Business Goal

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- Step 1 : Importing required Libraries
- Step 2: Reading and Understanding the Data

-A few of the variable are null and therefore will required further analysis for cleaning.

- Step 3: Cleansing the data for EDA

-Looking at the data we have some columns like Specialization, Lead Profile, City etc with values as 'Select' where we dont have the info, there are as good as null, so replacing them with Nulls.

- Step 4: Performing numerical and categorical analysis on the data

-Plotting a pairplot for 3 Numerical variables to analysize the trend with Target Converted i.e. Number of leads converted

- Numerical Inferences:
- Total Visits : Converted customers have a fewer number of visits compared to non-converted customers between 0-50 visits. High visiting customers also seem to be less likely to convert.
- Total Time Spent on Website : Converted customers are likely to spend between 1000 to 1500 minutes on the website. This means converted customers have an higher average time spend per session on the website. Non-converted customers only spend between 0-500 minutes.
- Page Views Per Visit : Both converted and non-converted customers have the same number of website visits.

- Categorical Inferences
- Lead Origin : Leads originated through 'add forms ' has a high conversion rate compared to other Origins, followed by Page submissions and API
- Lead Source : Leads generated from welingak website, Reference and Google are more likely to convert. Though Direct traffic is giving highest number of leads their coversion rate is less than 50%
- Do not call or Email; Though customer have selected No, their conversions rate is higher than customers who have said yes to both calls and emails.
- There is a very conversion rate for Leads with Last Activity : SMS sent
- Almost all leads are from India, few from US and UAE
- Leads are spread across all Specializations, Specilizations under Domain Management are highlights
- Majority of leads are of Unemployed customers, however this doesnt reflect the same in conversion rate. Working professionals have a very high conversion rate.

- Step 5: Preparing the data for Model Building

-Converting some binary variables (Yes/No) to 0/1

-Creating dummy features for multi level variables

Step 6: Splitting the data into test and train datasets

-

- Step 7: Feature Scaling

-From the data, we can see TotalVisits, Total Time Spent on Website and Page Views Per Visit have larger values compared to others. We can normalize the numbers using the StandardScaler method and have all the numbers within small range.

-Checking the Conversion Rate

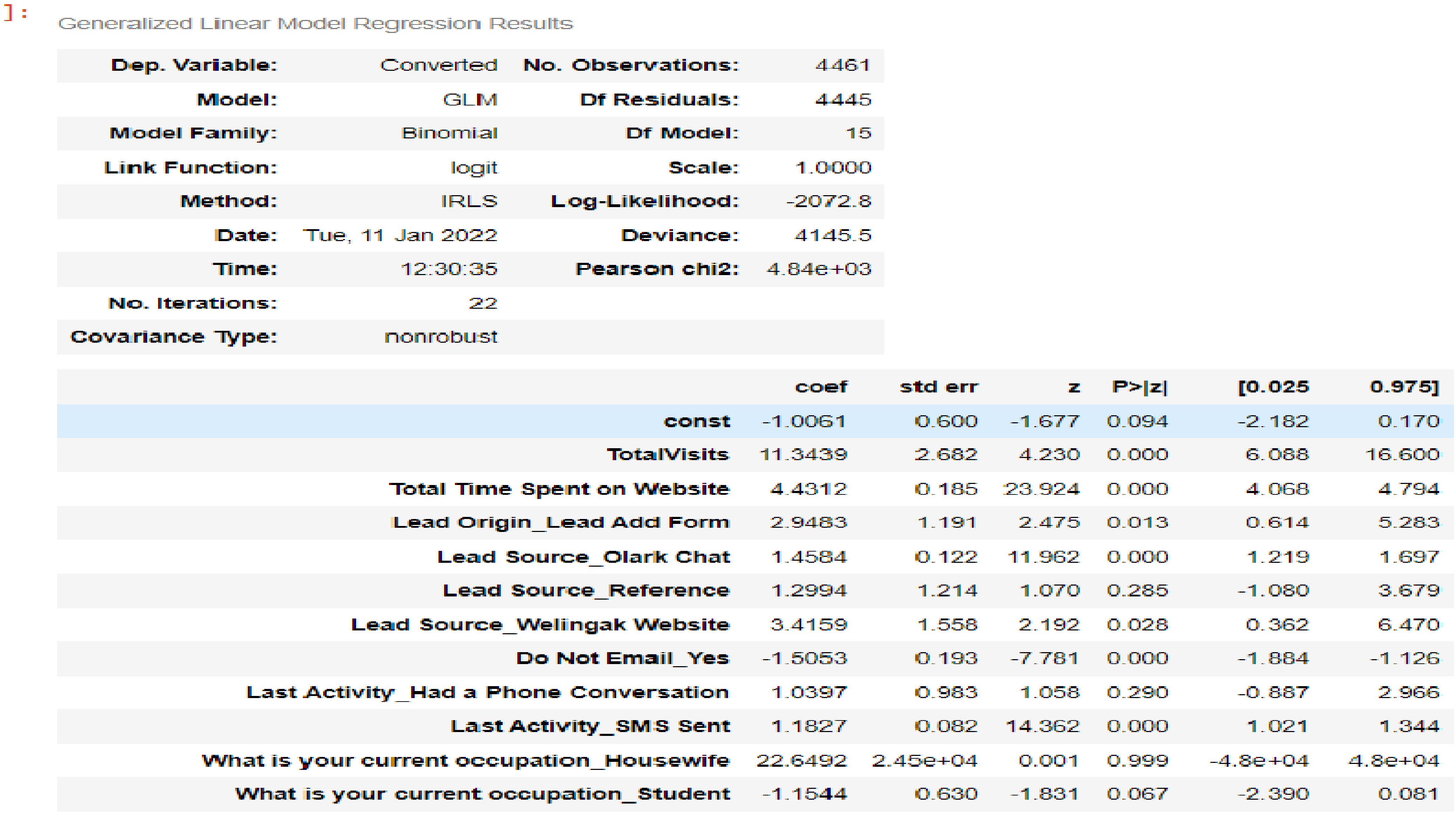
We are required to build a model such that customer would be assigned a lead score using which we achieve a target lead conversion rate to be around 80%

- Step 8: Looking at Correlations

- checking the correlation coefficients to see which variables are highly correlated

Step 9: Model Building

-Model 1 : Using all feature variables



]:

Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	4461
Model:	GLM	**Df Residuals:**	4445
Model Family:	Binomial	**Df Model:**	15
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2072.8
Date:	Tue, 11 Jan 2022	**Deviance:**	4145.5
Time:	12:30:35	**Pearson chi2:**	4.84e+03
No. Iterations:	22		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	-1.0061	0.600	-1.677	0.094	-2.182	0.170
TotalVisits	11.3439	2.682	4.230	0.000	6.088	16.600
Total Time Spent on Website	4.4312	0.185	23.924	0.000	4.068	4.794
Lead Origin_Lead Add Form	2.9483	1.191	2.475	0.013	0.614	5.283
Lead Source_Olark Chat	1.4584	0.122	11.962	0.000	1.219	1.697
Lead Source_Reference	1.2994	1.214	1.070	0.285	-1.080	3.679
Lead Source_Welingak Website	3.4159	1.558	2.192	0.028	0.362	6.470
Do Not Email_Yes	-1.5053	0.193	-7.781	0.000	-1.884	-1.126
Last Activity_Had a Phone Conversation	1.0397	0.983	1.058	0.290	-0.887	2.966
Last Activity_SMS Sent	1.1827	0.082	14.362	0.000	1.021	1.344
What is your current occupation_Housewife	22.6492	2.45e+04	0.001	0.999	-4.8e+04	4.8e+04
What is your current occupation_Student	-1.1544	0.630	-1.831	0.067	-2.390	0.081

Step 10: Feature Selection Using RFE

importing libraries for RFE

```
from sklearn.linear_model import LogisticRegression
from sklearn.feature_selection import RFE
```

- Step 11 : Making Predictions on train data
- Step 12 : Checking metrics using a confusion Matrix, Sensitivity and Specificity

We will import sklearn import metrics and find below points:

-Confusion matrix

-check the overall accuracy.

-we see the sensitivity of our logistic regression model

-calculate specificity

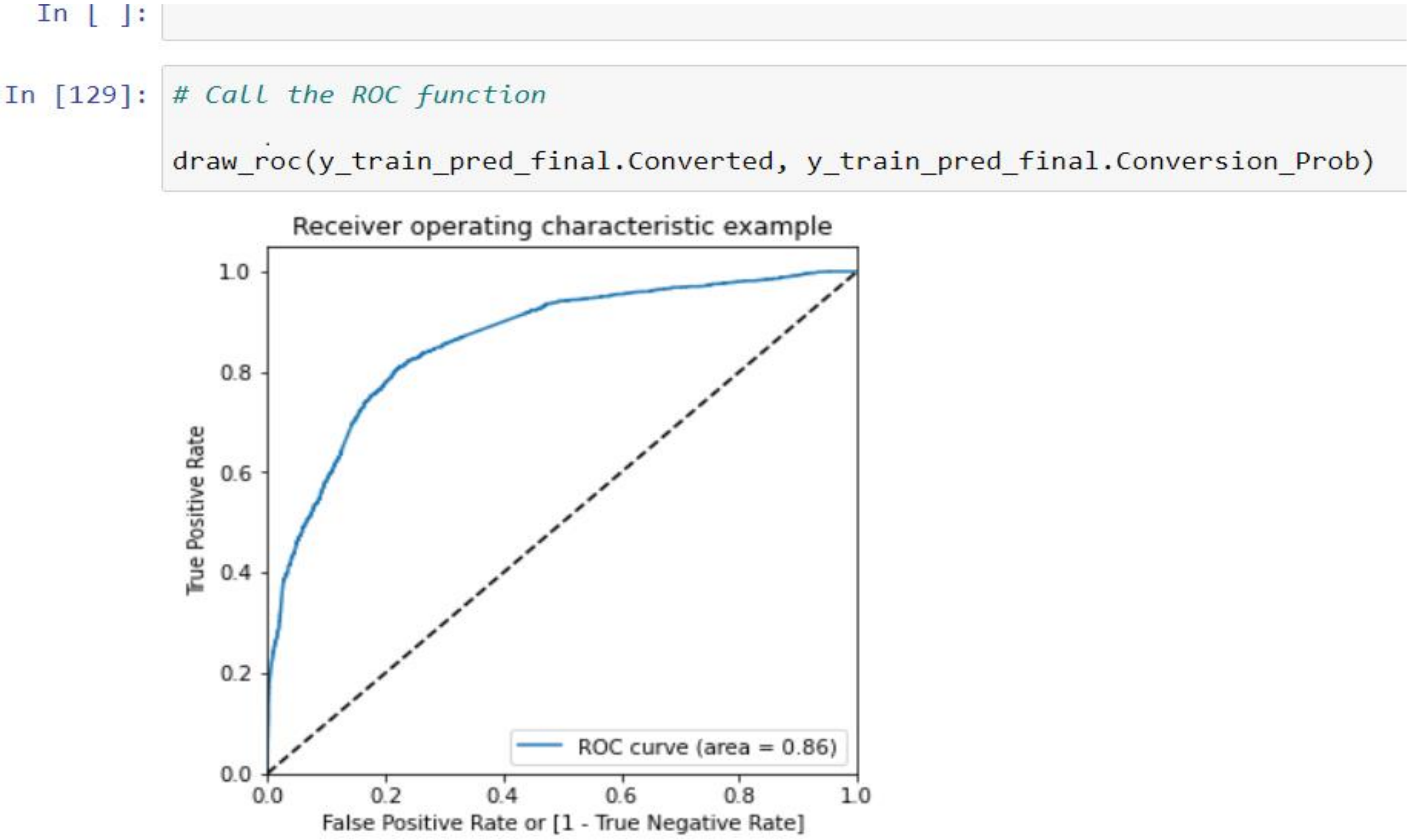
-positive predictive value

-Negative predictive value

- Step 14: Plotting a ROC curve to check AOC

-An ROC curve demonstrates several things:

It shows the tradeoff between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity). The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.



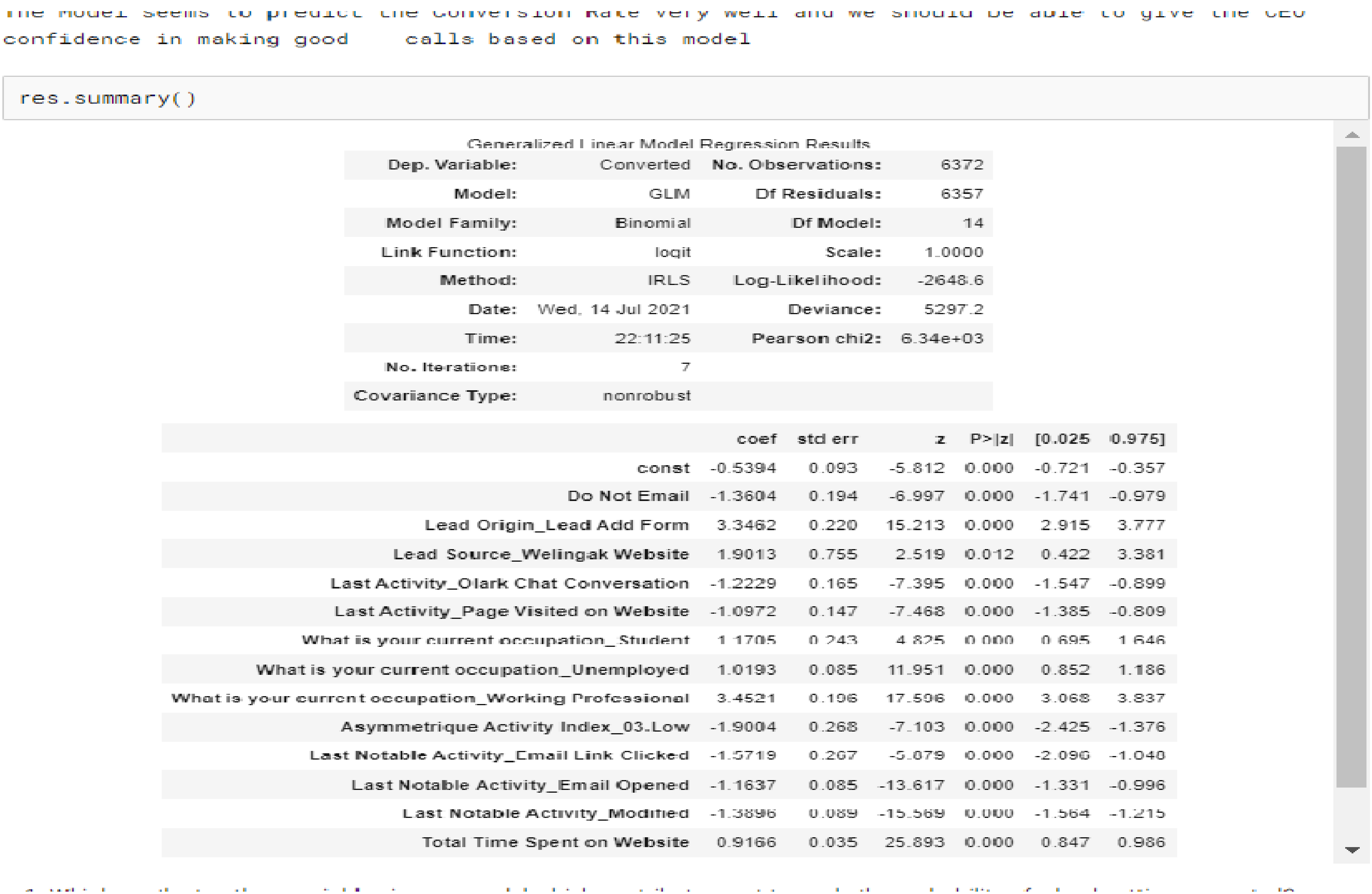
```
In [ ]:

In [129]: # Call the ROC function

          draw_roc(y_train_pred_final.Converted, y_train_pred_final.Conversion_Prob)
```

Receiver operating characteristic example
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate or [1 - True Negative Rate]
ROC curve (area = 0.86)

Step 15: Finding Optimal Cutoff Point

-we will plot accuracy sensitivity and specificity for various pro

-Creating a confusion matrix with the new thresholdbabilities.



The model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model

```
res.summary()
```

```
                 Generalized Linear Model Regression Results
   Dep. Variable:        Converted  No. Observations:      6372
           Model:              GLM     Df Residuals:      6357
    Model Family:         Binomial         Df Model:        14
   Link Function:            logit            Scale:    1.0000
          Method:             IRLS   Log-Likelihood:   -2648.6
            Date: Wed, 14 Jul 2021         Deviance:    5297.2
            Time:         22:11:25     Pearson chi2:  6.34e+03
  No. Iterations:                7
Covariance Type:         nonrobust

                                                 coef  std err       z  P>|z|  [0.025  0.975]
                                      const   -0.5394    0.093  -5.812  0.000  -0.721  -0.357
                               Do Not Email   -1.3604    0.194  -6.997  0.000  -1.741  -0.979
                   Lead Origin_Lead Add Form   3.3462    0.220  15.213  0.000   2.915   3.777
                 Lead Source_Welingak Website   1.9013    0.755   2.519  0.012   0.422   3.381
        Last Activity_Olark Chat Conversation  -1.2229    0.165  -7.395  0.000  -1.547  -0.899
        Last Activity_Page Visited on Website  -1.0972    0.147  -7.468  0.000  -1.385  -0.809
     What is your current occupation_Student   1.1705    0.243   4.825  0.000   0.695   1.646
  What is your current occupation_Unemployed   1.0193    0.085  11.951  0.000   0.852   1.186
What is your current occupation_Working Professional  3.4521  0.196  17.596  0.000  3.068  3.837
       Asymmetrique Activity Index_03.Low     -1.9004    0.268  -7.103  0.000  -2.425  -1.376
       Last Notable Activity_Email Link Clicked  -1.5719  0.267  -5.079  0.000  -2.096  -1.040
          Last Notable Activity_Email Opened  -1.1637    0.085 -13.617  0.000  -1.331  -0.996
              Last Notable Activity_Modified  -1.3896    0.089 -15.569  0.000  -1.564  -1.215
                 Total Time Spent on Website   0.9166    0.035  25.893  0.000   0.847   0.986
```

1. Which are the top three variables in your model which contribute most towards the probability of a lead getting converted?

paying customers from the bottom.

In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

First, sort out the best prospects from the leads you have generated. 'TotalVisits' , 'Total Time Spent on Website' , 'Page Views Per Visit' which contribute most towards the probability of a lead getting converted.

Then, You must keep a list of leads handy so that you can inform them about new courses, services, job offers and future higher studies.

Monitor each lead carefully so that you can tailor the information you send to them.

Carefully provide job offerings, information or courses that suits best according to the interest of the leads.

A proper plan to chart the needs of each lead will go a long way to capture the leads as prospects.

Focus on converted leads.

Hold question-answer sessions with leads to extract the right information you need about them.

Make further inquiries and appointments with the leads to determine their intention and mentality to join online courses.