduation degree by the student
e. 'hsc_p': the percentage scored in higher secondary education by the student
f. 'ssc_p': the percentage scored in secondary education by the student

which were also the features that were obtained by determination of the least significant variable by R^2 score method.

The summary of the fitted OLS model obtained is given in table 1.1.



OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.424
Model:	OLS	**Adj. R-squared:**	0.407
Method:	Least Squares	**F-statistic:**	25.50
Date:	Mon, 18 May 2020	**Prob (F-statistic):**	1.30e-22
Time:	03:18:21	**Log-Likelihood:**	-2814.5
No. Observations:	215	**AIC:**	5643.
Df Residuals:	208	**BIC:**	5667.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-1.003e+05	3.07e+04	-3.270	0.001	-1.61e+05	-3.98e+04
ssc_p	2.513e+05	4.73e+04	5.308	0.000	1.58e+05	3.45e+05
hsc_p	1.854e+05	5.5e+04	3.369	0.001	7.69e+04	2.94e+05
degree_p	1.523e+05	5.69e+04	2.677	0.008	4.01e+04	2.64e+05
mba_p	-9.714e+04	4.44e+04	-2.187	0.030	-1.85e+05	-9588.688
Male?	4.702e+04	1.82e+04	2.584	0.010	1.12e+04	8.29e+04
Experience?	6.286e+04	1.77e+04	3.558	0.000	2.8e+04	9.77e+04

Omnibus:	89.862	**Durbin-Watson:**	2.011
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	490.261
Skew:	1.533	**Prob(JB):**	3.48e-107
Kurtosis:	9.732	**Cond. No.**	12.1

Table 1.1

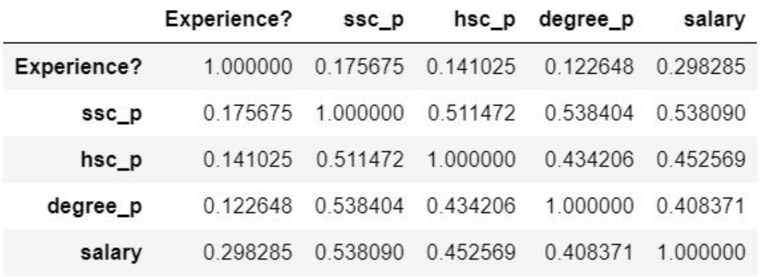
Assumptions of Multiple Linear Regression:

A. MULTICOLLINEARITY:

Multicollinearity is a phenomenon which occurs when one or more independent/ predictor variable(s) can be predicted from the other independent/ predictor variables with some accuracy. In such a situation, a small change in the model or the data may cause significant changes in the coefficient estimates of the model. It does not degrade the predictive power or the reliability of the model as a whole, but it affects the individual predictor calculations. When ‘no multicollinearity’ is used in the text, it means that perfect multicollinearity is absent there, not that there is absolutely no correlation between them. Two central criteria are used to test for multicollinearity in this model: Correlation Matrix and VIF (Variation Inflation Factor).

- Correlation matrix is a table that shows correlation coefficients between the variables in the model. Each variable is correlated with each of the other variables. Correlation coefficient varies between -1 and 1. The variables which have 0 correlation between them are totally uncorrelated, whereas the variables having a correlation of -1 and 1 represent strong negative and positive correlation respectively.

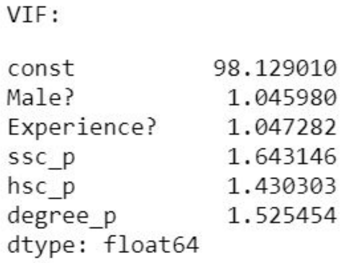
  In the model, on inspection, it was found that mba_p and gender does not correlate substantially with the dependent variable, so it was dropped from the model and the correlation coefficients between the remaining variables are given in the table below. In this case, all the independent variables correlate substantially (more than 0.29) with the dependent variable, ‘salary’, ‘ssc_p’ with the highest correlation coefficient of 0.538 (approx.). Bivariate correlation between the independent variables is less than 0.538, therefore all the variables are retained.



	Experience?	ssc_p	hsc_p	degree_p	salary
Experience?	1.000000	0.175675	0.141025	0.122648	0.298285
ssc_p	0.175675	1.000000	0.511472	0.538404	0.538090
hsc_p	0.141025	0.511472	1.000000	0.434206	0.452569
degree_p	0.122648	0.538404	0.434206	1.000000	0.408371
salary	0.298285	0.538090	0.452569	0.408371	1.000000

Table 1.2

- VIF (Variation Inflation Factor) is quotient of the variance in a model with multiple terms by the variance of a model with one term. It quantifies how much the variance of a regression coefficient changes due to multicollinearity in the model. If VIF is more than 10, then multicollinearity is taken to be high. Variance Inflation Factor for the independent variables is given in the table below. It can be seen that all the values are very less than 10, which again, indicates the absence of multicollinearity.



```
VIF:

const          98.129010
Male?           1.045980
Experience?     1.047282
ssc_p           1.643146
hsc_p           1.430303
degree_p        1.525454
dtype: float64
```

Table 1.3

B. NORMALITY:

Normality is the random error in the relation between the dependent and independent variables in a model. Every case in the sample has a different random variable that encompasses the "noise" which accounts for the differences in the theoretical and predicted values produced by the regression equation; and the distribution of this variable should be normally distributed. For this model, histogram, probability-probability (P-P) plot, and quantile- quantile (Q-Q) plot are plotted to check for normality of the residuals.

- Histogram is used to determine the shape and spread of the data. It is a frequency plot that is obtained by plotting the data frequency with the center of the data value.

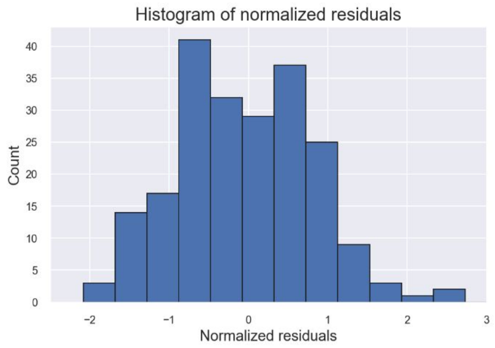


Figure 1.2

Figure 1.3 shows an approximately normal distribution of residuals from the model since the distribution is bell- shaped and symmetric.

- P-P Plot is a graphical technique that is used to compare the observed cumulative distribution function of the residual to the expected cumulative distribution function of the normal distribution, since checking for the normality of the residuals is needed, not of the predictors. The plotted data should form approximately a straight line. The plot does not show any strong deviations from the straight line, so the assumption is met.

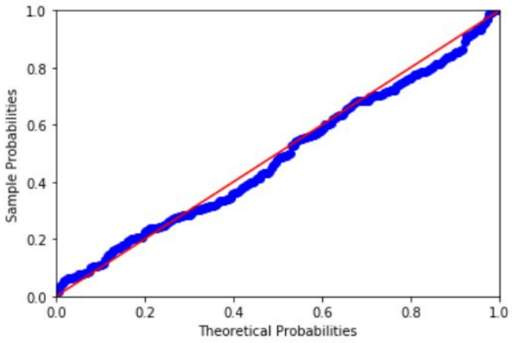


Figure 1.3

- Q-Q Plot compares the distribution of the data to a normal distribution by plotting the quartiles of the residual data versus the quartiles of a normal distribution. It is used to assess if the residuals are normally distributed. The following plot does not show any strong deviations from the straight line so the assumption is met.

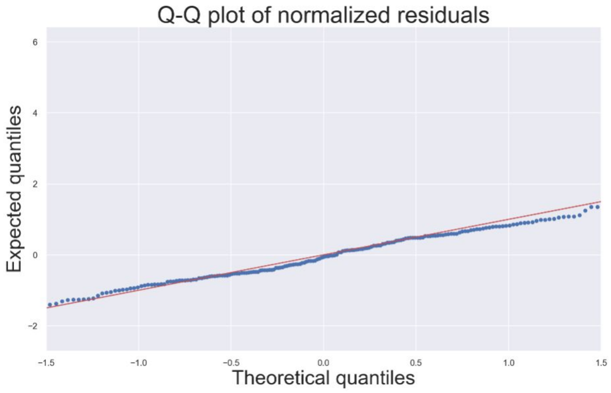


Figure 1.4

C. HOMOSCEDASTICITY:

The assumption of homoscedasticity also applies on the residuals of the model and can be tested with a scatterplot of the residuals. It means that the errors exhibit a constant variance, which is a key assumption of a linear regression model. The assumption of homoscedasticity is valid when the noise of the model is random and do not follow a pattern.

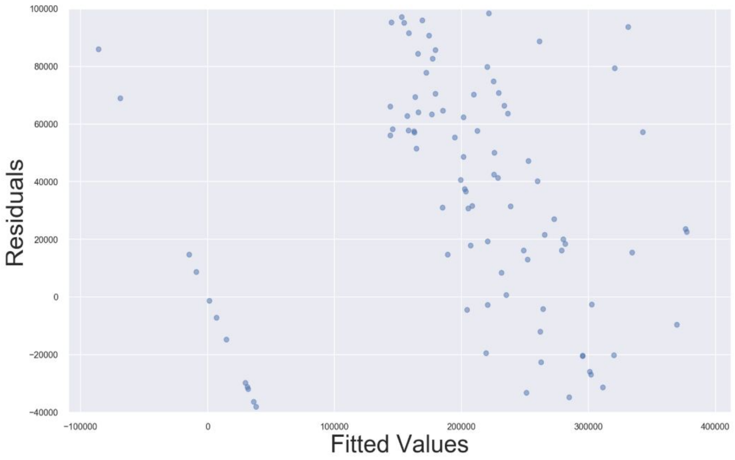


Figure 1.5

D. LINEARITY:

The dependent variable should be a linear function of the predictor variables/ features in the model. The plot of dependent versus independent variables should be a symmetrical distribution of the points around a diagonal line. Figure 1.7 shows that this assumption has been met.

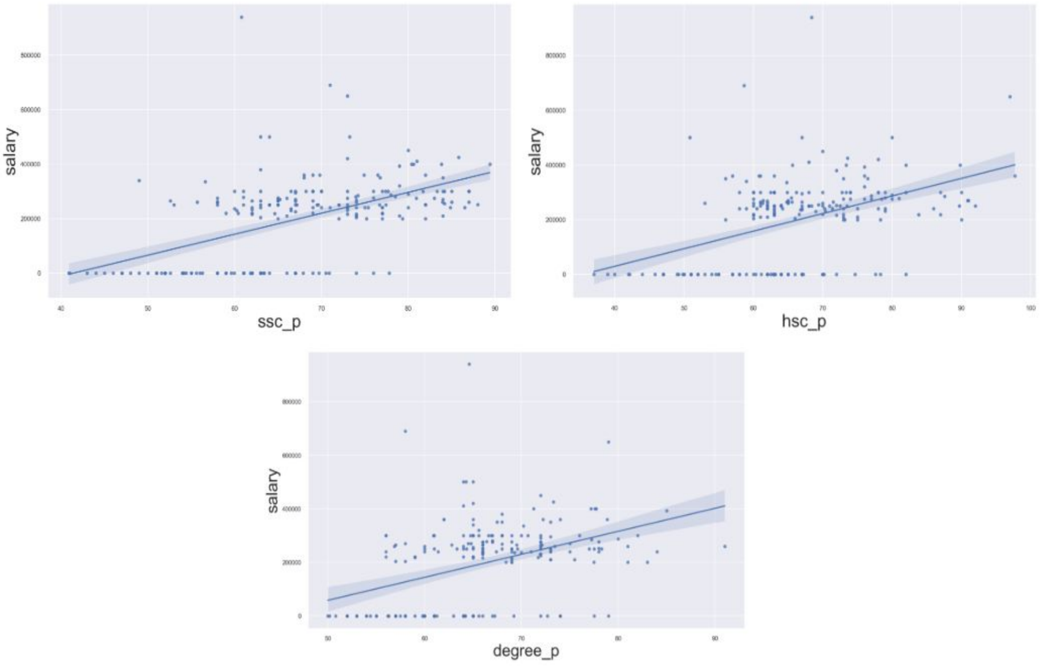
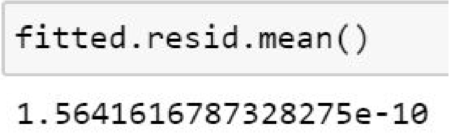


Figure 1.6

Work Experience versus salary is not plotted as it is a categorical variable.

E. EXPECTATION OF RESIDUALS IS ZERO:

The expectation (mean) of residuals (difference between actual value and predicted value) should be equal to zero. Since this value has come out to be very small (close to zero), this assumption is satisfied.



```
fitted.resid.mean()

1.5641616787328275e-10
```

Figure 1.7

F. VALUES OF RESIDUALS ARE UNCORRELATED:

To test this assumption, Durbin- Watson test statistic is used. It is a measure of correlation in residuals of a model. The value of this test ranges between 0 and 4. For the assumption to be valid, the value should be close to 2. In this case, the value is equal to 2.011, as can be verified from Table 1.1, so the assumption is met.

G. INFLUENTIAL DATA POINTS:

These are observations that bring into play an unexpectedly large effect on the outcome of the regression analysis. They can be classified as outliers. Cook's Distance values are observed to test the presence of Influential Data Points. Values larger than one are a problem. Since no value is higher than 0.16, there is no major problem.

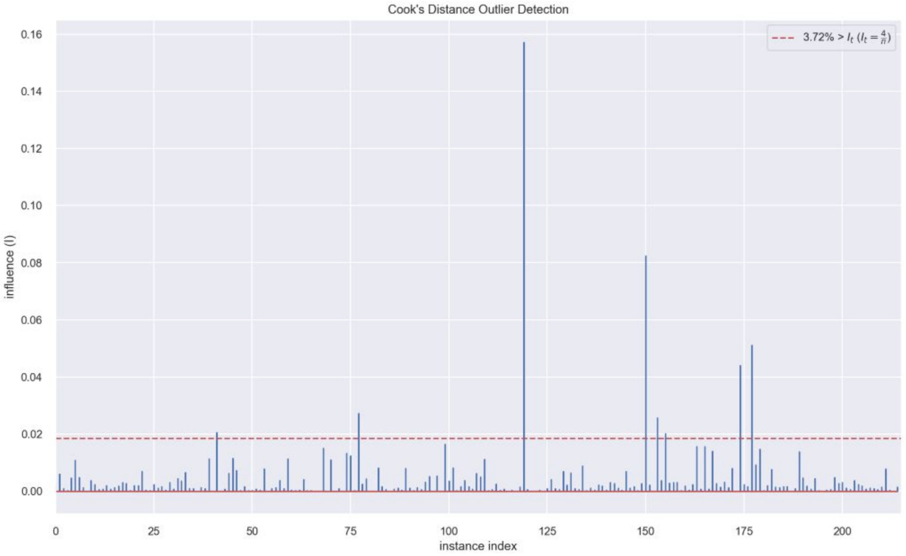


Figure 1.8

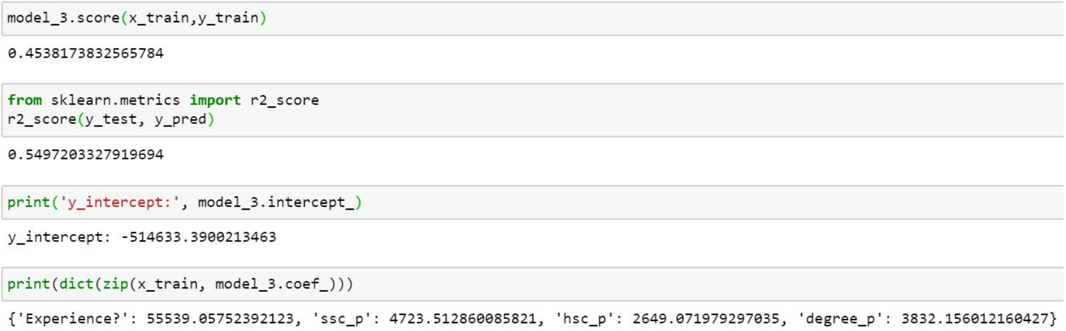
Multiple Regression Model:

$$Salary = \beta_0 + \beta_1 \times work\ ex + \beta_2 \times secondary\ percentage + \beta_3 \times higher\ secondary\ percentage + \beta_4 \times degree\ percentage$$

where $\beta_0,\ \beta_1,\ \beta_2,\ \beta_3,\ and\ \beta_4$ are Regression Coefficients.

In order to estimate the parameters β_i, for i= 0,1,2,... the model needs to be fitted to a set of data. To achieve that, the OLS (Ordinary Least Squares) method is used in the case of Multiple Linear Regression. OLS is a linear least squares method which minimises the sum of the square of the differences, also known as the principle of least squares, between the theoretical and predicted values of the dependent variable in the data.

After importing LinearRegression and train_test_split from sklearn, train and test variables of x and y are made with a test size of 0.15 (training size= 0.85) to predict the value of y.



```
model_3.score(x_train,y_train)

0.4538173832565784

from sklearn.metrics import r2_score
r2_score(y_test, y_pred)

0.5497203327919694

print('y_intercept:', model_3.intercept_)

y_intercept: -514633.3900213463

print(dict(zip(x_train, model_3.coef_)))

{'Experience?': 55539.05752392123, 'ssc_p': 4723.512860085821, 'hsc_p': 2649.071979297035, 'degree_p': 3832.156012160427}
```

Figure 1.9

The accuracy of the multiple linear regression model on the test set has come out to be 45.382% and the value of R^2 is 0.549. The values of β_0, β_1, β_2, β_3, and β_4 are -514644.39, 55555.06, 4723.513, 2649.072, 2649.072, and 3832.156 respectively.

2. PREDICTION OF PLACEMENT STATUS AND ASSUMPTIONS OF BINARY LOGISTIC REGRESSION:

Now the analysis is conducted by taking Y (Placement status) as the dependent variable and X_i (all the other factors) as independent variables. Since in this case, Y consists of binary categorical data, Binary Logistic Regression will be used. The pre-processed data is used for this analysis, and the salary column is dropped as it would have introduced bias in the analysis.

Now, as this data contains a lot of factors, the number of features required for best performance are determined by two methods: Determining the Least Significant variable by R^2 Score, and Eliminating the Least Significant variable by p- value.

➢ **Determining the Least Significant variable by R^2 score:** Sequential Backward Elimination is used to determine the number of factors for the best performance of the model for placement status as the dependent variable. The graph of this technique plotted by importing Sequential Feature Selection from mlxtend shows that five factors/ independent variables will give the best performance in this model.

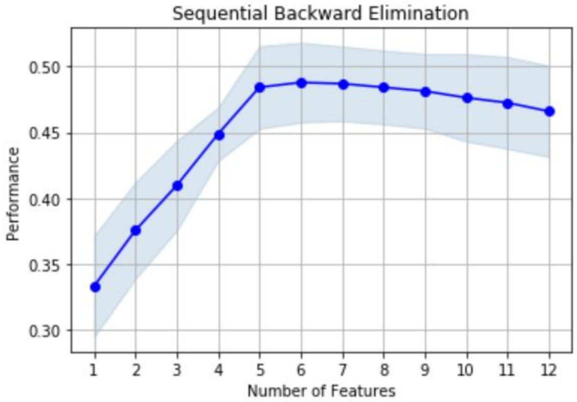


Figure 2.1

The most significant five features found on running the algorithm are:

a. '<u>Experience?</u>': if the student has work experience or not

b. '<u>mba_p</u>': the percentage scored in MBA degree by the student

c. '<u>degree_p</u>': the percentage scored in under graduation degree by the student

d. '<u>hsc_p</u>': the percentage scored in high secondary education by the student

e. '<u>ssc_p</u>': the percentage scored in secondary education by the student

➢ **Eliminating the Least Significant variable by p- value:** For this model, the null hypothesis,

H_0 is taken that there is no significant difference between the samples of R^2 scores, and the alternate hypothesis,

H_1 is that there is significant difference between the samples of R^2 scores,

where the different R^2 scores are obtained by taking the R^2 value of the fitted OLS (ordinary least square) base model first, then find out the factor with the highest p- value while looking at the summary of it, and drop that factor from the set of independent variables. Again take the R^2 value of the fitted OLS model after the elimination of the factor, find the factor with the highest p- value, drop it, and continue with this process until the p- values of leftover variables is less than the previously determined significance value.

For this model, the significance value is taken to be 0.01. If the calculated p- value comes out to be less than 0.01, the null hypothesis is rejected, and we consider the alternative hypothesis to be correct. Here, a small p- value indicates that there is a rare chance of observing a relationship between the independent/ predictor and dependent/ predicted variables just due to chance.

On the execution of the base model, the highest p- value (0.740) was found to be of the feature 'hsc_s' so that feature is dropped from the model. Then 'HSCentralBoard?' is dropped with a p- value of 0.443 from the resulting model. Continuing this way, seven features are left: 'Experience?' with a p- value of 0.000, 'ssc_p' with a p- value of 0.000, 'mba_p' with a p- value of 0.000, 'degree_p' with a p- value of 0.000, 'hsc_p' with a p- value of 0.001, 'Male?' with a p- value of 0.117, and 'degree_t' with a p- value of 0.026. Since a lot of these features still have a p- value more than 0.01, the elimination process is continued, until only 'ssc_p', 'hsc_p', 'degree_p', 'mba_p' and 'Experience?' features are left with p- values of 0.000, 0.000, 0.000, 0.000 and 0.000 respectively.

Thus, the most significant two features found by elimination of the least significant variable by p- value method are:

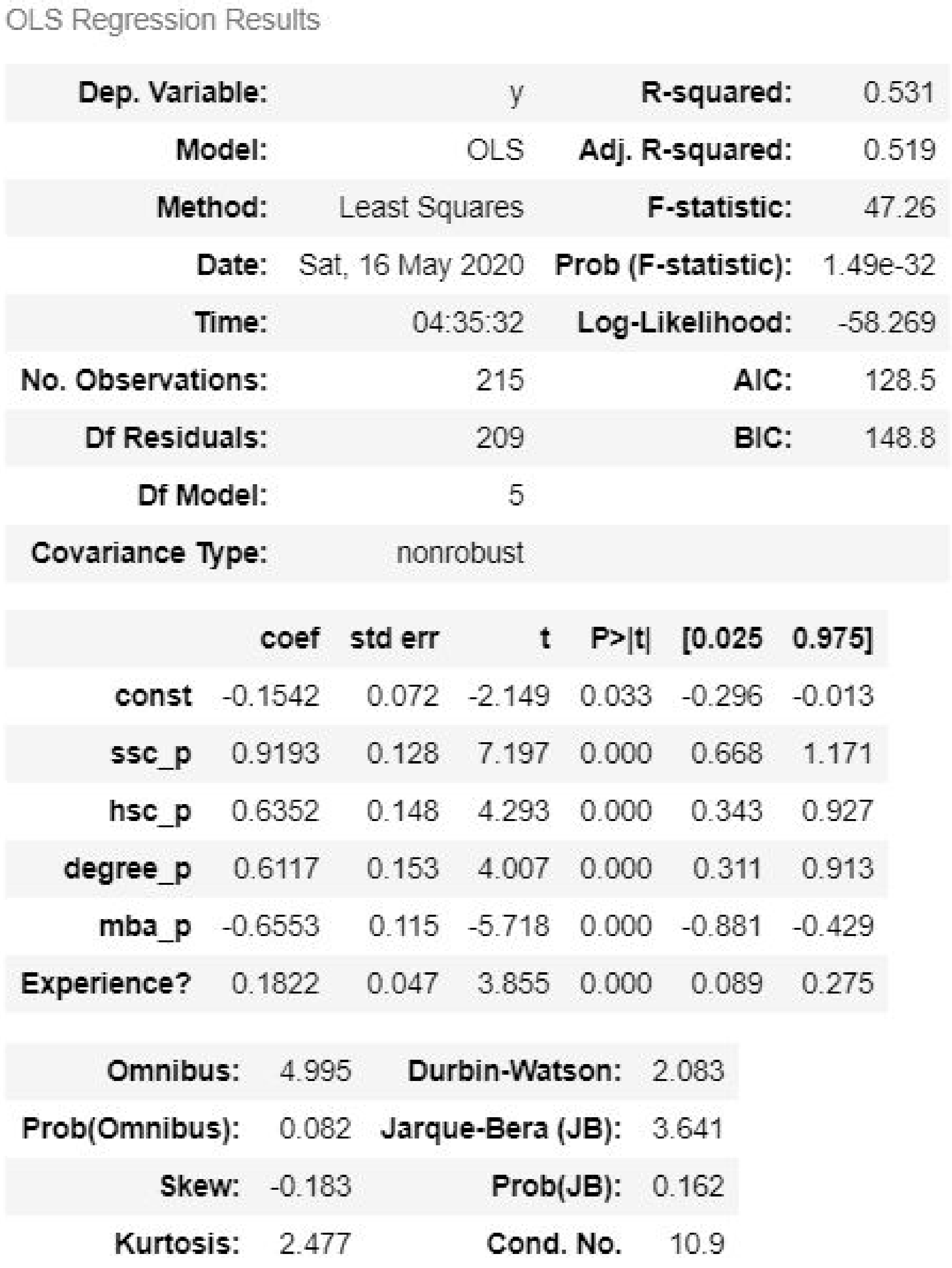
a. 'Experience?': if the student has work experience or not
b. 'mba_p': the percentage scored in MBA degree by the student
c. 'degree_p': the percentage scored in under graduation degree by the student

d. ‘hsc_p’: the percentage scored in high secondary education by the student

e. ‘ssc_p’: the percentage scored in secondary education by the student

which were also the features that were obtained by determination of the least significant variable by R^2 score method. So, these features will be used as the independent variables for the binary logistic regression analysis of the data.

The summary of the fitted OLS model obtained is given in the below table:



OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.531
Model:	OLS	**Adj. R-squared:**	0.519
Method:	Least Squares	**F-statistic:**	47.26
Date:	Sat, 16 May 2020	**Prob (F-statistic):**	1.49e-32
Time:	04:35:32	**Log-Likelihood:**	-58.269
No. Observations:	215	**AIC:**	128.5
Df Residuals:	209	**BIC:**	148.8
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1542	0.072	-2.149	0.033	-0.296	-0.013
ssc_p	0.9193	0.128	7.197	0.000	0.668	1.171
hsc_p	0.6352	0.148	4.293	0.000	0.343	0.927
degree_p	0.6117	0.153	4.007	0.000	0.311	0.913
mba_p	-0.6553	0.115	-5.718	0.000	-0.881	-0.429
Experience?	0.1822	0.047	3.855	0.000	0.089	0.275

Omnibus:	4.995	**Durbin-Watson:**	2.083
Prob(Omnibus):	0.082	**Jarque-Bera (JB):**	3.641
Skew:	-0.183	**Prob(JB):**	0.162
Kurtosis:	2.477	**Cond. No.**	10.9

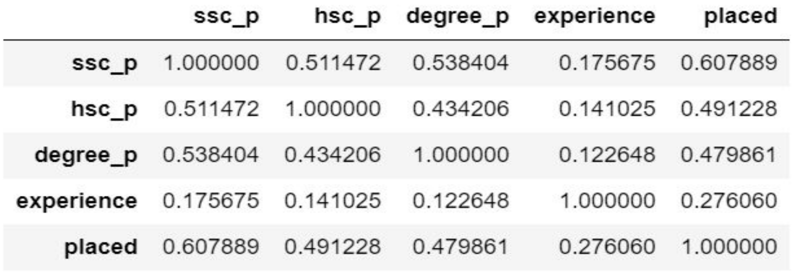
Table 2.1

Assumptions of Multiple Logistic Regression:

A. MULTICOLLINEARITY:

Two central criteria are used to test for multicollinearity in this model: Correlation Matrix and VIF (Variation Inflation Factor).

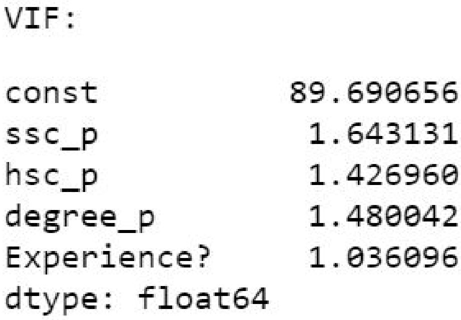
- Correlation matrix: In the model, on inspection, it was found that mba_p does not correlate substantially with the dependent variable, so it was dropped from the model and the correlation coefficients between the remaining variables are given in the table below. In this case, all the independent variables correlate substantially (more than 0.27) with the dependent variable, ‘placed’, ‘ssc_p’ with the highest correlation coefficient of 0.608 (approx.). Bivariate correlation between the independent variables is less than 0.608, therefore all the variables are retained.



	ssc_p	hsc_p	degree_p	experience	placed
ssc_p	1.000000	0.511472	0.538404	0.175675	0.607889
hsc_p	0.511472	1.000000	0.434206	0.141025	0.491228
degree_p	0.538404	0.434206	1.000000	0.122648	0.479861
experience	0.175675	0.141025	0.122648	1.000000	0.276060
placed	0.607889	0.491228	0.479861	0.276060	1.000000

Table 2.2

- VIF (Variation Inflation Factor): If VIF is more than 10, then multicollinearity is taken to be high. Variance Inflation Factor for the independent variables is given in the table below. It can be seen that all the values are very less than 10, which again, indicates the absence of multicollinearity.



```
VIF:

const          89.690656
ssc_p           1.643131
hsc_p           1.426960
degree_p        1.480042
Experience?     1.036096
dtype: float64
```

Table 2.3

B. OUTLIERS: The assumption of the lack of outliers can be checked by using a box plot. Separate box plots can be produced for each independent variable.

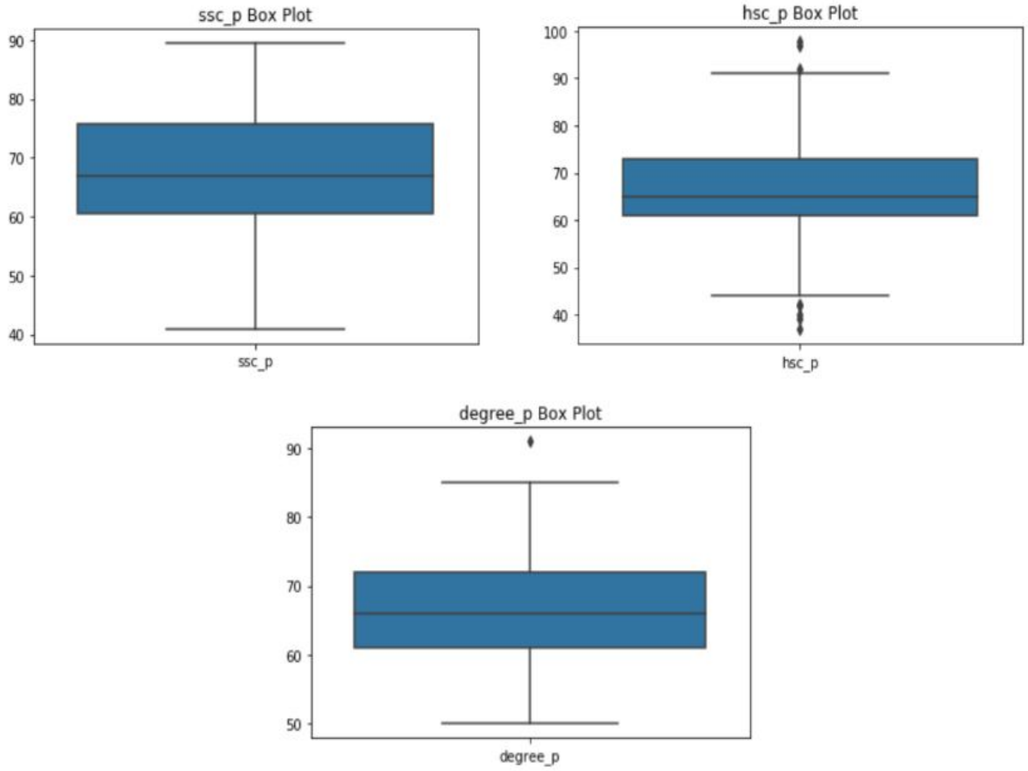


Figure 2.2

As it can be noticed from Figure 3.2, there are no outliers in ‘ssc_p’, a few outliers in ‘hsc_p’, and one outlier in ‘degree_p’. The outliers that are present in ‘hsc_p’ and ‘degree_p’ lie very close to the rest of the values, thus, the values can be kept and used in analysis. There is no use of plotting a box plot of independent variable ‘experience’ as it only takes the value of 0 and 1, so there can not be any outlier present in that variable.

C. CONTINUOUS INDEPENDENT VARIABLES ARE LINEARLY RELATED TO THE LOG- ODDS: Log- odds (or the logit) means the logarithm of the odds. If p is the probability, then $\frac{p}{1-p}$ is the corresponding odds; hence the log- odds is

$$log(\frac{p}{1-p}) = log(p) - log(1-p) = \beta_0 + x_1 \times \beta_1 + x_2 \times \beta_2 + ...$$

A requirement of logistic regression is for the continuous independent variable to be linearly related to the log- odds of the independent variable. To do this, a regression plot is constructed using the ‘seaborn’ package. The plot should look like an S-shaped curve. Since in the data, 'ssc_p', 'hsc_p', and 'degree_p' are continuous variables, their plots are constructed and the results seem to be promising.

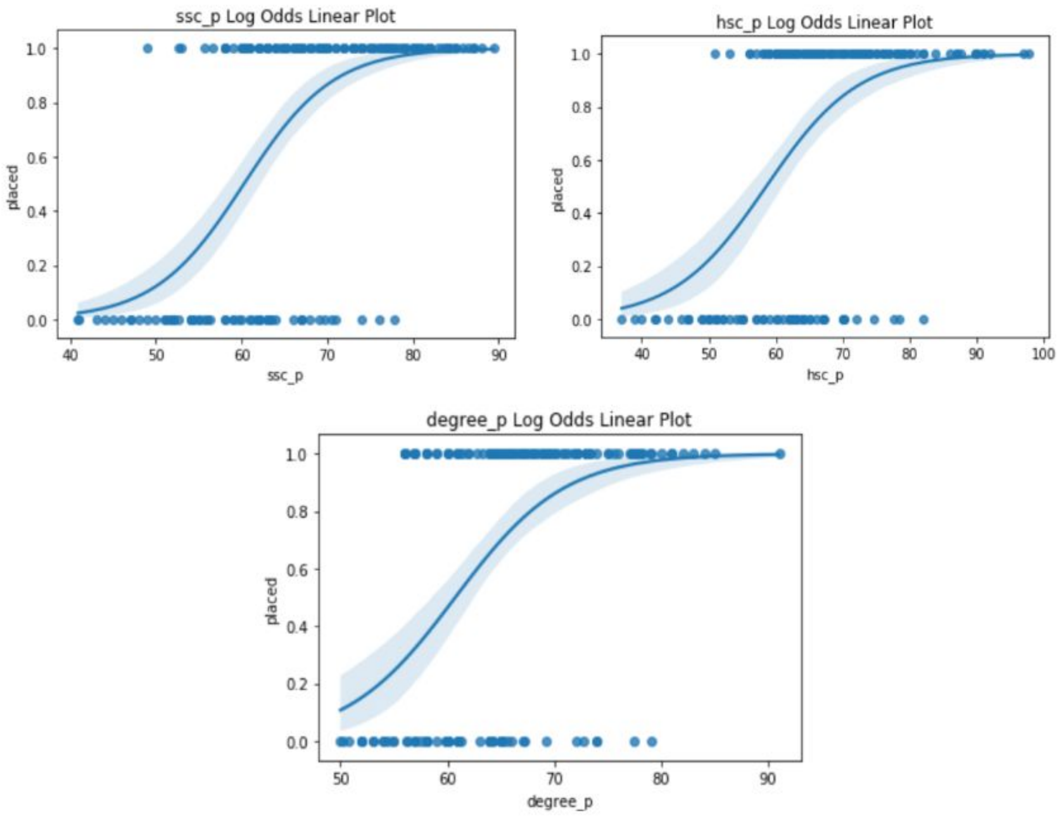
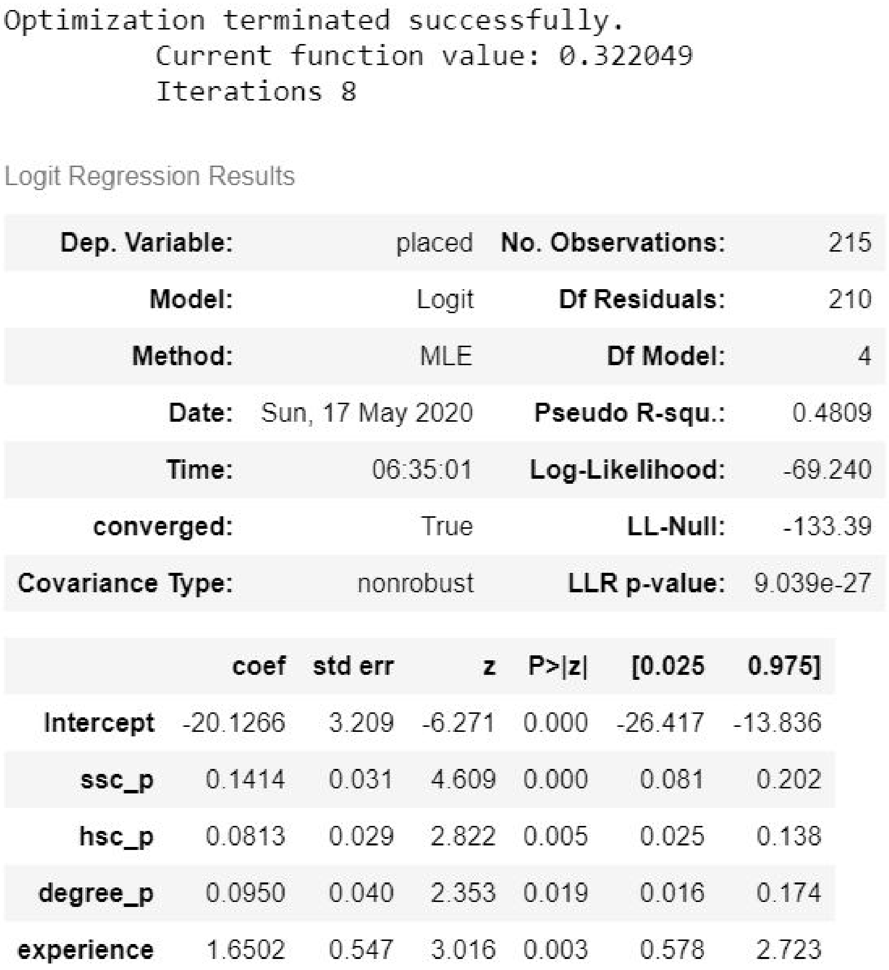


Figure 2.3

Also, we know that the dependent variable is categorical with two values as categories, so it is a binary logistic regression model. The sample size is also large enough (215 values). So all the assumptions of Multiple Logistic Regression are met and the data is good to run.

After implementation of the fitted logit model, summary is obtained as shown in Table 2.4:



```
Optimization terminated successfully.
         Current function value: 0.322049
         Iterations 8
```

Logit Regression Results

Dep. Variable:	placed	**No. Observations:**	215
Model:	Logit	**Df Residuals:**	210
Method:	MLE	**Df Model:**	4
Date:	Sun, 17 May 2020	**Pseudo R-squ.:**	0.4809
Time:	06:35:01	**Log-Likelihood:**	-69.240
converged:	True	**LL-Null:**	-133.39
Covariance Type:	nonrobust	**LLR p-value:**	9.039e-27

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	-20.1266	3.209	-6.271	0.000	-26.417	-13.836
ssc_p	0.1414	0.031	4.609	0.000	0.081	0.202
hsc_p	0.0813	0.029	2.822	0.005	0.025	0.138
degree_p	0.0950	0.040	2.353	0.019	0.016	0.174
experience	1.6502	0.547	3.016	0.003	0.578	2.723

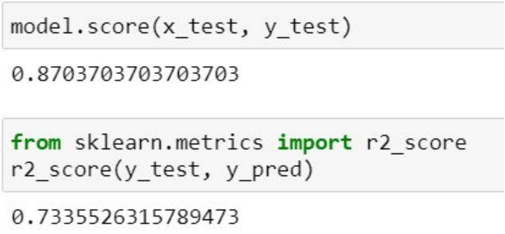
Table 2.4

It can be inferred from this table that since the p- value of all the predictor variables is less than 0.025, all the independent/ predictor variables have a significant effect on the log odds of being placed. For every one unit increase in the percentage of secondary education, higher secondary education, and under graduation degree, the log odds of placement status increases by 0.1414, 0.0813, and 0.0950 respectively. Since experiment is a categorical variable, it affects the placement status differently and its coefficient is given by 1.6502. Also, the intercept of the model is equal to -20.1266.

Logistic Regression Model Fitting:

In order to estimate the parameters β_i , for i= 0,1,2,... the model needs to be fitted to a set of data. To achieve that, the MLE (Maximum Likelihood Estimator) method is used in the case of Logistic Regression. In this method, an iterative process is used, starting with the tentative solution, then some revision is made so as to improve it, then keep on repeating this process until no more improvement can be made, at which point, the process is said to have converged.

As can be seen from table 3.4, the model has converged. After importing LogisticRegression and train_test_split from sklearn, train and test variables of x and y are made with a test size of 0.25 (training size= 0.75) to predict the value of y. The accuracy of the logistic regression classifier on the test set has come out to be 87.037% and the value of R^2 is 0.733.



```
model.score(x_test, y_test)
```

```
0.8703703703703703
```

```
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

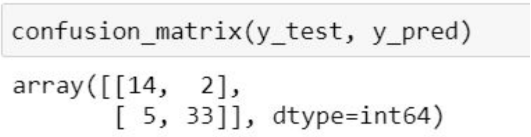
```
0.7335526315789473
```

Figure 2.4

Confusion Matrix:

A Confusion Matrix (error matrix) is a performance measurement for classification of a model on some of the test data for which the theoretical values are known. It is a tabular representation of Predicted vs Actual values. It helps in finding out how accurate the model is, and avoiding over- fitting. The number of right and wrong predictions are summarised with count values, and broken down by each class. The Confusion Matrix is used to compute most performance measures.

The Confusion Matrix from the model is given by:



```
confusion_matrix(y_test, y_pred)
```

```
array([[14,  2],
       [ 5, 33]], dtype=int64)
```

Table 2.5

From the above Confusion matrix, it can be concluded that 14+33 values were predicted correctly, and 2+5 values were predicted incorrectly, where

- True Positive: 14 (Positive result was predicted to be positive)
- True Negative: 33 (Negative result was predicted to be negative)
- False Positive: 2 (Negative result was predicted to be positive)
- False Negative: 5 (Positive result was predicted to be negative)

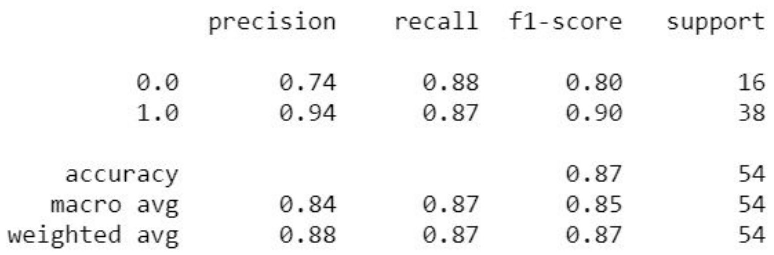
Computation of Precision, Recall, F- measure and Support:

Precision is the ratio of True Positive with the sum of True Positive and False Positive. It is the ability of the classifier to not label the sample as positive when it is negative.

Recall is the ratio of True Positive with the sum of True Positive and False Negative. It is the ability of the classifier to find positive samples.

The F- Beta score is the weighted harmonic mean of precision and recall. It ranges between 0 and 1 where 0 is its worst score and 1 is its best score.

Support is the frequency of each class in y_test. The table is given as follows:



```
              precision    recall  f1-score   support

         0.0       0.74      0.88      0.80        16
         1.0       0.94      0.87      0.90        38

    accuracy                           0.87        54
   macro avg       0.84      0.87      0.85        54
weighted avg       0.88      0.87      0.87        54
```

Table 2.6

From the table, we can conclude that 87% of the placed students were accurately classified by the model.

ROC Curve:

The ROC (Receiver Operating Characteristic) Curve is a common tool that is used in Binary Classifiers. It is plotted between the False Positive rate (1- specitivity) and the True Positive rate (sensitivity), where specificity is the ratio of True Negative with the sum of True Negative and False Positive, and sensitivity is the ratio of True Positive with the sum of True Positive and False Negative.

The dotted line in Figure 3.5 represents the curve of a purely random classifier. A good classifier stays towards the top- left corner, as far away from that line as possible, and as it can be noticed that the curve obtained in the model is very far from that line. So, it is a good classifier. The area under the curve (AUC) is referred to as the index of accuracy or the concordance index, where higher area represents better prediction power. It is equal to 0.72 for this model, which is a good measure of accuracy as well.

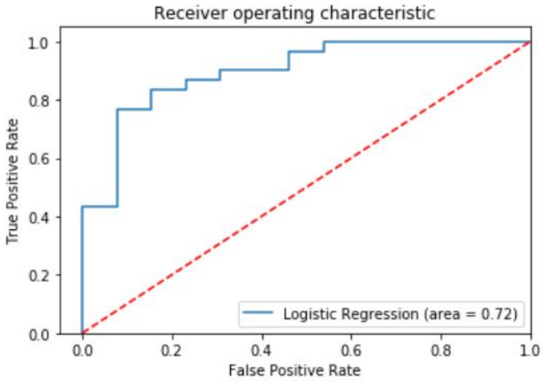


Figure 2.5

DISCUSSION AND CONCLUSIONS

The standard multiple linear regression method is used to predict the dependent variable (salary) using Python Programming Language. For the prediction of Salary using Multiple Linear Regression, the four most significant factors that affect the salary are the secondary education percentage, the higher secondary education percentage, the undergraduate degree percentage, and work experience of the student. The model thus obtained is:

$$Salary = \beta_0 + \beta_1 \times sscp + \beta_2 \times hscp + \beta_3 \times degreep + \beta_4 \times workex$$

All the assumptions of Multiple Linear Regression are met. Linear Regression is a good method for the prediction of a continuous dependent variable. It uses Ordinary Least Squares (OLS) method for estimating the required parameter which is powerful even at small sample sizes.

Binary Logistic Regression is used to analyze the model to demonstrate placement status (binary categorical variable). For this method, the four most significant factors that affect the placement status are work experience, the under graduation degree percentage, the High Secondary Education percentage, and the Secondary Education percentage.

The assumptions of Binary Logistic Regression model, such as the presence of binary dependent variable, no Multi- collinearity, linear relation of the independent variables to the log- odds, and no significant outliers are all met. Logistic Regression can be thought of as a special case of linear regression, where the outcome variable is categorical, with log of odds as the dependent variable. Large Sample sizes are needed for this model as it makes the use of Maximum Likelihood Estimates and MLE are less powerful at small sample sizes.