LEAD SCORE CASE STUDY

Submitted By:

Alekya Seerapu Rani

Ananya

Anil

Problem Statement

X Education sells online courses to industry professionals. They need help in selecting the most promising leads, i.e. the ‘leads’ that are most likely to convert into paying customers.

Customers are called as ‘leads’ when they provide their contact details to the company through various channels.
The Sales team of the company follows up with these Leads to sell them the course and convert into paying customers.

The conversion to lead ratio is currently 30%. The company wants to improve this ratio by identifying the most potential leads, also known as ‘Hot Leads’.

Business Goals

As a Data Scientist, we have to build a Machine Learning model using Logistic Regression that helps the company in identifying the most potential lead and reducing the manpower in irrelevant follow ups by the sales team.

The CEO, in particular, has given a ballpark of target lead conversion rate to be around 80%.

Lead Conversion Process Demonstrated as a funnel

Overall Analysis approach in a nutshell:

1) Importing Data and Inspecting the Dataframe.

2) Data Preparation (Encoding Categorical Variables, Handling Null Values)

3) EDA (univariate analysis, outlier detection, checking data imbalance)

4) Dummy Variable Creation

5) Test-Train Split

6) Feature Scaling

7) Looking at Correlations

8) Model Building (Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-values)

9) Build final model and verifying Logistic Regression Assumptions for the model.

10) Model evaluation with different metrics Specificity , Sensitivity, Precision and Recall

Detailed Solution (Implementation done in Python)

1.Importing the data and Inspecting the dataframe :

Looking at the dataset we observed that following things need to be done:
Missing value handling is required for some of the features.
As we can see there are a lot of columns which have high number of missing values. Clearly, these columns are not useful.
Since, there are 9240 datapoints in our data frame, let's eliminate the columns having greater than 3200(~35%) missing values as they are of no use to us.

2.Data Preparation : Missing Value Handling :

After identifying all the missing data, dropped columns having more than 35% null values.

Lead Quality	52.0
Asymmetrique Profile Score	46.0
Asymmetrique Activity Score	46.0
Asymmetrique Profile Index	46.0
Asymmetrique Activity Index	46.0
Tags	36.0

There are a few columns in which there is a level called 'Select' which basically means that the student had not selected the option for that particular column which is why it shows 'Select'. These values are as good as missing values and hence we need to identify the value counts of the level 'Select' in all the columns that it is present.

The following 4 columns have level **'Select**' with the below mentioned value counts:
Specialization: Select 1942
How did you hear about X Education: Select 5043
Lead Profile: Select 4146
City: Select 2249

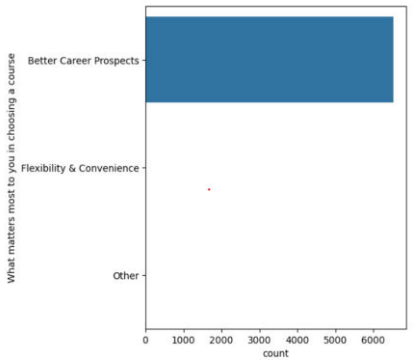
Further analysis , we concluded that the columns **"Lead Profile"** and "**How did you hear about X Education**" have a lot of rows which have the value Select which is of no use to the analysis so it's best that we drop them.

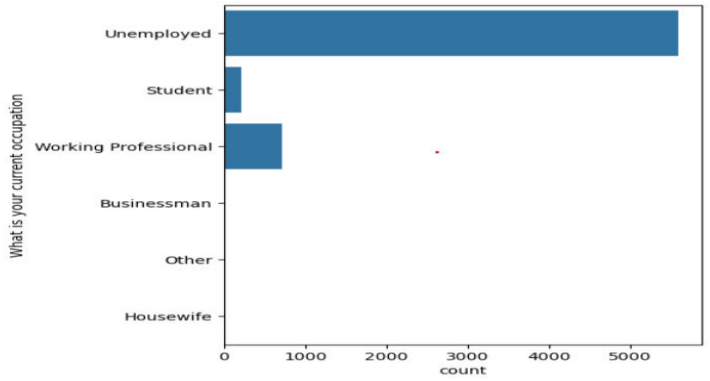
Also **City** has mostly Mumbai and Select as values, if we replace Select with Mumbai or Unknown still it will be skewed. Hence dropping this column as well.

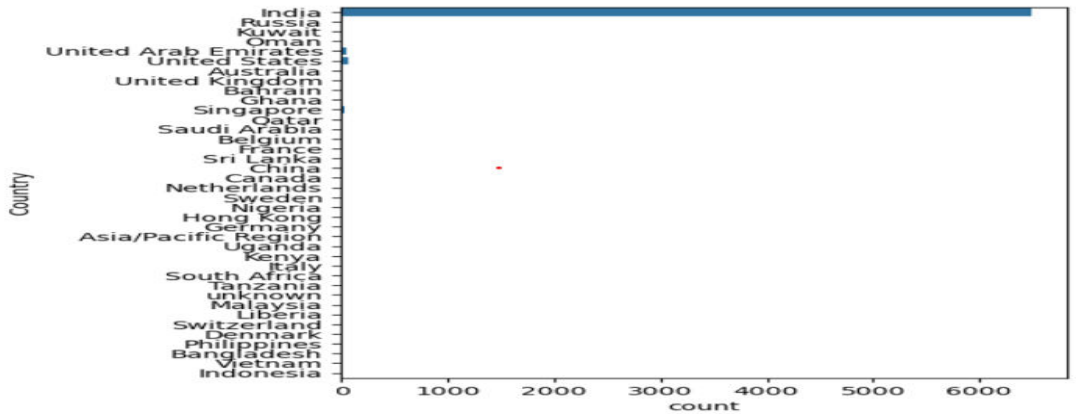
Will talk about **Specialization** in the upcoming slide.

Lets proceed with the next set of columns with missing value counts as below:

```
What matters most to you in choosing a course    2709
What is your current occupation                  2690
Country                                          2461
```







We can see that **“What matters most to you in choosing a course**’” is highly skewed column we can drop this column.

85% of the values under **'What is your current occupation'** column is 'Unemployed', hence we can impute missing values in this column with this value

We can see that **Country** is highly skewed column but it is an important information w.r.t. to the lead. Since most values are 'India' , we can impute missing values in this column with this value.

We saw that **'Specialization'** has 37% missing values

Select values are as good as missing values

We have replace missing values with 'Select' since that is also a missing value in this column. So merging all missing values under one umbrella for easy identification in further analysis.

We can observe that the below columns have very minimal amount of missing values, we preferred dropping those rows with missing values:

TotalVisits	137
Page Views Per Visit	137
Last Activity	103
Lead Source	36

Checking for duplicates

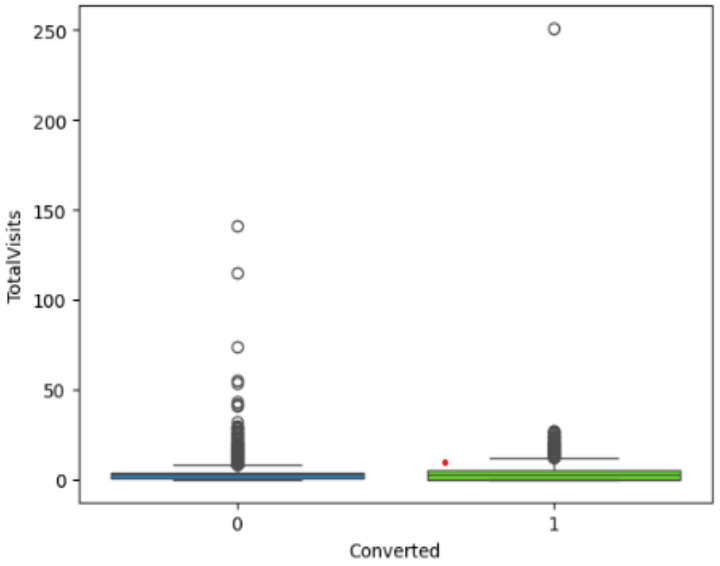
Prospect ID	Lead Number	Lead Origin	Lead Source	Do Not Email	Do Not Call	Converted	TotalVisits	Total Time Spent on Website	Page Views Per Visit	...	X Education Forums	Newspaper	Digital Advertisement	Through Recommendations	Receive More Updates About Our Courses	Up m Su C Cor

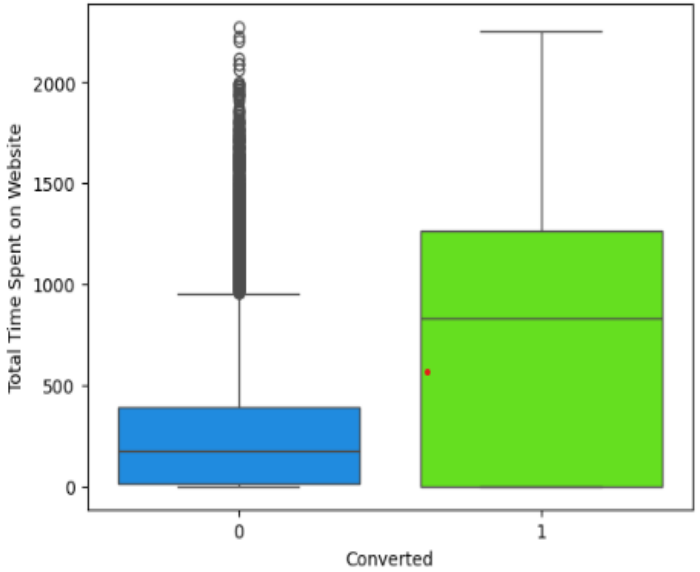
0 rows × 27 columns

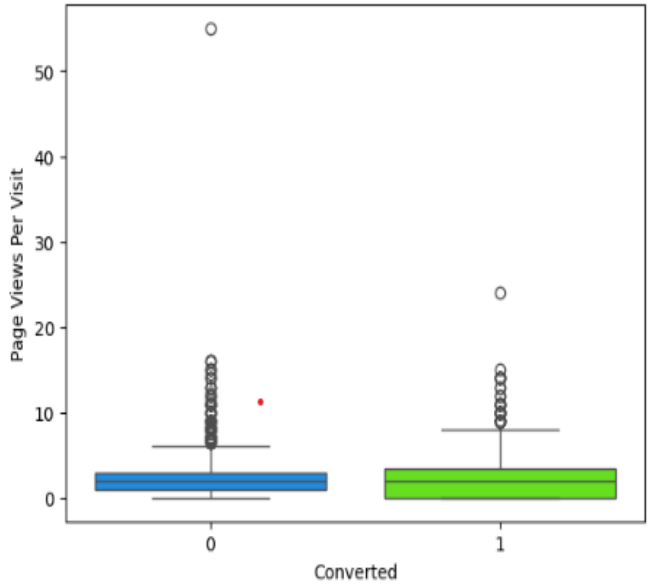
There are no duplicate records in the dataset

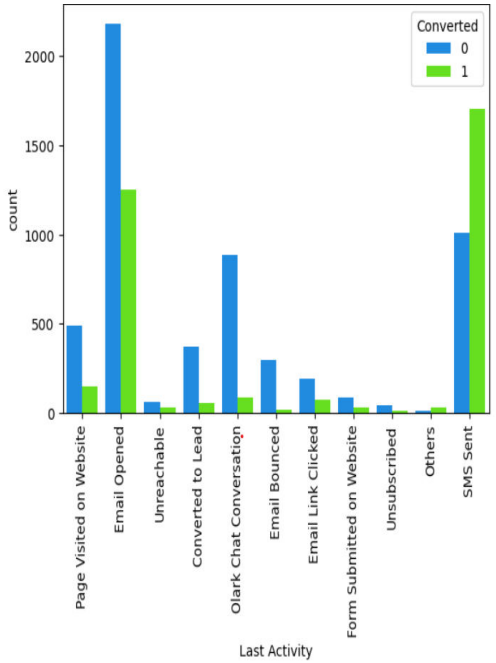
Exploratory Data Anaysis

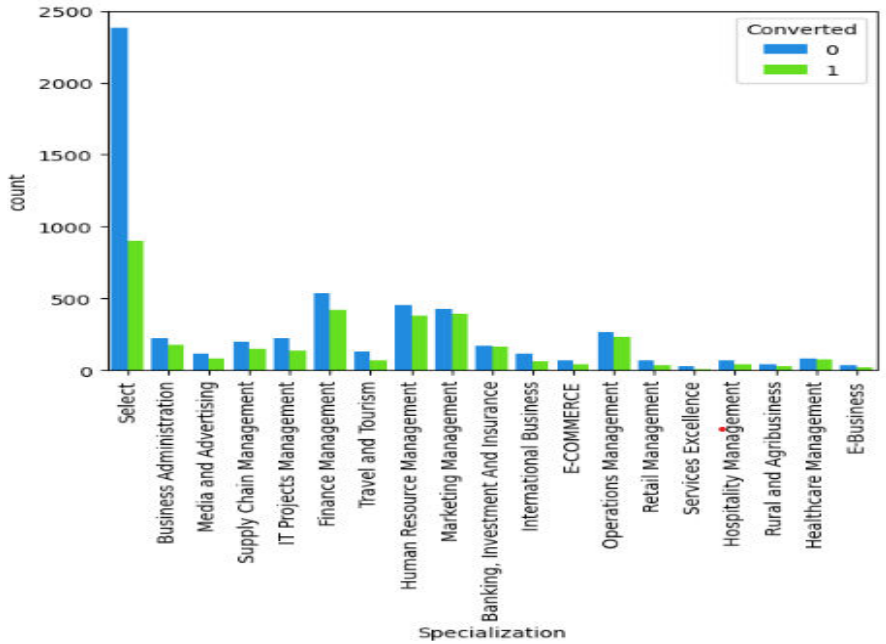
Univariate Analysis and Bivariate Analysis

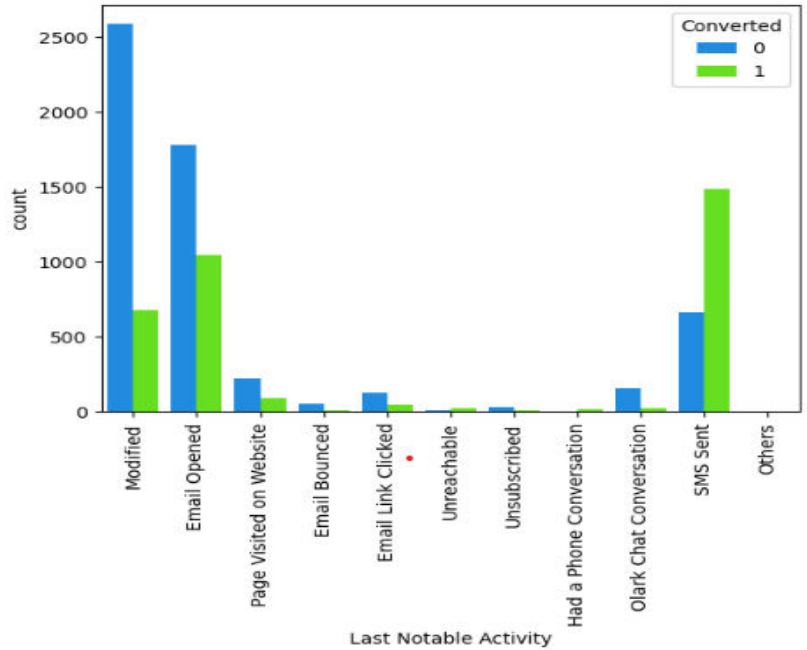


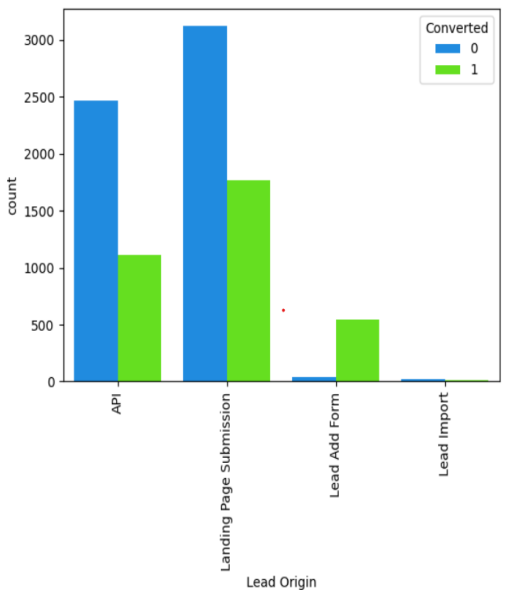


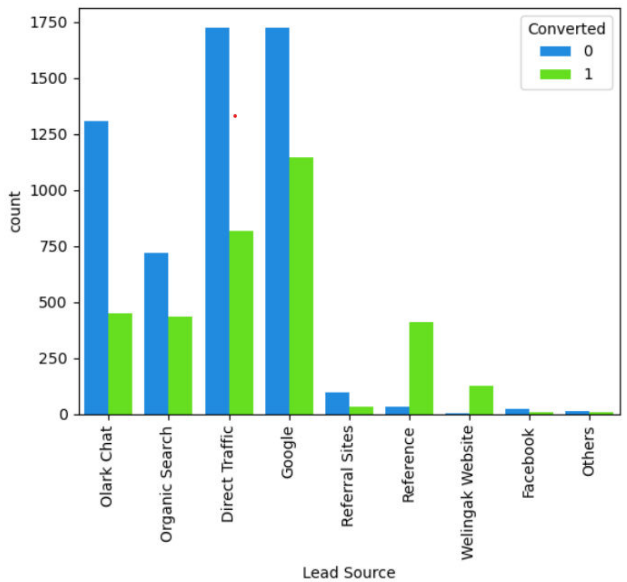


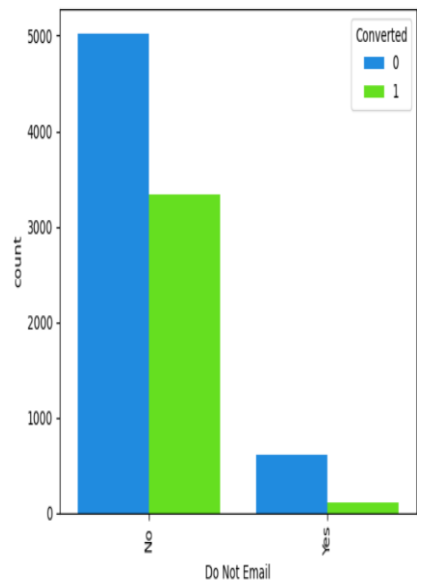

Converted
0
1
count
2000
1500
1000
500
0
Page Visited on Website
Email Opened
Unreachable
Converted to Lead
Olark Chat Conversation
Email Bounced
Email Link Clicked
Form Submitted on Website
Unsubscribed
Others
SMS Sent
Last Activity

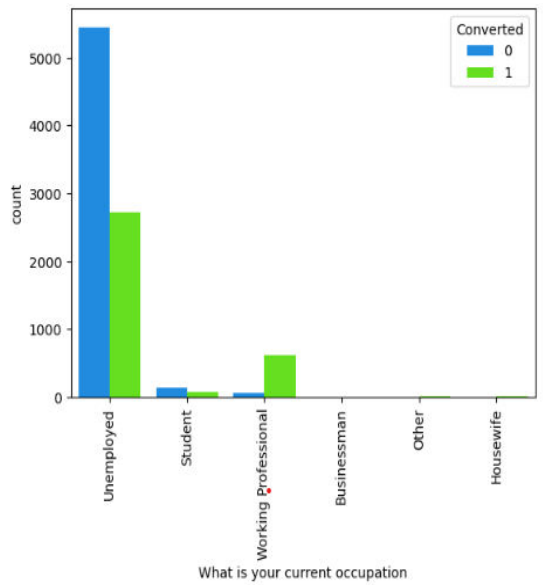

Converted
0
1
count
2500
2000
1500
1000
500
0
Select
Business Administration
Media and Advertising
Supply Chain Management
IT Projects Management
Finance Management
Travel and Tourism
Human Resource Management
Marketing Management
Banking, Investment And Insurance
International Business
E-COMMERCE
Operations Management
Retail Management
Services Excellence
Hospitality Management
Rural and Agribusiness
Healthcare Management
E-Business
Specialization


Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Page Visited on Website
Email Bounced
Email Link Clicked
Unreachable
Unsubscribed
Had a Phone Conversation
Olark Chat Conversation
SMS Sent
Others
Last Notable Activity


Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Lead Origin


Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Referral Sites
Reference
Welingak Website
Facebook
Others
Lead Source


Converted
0
1
count
5000
4000
3000
2000
1000
0
No
Yes
Do Not Email


Converted
0
1
count
5000
4000
3000
2000
1000
0
Unemployed
Student
Working Professional
Businessman
Other
Housewife
What is your current occupation

Inferences from EDA :

To improvise the lead conversion rate, we need to focus more on improving lead converion of API and Landing Page Submission origin and generate more leads from Lead Add Form.

To improve overall lead conversion rate, focus should be on improving lead converion of 'olark chat', 'organic search', 'direct traffic', and 'google' leads and generate more leads from 'reference' and 'welingak website'.

Focus can be made on enhancing the customer experience on website, in turn increasing the time customer spends on the websites, resulting in increasing the conversion rates.

We should focus on increasing the conversion rate of those having Last Activity as Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as SMS sent.

We can focus on increasing Working Professional leads by reaching out to them through different channels as they show good conversion rate and also focus should be made on increasing the conversion rate of Unemployed leads

Customers having job experience in certain Industry Domains have shown considerable conversion rates.

'Will revert after reading the email' and 'Closed by Horizzon' categories under Tags have high conversion rate
We should focus on increasing the conversion rate of those having 'Last Notable Activity' as Modified and Email Opened by making a call to those leads and also try to increase the count of the ones having last activity as 'SMS sent'.

Inference from the Univariate Analysis

•We have observed that for the following columns, we could not draw and inferences, as the data was either highly skewed, or had single value hence we will drop those columns to proceed further analysis.

‘Country’','Search','Magazine','Newspaper Article','X Education Forums', 'Newspaper','Digital Advertisement', 'Through Recommendations','Receive More Updates About Our Courses','Update me on Supply Chain Content', 'Get updates on DM Content','I agree to pay the amount through cheque','A free copy of Mastering The Interview’,
'Do Not Call’

•Also dropping Prospect ID and Lead Number as we dont need it for our analysis.

After cleaning the data, we have retained 98% of the rows. Which is very good for our analysis as it helps in building a good prediction model.

Data Preparation:

Creating Dummy Variables:

As logistic regression can work with numeric data only, creating dummy variables for the categorical columns listed below.

#	Column	Non-Null Count	Dtype
0	Lead Origin	9074 non-null	object
1	Lead Source	9074 non-null	object
2	Do Not Email	9074 non-null	object
3	Converted	9074 non-null	int64
4	TotalVisits	9074 non-null	float64
5	Total Time Spent on Website	9074 non-null	int64
6	Page Views Per Visit	9074 non-null	float64
7	Last Activity	9074 non-null	object
8	Specialization	9074 non-null	object
9	What is your current occupation	9074 non-null	object
10	Last Notable Activity	9074 non-null	object

We created dummy variable separately for the variable 'Specialization' since it has Select as Unknown values, which is not required for analysis, so we drop that level by specifying it explicitly.

Splitting Data into Training and Test set

Next, the dataset was split into training and test set, to train model first with a chunk of data and then evaluate its performance on unseen data.

Feature Scaling

Feature Scaling is required before Logistic Regression to bring all the features in same scale, this ensures that features with high magnitude are not given higher importance by Logistic Regression Model.

```
# Checking the Lead Conversion rate

Converted = (sum(lead['Converted'])/len(lead['Converted'].index))*100
Converted
```

```
37.85541106458012
```

Currently the company has 37.85% conversion rate. This means among the targeted people only 37% are converting into customers. This is our focus area to improvise this percentage.

Checking Correlation

Since Logistic Regression Model is high affected by multi-collinearity, removing the features showing high correlation.

Lead Origin_Lead Import	Lead Source_Facebook	0.983684
Last Activity_Unsubscribed	Last Notable Activity_Unsubscribed	0.872656
Lead Origin_Lead Add Form	Lead Source_Reference	0.866191
Last Activity_Email Opened	Last Notable Activity_Email Opened	0.861636
Last Activity_SMS Sent	Last Notable Activity_SMS Sent	0.853102
Last Activity_Email Link Clicked	Last Notable Activity_Email Link Clicked	0.800686
Last Activity_Page Visited on Website	Last Notable Activity_Page Visited on Website	0.691811
Do Not Email_Yes	Last Activity_Email Bounced	0.620041
Last Activity_Unreachable	Last Notable Activity_Unreachable	0.594369
Last Activity_Others	Last Notable Activity_Had a Phone Conversation	0.576457

We can see few features are highly correlated, hence dropping highly correlated features from Train and Test data.

'Lead Source_Facebook','Last Notable Activity_Unsubscribed','Last Notable Activity_SMS Sent', 'Last Notable Activity_Email Opened','Last Notable Activity_Unreachable','Last Notable Activity_Email Link Clicked','Last Notable Activity_Page Visited on Website'

Model Building

(Feature Selection Using RFE, Improvising the model further inspecting adjusted R-squared, VIF and p-vales)

Recursive feature elimination is a method removes the weakest feature (or features) until the specified number of features is reached.
We used RFE to obtain top 15 features to begin with.

After that, manually inspecting p-values and VIF to improve the model even further.

We will be dropping columns (highly collinear ones) with high p-value and VIF and building new model until we get set of features within acceptable ranges of p-value(0.05) and VIF(0.5).

A low p-value (<0.05) suggests that the coefficient is statistically significant, implying a meaningful association between the variable and the response.

Multicollinearity can prevent predictive models from producing accurate predictions by increasing model complexity and overfitting. Hence we focus to achieve VIF<0.5.

All variables have a good value of VIF. And p-values. So we ended up with the below mentioned 13 Features. And this is our final model, which we will use to make the predictions.

	Features	VIF
11	What is your current occupation_Unemployed	2.41
12	Last Notable Activity_Modified	1.81
3	Lead Source_Olark Chat	1.73
6	Last Activity_Olark Chat Conversation	1.56
8	Last Activity_SMS Sent	1.52
1	Lead Origin_Lead Add Form	1.46
4	Lead Source_Welingak Website	1.31
0	Total Time Spent on Website	1.28
5	Do Not Email_Yes	1.20
9	Last Activity_Unsubscribed	1.08
10	What is your current occupation_Student	1.03
2	Lead Origin_Lead Import	1.01
7	Last Activity_Others	1.01

	P>|z|
const	0.000
Total Time Spent on Website	0.000
Lead Origin_Lead Add Form	0.000
Lead Origin_Lead Import	0.001
Lead Source_Olark Chat	0.000
Lead Source_Welingak Website	0.008
Do Not Email_Yes	0.000
Last Activity_Olark Chat Conversation	0.000
Last Activity_Others	0.000
Last Activity_SMS Sent	0.000
Last Activity_Unsubscribed	0.003
What is your current occupation_Student	0.000
What is your current occupation_Unemployed	0.000
Last Notable Activity_Modified	0.000

Final Model and Model Evaluation

Now that we have the final set of features obtained by removing highly collinear ones, using RFE, inspecting p-values and VIF- we can build the final logistic regression model and evaluate its performance. Lets have a look at the evaluation metrices for Logistic Regression.

Confusion matrix:

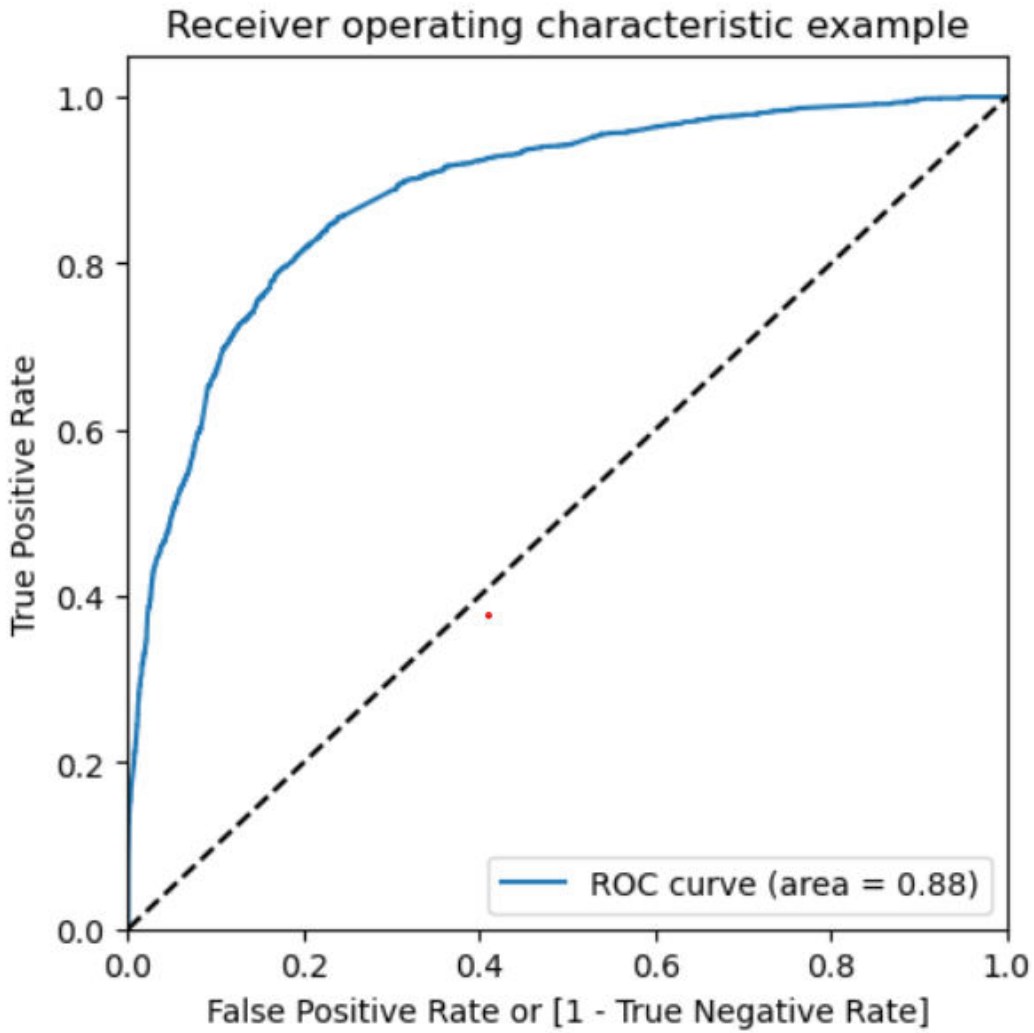
```
[[3475  430]
 [ 735 1711]]
```

Sensitivity : 69.95%
Specificity : 88.89%
Accuracy : 81.65%

Accuracy is the percentage of correctly predicted labels (positive and negative) among all the labels.
Sensitivity gives us the percentage of correctly predicted conversion out of total conversions.
Sensitivity = True Positive/(True Positive + False Negative)
Specificity gives us the percentage of correctly predicted non-conversion out of total non-conversions.
Specificity = True Negative/(True Negative + False Positive)

We will try to improvise the Sensitivity score as number of Conversions that are the top most priority in this Business Use Case.

Plotting the ROC Curve



ROC curve shows tradeoff between sensitivity and specificity (increase in one will cause decrease in other).

•The closer the curve follows the y-axis and then the top border of the ROC space, means more area under the curve and the more accurate the test.

•The closer the curve comes to the 45-degree diagonal of the ROC space i.e. the reference line, means less area and the less accurate is the test.

•After plotting the ROC curve, we see that the area under the curve is 0.88 which is high and this is indicates that the model that we have built is a good model.

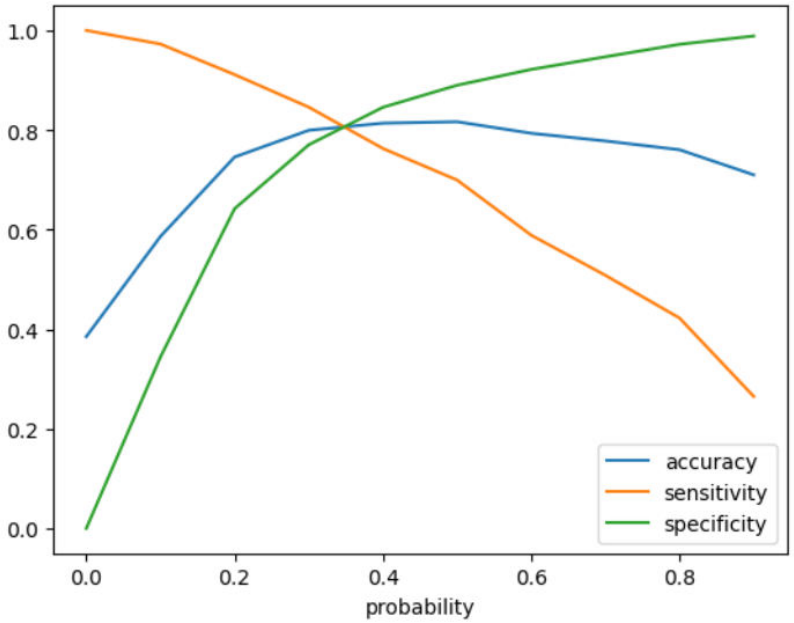
Optimal cutoff probability is that probability where we get balanced sensitivity and specificity.Since we choose an arbitrary cut-off value of 0.5 earlier, we need to determine the best cut-off value and the below section deals with that.

Next, we need to find out the optimal cut-off. Here's the predicted conversion status based on different values of cut-off.

	Converted	Converted_prob	Prospect ID	predicted	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0	0.198519	3009	0	1	1	0	0	0	0	0	0	0	0
1	0	0.301167	1012	0	1	1	1	1	0	0	0	0	0	0
2	0	0.358675	9226	0	1	1	1	1	0	0	0	0	0	0
3	1	0.873619	4750	1	1	1	1	1	1	1	1	1	1	0
4	1	0.811317	7987	1	1	1	1	1	1	1	1	1	1	0

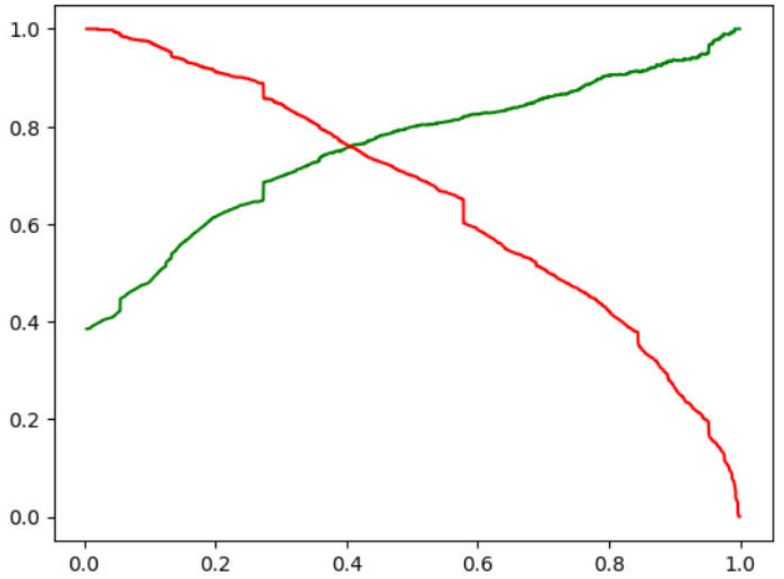
For different probability, we calculate the Accuracy, Sensitivity and Specificity for various probability cutoffs.

	probability	accuracy	sensitivity	specificity
0.0	0.0	0.385136	1.000000	0.000000
0.1	0.1	0.586207	0.972608	0.344174
0.2	0.2	0.745709	0.911284	0.641997
0.3	0.3	0.799402	0.845871	0.770294
0.4	0.4	0.813730	0.762878	0.845583
0.5	0.5	0.816564	0.699509	0.889885
0.6	0.6	0.793418	0.588716	0.921639
0.7	0.7	0.777988	0.508177	0.946991
0.8	0.8	0.760353	0.422322	0.972087
0.9	0.9	0.709967	0.264922	0.988732



The above graph is the accuracy, sensitivity and specificity plotted for various probabilities.

We see that the point of intersection is approximately around 0.38 and try to predict the final probability of the train set.



The above graph shows the trade-off between the Precision and Recall .
The optimal cutoff point is where the values of precision and recall will be equal The optimal value shows around 0.38.
So, when precision and recall are both around 0.38, the two curves are intersecting.

We are taking 0.38 is the optimum point as a cutoff probability, this cut-off helps in assigning a Lead Score in training data, which acts as an indicator if the Lead is a Hot Leads or not
i.e., any lead with greater than 0.38 probability of converting is predicted as Hot Lead.

Results:

Comparing the values obtained for Train & Test:

```
Train Data:
Accuracy : 81.31 %
Sensitivity : 69.95 %
Specificity : 88.98 %
Precision :  79.92%

Test Data:
Accuracy : 78.89 %
Sensitivity : 80.02 %
Specificity : 77.79 %
Precision : 72.15%
```

We have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% .

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 to get a higher lead conversion rate of 80%.

Important Features from our final mode:

Converted = 1.471697+ Lead Origin_Lead Add Form*3.908569+Last Activity_Others*2.281693+Lead Source_Welingak Website*2.012467+ Lead Origin_Lead Import*1.477128+ Last Activity_Unsubscribed*1.435453+Last Activity_SMS Sent*1.294636+ Lead Source_Olark Chat*1.238324+ Total Time Spent on Website*1.127928 -Last Notable Activity_Modified*0.896926- Last Activity_Olark Chat Conversation*0.970622- Do Not Email_Yes*1.704308 - What is your current occupation_Student*2.272949- What is your current occupation_Unemployed*2.686516

Conclusion:

The logistic regression model predicts the probability of the target variable having a certain value, rather than predicting the value of the target variable directly.

Then a cutoff of the probability is used to obtain the predicted value of the target variable.

Here, the logistic regression model is used to predict the probabilty of conversion of a customer.

Optimum cut off is chosen to be 0.38 i.e. any lead with greater than 0.38 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.38 or less probability of converting is predicted as Cold Lead (customer will not convert)

Our final Logistic Regression Model is built with 13 features.
Features used in final model are:" What is your current occupation_Unemployed , Last Notable Activity_Modified, Lead Source_Olark Chat, Last Activity_Olark Chat Conversation, Last Activity_SMS Sent, Lead Origin_Lead Add Form, Lead Source_Welingak Website, Total Time Spent on Website, Do Not Email_Yes, Last Activity_Unsubscribed, What is your current occupation_Student, Lead Origin_Lead Import, Last Activity_Others"

The top three categorical/dummy variables in the final model are

‘Lead Origin_Lead Add Form’, ‘Last Activity_Others’, ‘Lead Source_Welingak Website’ with respect to the absolute value of their coefficient factors.

‘Lead Origin_Lead Add Form’ is obtained by Dummy Encoding of original categorical variable 'Lead Origin’

‘Last Activity_Others’, is obtained by Dummy Encoding of original categorical variable 'Last Activity’

‘Lead Source_Welingak Website’ is obtained by Dummy Encoding of original categorical variable 'Lead Source’

Lead Origin_Lead Add Form having Coefficient factor = 3.908569
Last Activity_Others having Coefficient factor = 2.281693
Lead Source_Welingak Website having Coefficient factor = 2.012467

The final model has Sensitivity of 69.92%, this means the model is able to predict 69.92% customers out of all the converted customers, (Positive conversion) correctly.
The final model has Precision of 78.15%, this means 78.15% of predicted hot leads are True Hot Leads.

We have also built an reusable code block which will predict Convert value and Lead Score given training, test data and a cut-off. Different cutoffs can be used depending on the use-cases (for eg. when high sensitivity is required, when model have optimum precision score etc.)

Recommendations:

•The company **should make calls and nurture** the leads coming from **Lead Origin - Lead Add Form and Lead Import** as these are more likely to get converted.

•The company **should make calls and nurture** the leads with **Last Activity SMS Sent and Others like customers who Had a Phone Conversation, View in browser link, Visited Booth in Tradeshow, Approached upfront,Resubscribed to emails,Email Received** as these are more likely to get converted.

•The company **should make calls and nurture** the leads coming from **Lead Source - Welingak Website and Olark Chat** , as these are more likely to get converted.

•The company **should make calls and nurture** the leads who spent **more time on the websites** , as these are more likely to get converted.

•The company **should not make calls** to the leads whose **last Notable activity was Modified**, as they are not likely to get converted.

•The company **should not make calls** to the leads whose **last activity was Olark Chat Conversation**, as they are not likely to get converted.

•The company **should not make calls** to the leads who **opted out of Emails**, as they are not likely to get converted.

•The company **should not make calls** to the leads who are **"Students" or "Unemployed"**, as they are not likely to get converted.