



# Lead Score Case Study

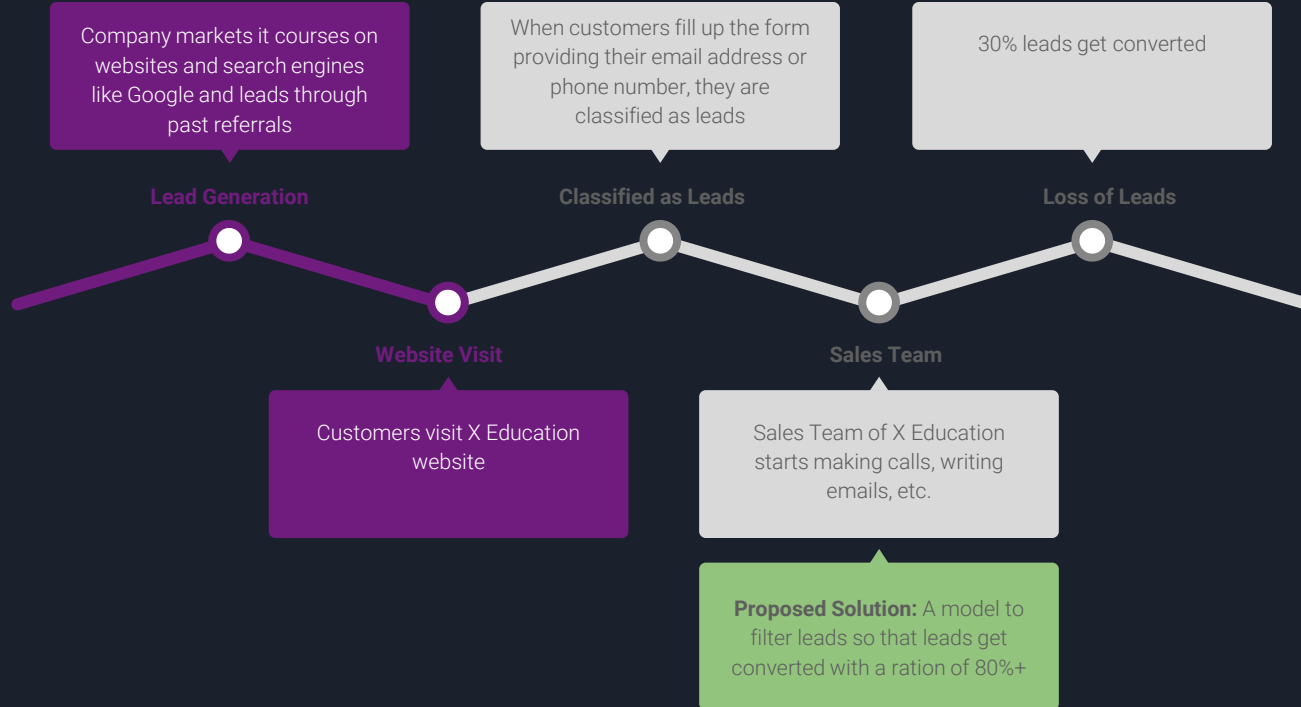
By: Shikhar Verma (APFE20806252)  
Uttkarsh Mishra (DDS2070327)



# Background

- An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.
- The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- The company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

# Lead Conversion Rate





# Proposed Solution

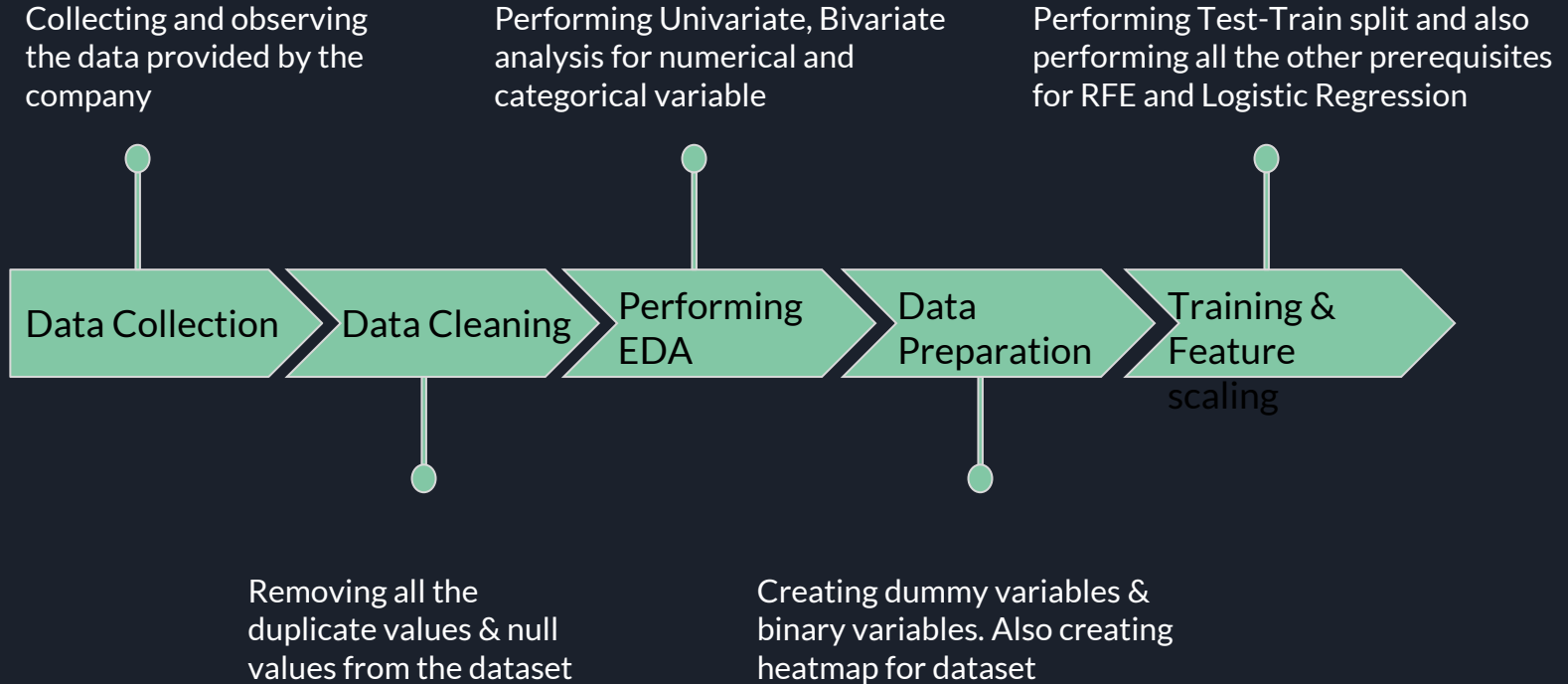
01	Selection of Hot Leads	Clustering the leads into certain categories based on their probability of conversion. Thus, getting a smaller section of hot leads to focus more on.
02	Communication with Hot Leads	Since after probability conversion, we get a small set of leads to have a communication with.
03	Conversion of Hot Leads	Since we focused on a small set of leads i.e. hot leads, which have more probability to convert which gives a better conversion rate and hence we can achieve the 80% target




# Solution

- The crucial part for our Problem Solution is to accurately identify hot leads.
- The more accurate we obtain the Hot Leads, the more chance we get of getting a higher conversion rate.
- Since we have a target of 80% conversion rate, we would want to obtain a high accuracy rate in order to obtain Hot Leads.

# Implementation

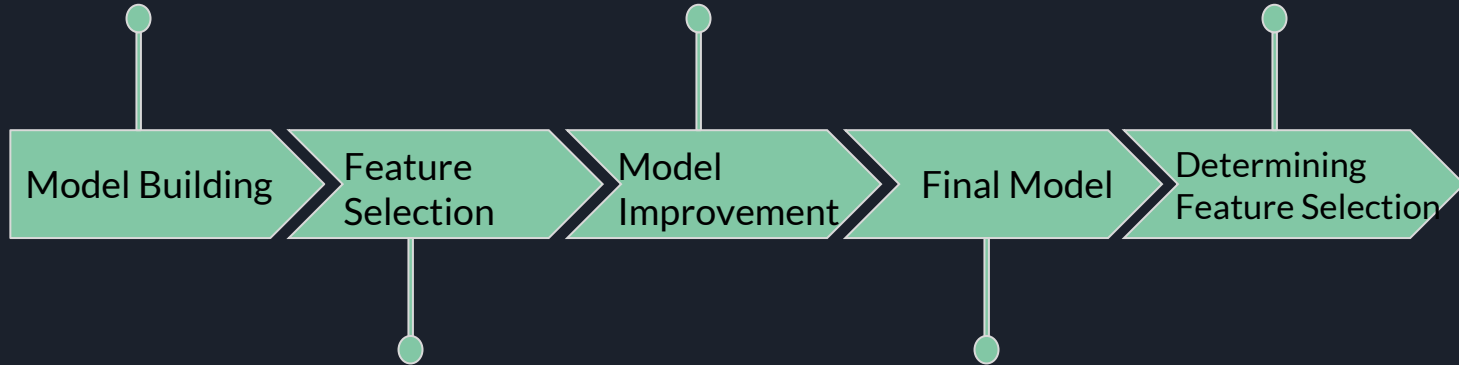




Running the First  
Training Model

Reducing the columns  
and re-building the model

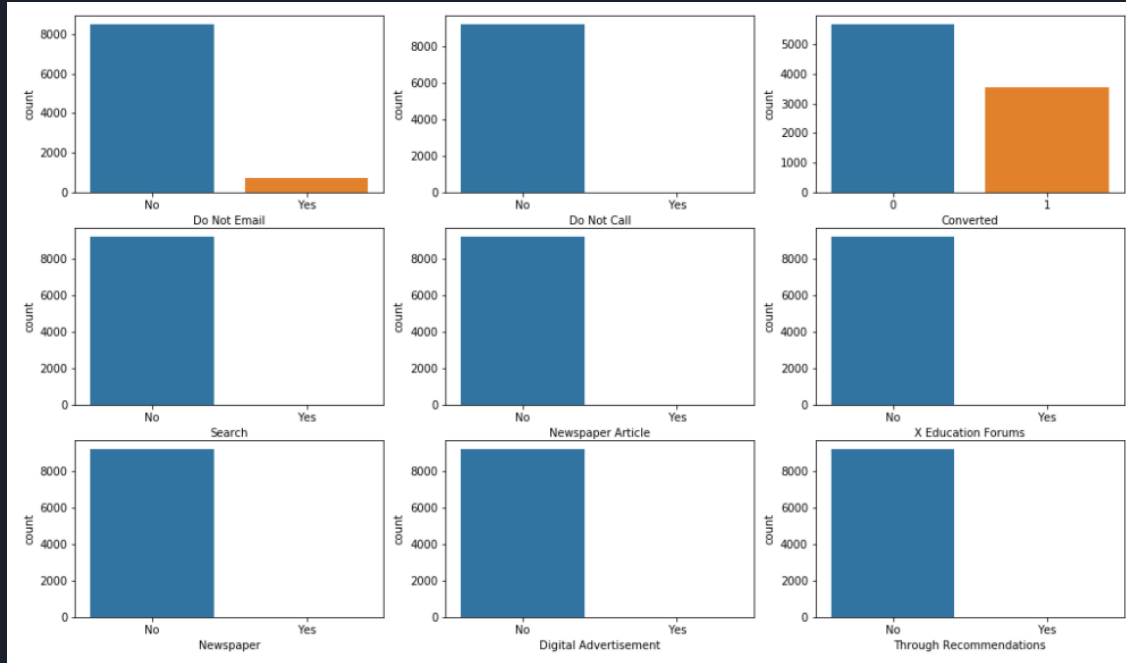
Determining the feature  
importance on the basis  
of coefficients



Selecting the top 15  
feature using RFE

Doing the Final Model  
Evaluation and Making  
predictions on test  
dataset

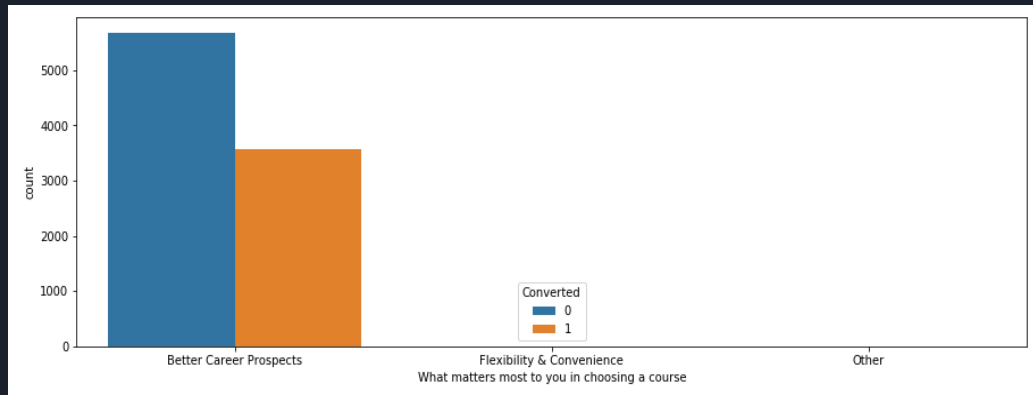
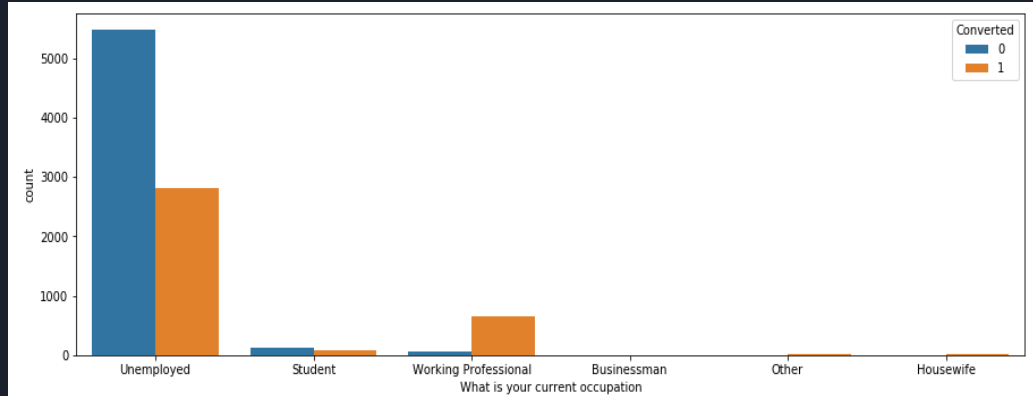
# Exploratory Data Analysis (EDA) + Univariate Analysis



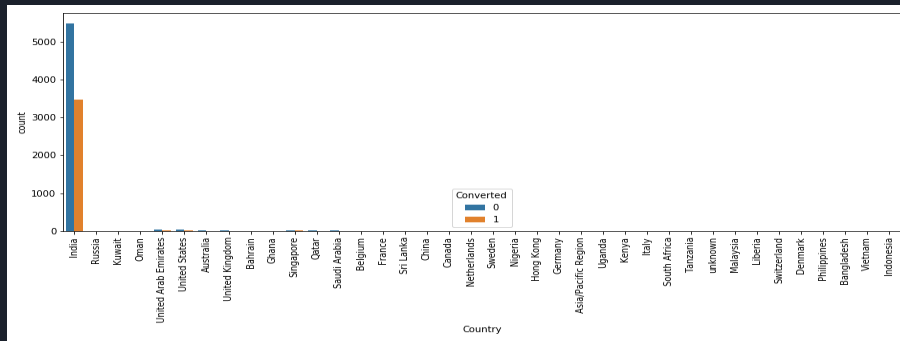
- Most customers don't want to be called about the course
- Very few customers want to get email regarding the course



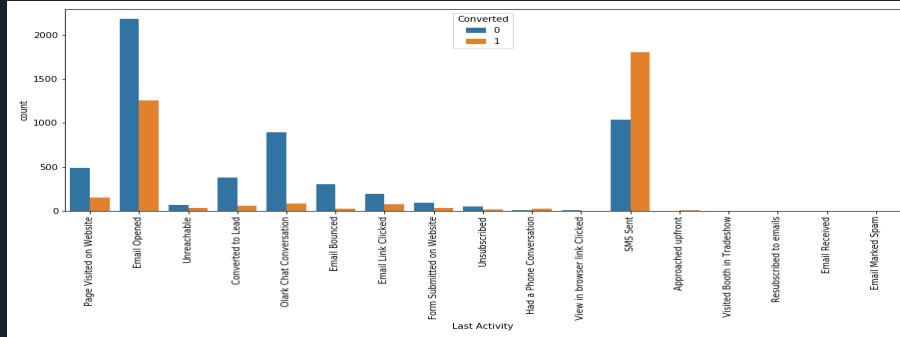
# Exploratory Data Analysis (EDA) + Bivariant Analysis



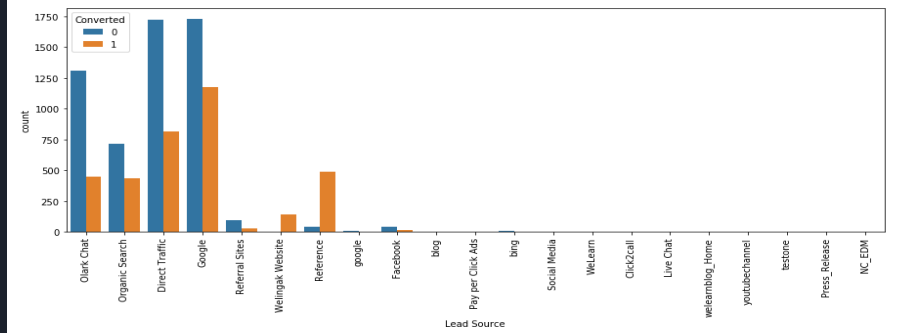
- Many leads that are 'Unemployed' take up a course
- Some of 'Working Professional' also take up courses
- For leads 'Better Career Prospects' matters most for choosing a course



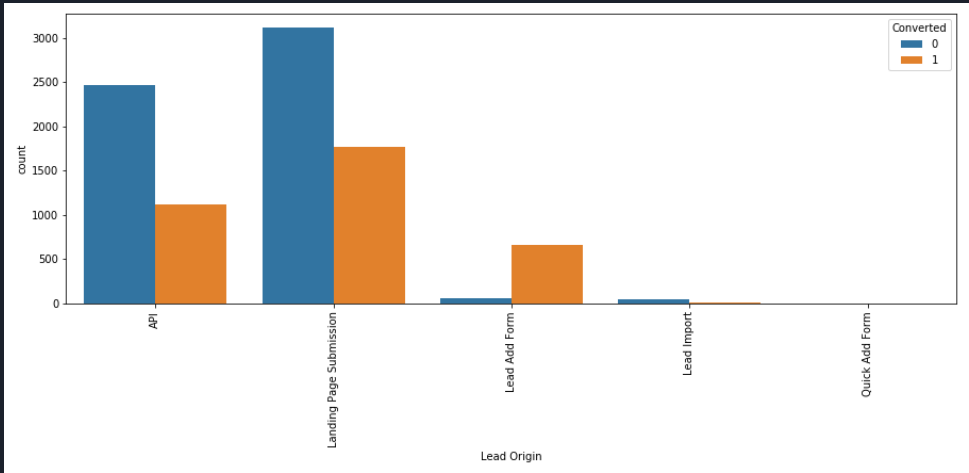
- India has the highest conversion



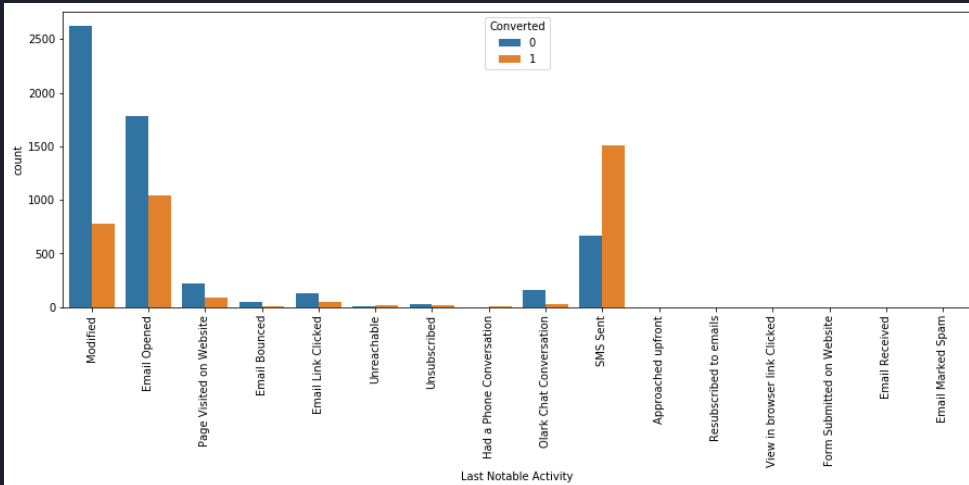
- 'SMS Sent' leads to highest conversion



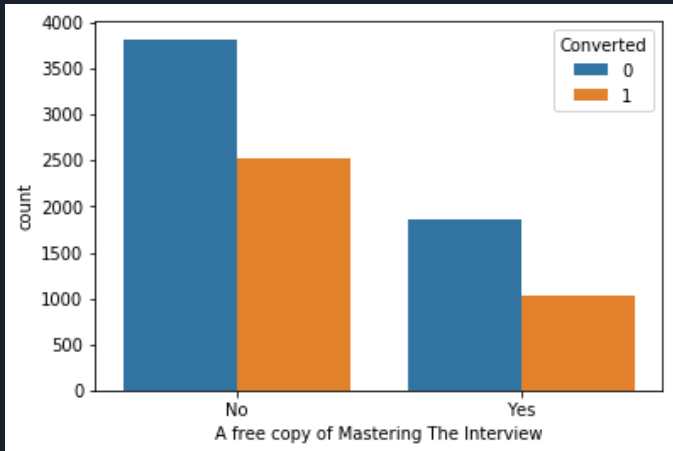
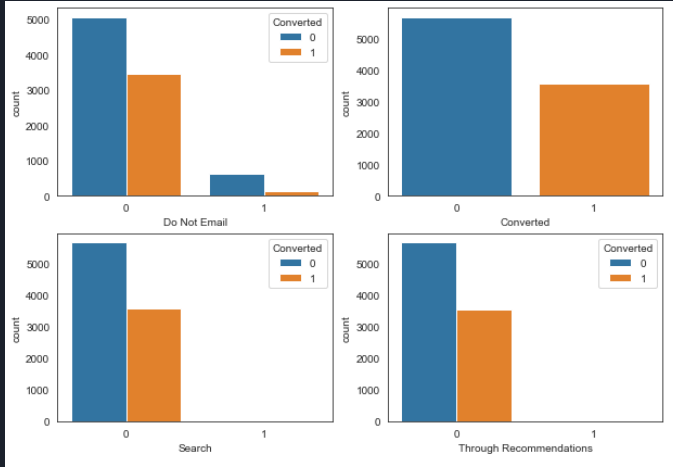
- 'Google' leads to highest conversion



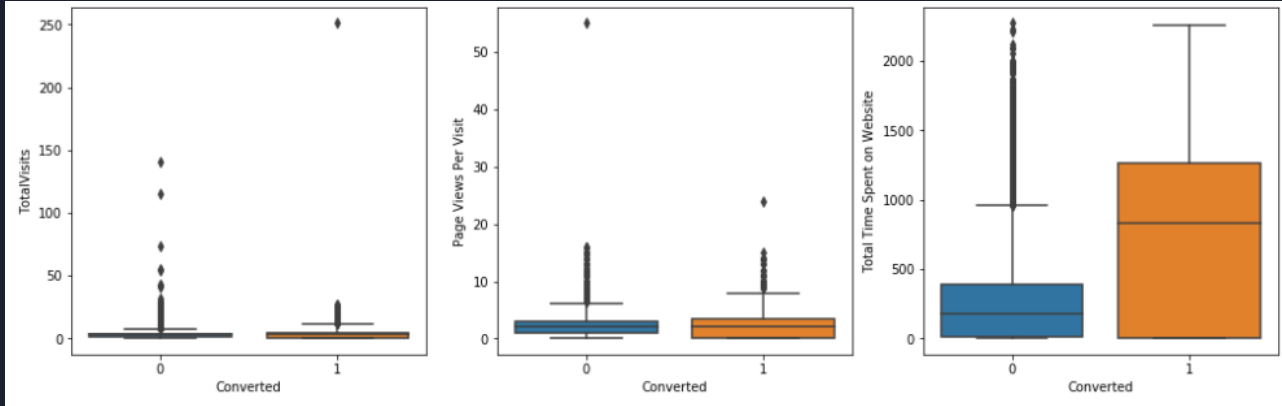
- 'Landing Page Submission' leads to highest conversion



- 'SMS Sent' leads to highest conversion

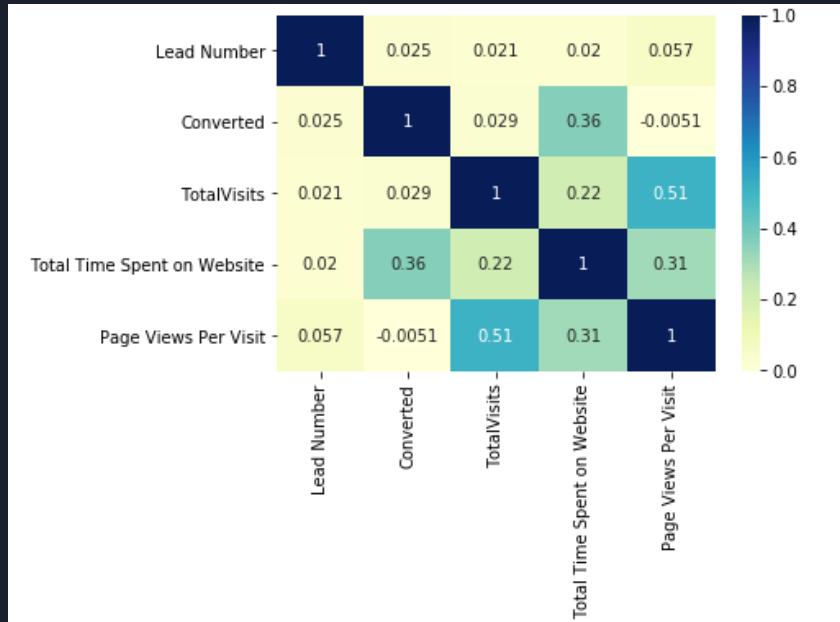


- Leads that don't want email regarding the course get converted the most
- Customers did not see any ad in 'Search' and customers did not come to website through recommendations
- Customers that do not want 'A free copy of Mastering The Interview' convert the most

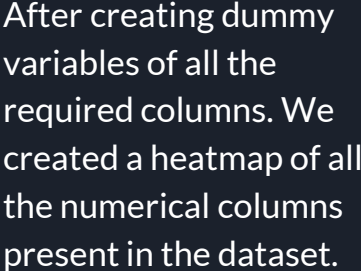


- Customers who spend more time on website tends to convert more

# Data Preparation

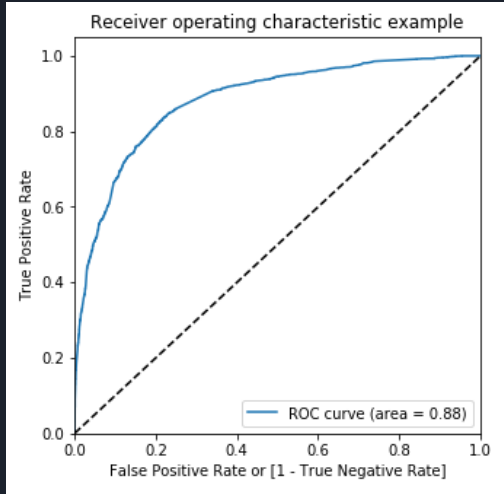


- Highest correlation can be seen between TotalVisits and Page Views Per Visit of 0.51
- Second highest correlation can be seen between Converted and Total Time Spend on Website of 0.36
- Lowest correlation can be seen between Page Views Per Visit and Converted of -0.0051



After creating dummy variables of all the required columns. We created a heatmap of all the numerical columns present in the dataset.

# Prerequisite before Final Model



Area under ROC = 0.88

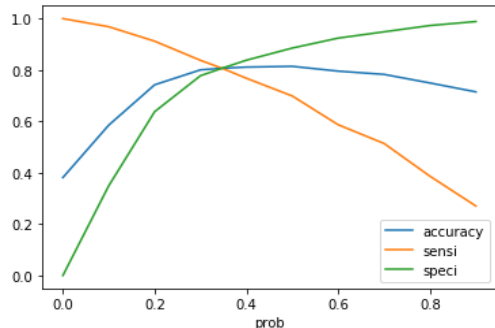
An ROC curve demonstrates several things:

- 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.

The ROC curve shows the tradeoff between True Positive Rate and False Positive Rate which essentially can also be viewed as a tradeoff between Sensitivity and Specificity. As you can see, on the Y-axis, you have the values of Sensitivity and on the X-axis, you have the value of (1 - Specificity). Notice that in the curve when Sensitivity is increasing, (1 - Specificity) is increasing, it simply means that Specificity is decreasing.



```
# Let's plot accuracy sensitivity and specificity for various probabilities.
cutoff_df.plot.line(x='prob', y=['accuracy', 'sensi', 'speci'])
plt.show()
```

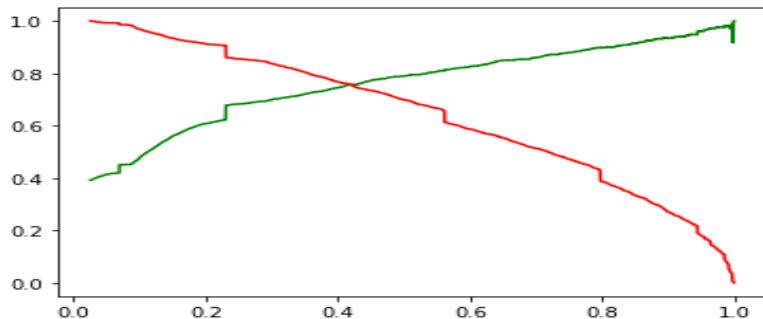


## Optimal Cutoff Point:

It show the Cutoff Point between all the three accuracy, sensitivity and specificity.

So the optimal cutoff point is 0.35.

```
plt.plot(thresholds, p[:-1], "g-") #precision
plt.plot(thresholds, r[:-1], "r-") #recall
plt.show()
```



Precision - Also called positive predictive value. It is probability that a predicted 'Yes' is actually 'Yes'

Recall - is exactly same as sensitivity. It is probability that an actual 'Yes' case is predicted correctly



# Inferences from Model

- After running the model on the Train and Test data we have these final observations:

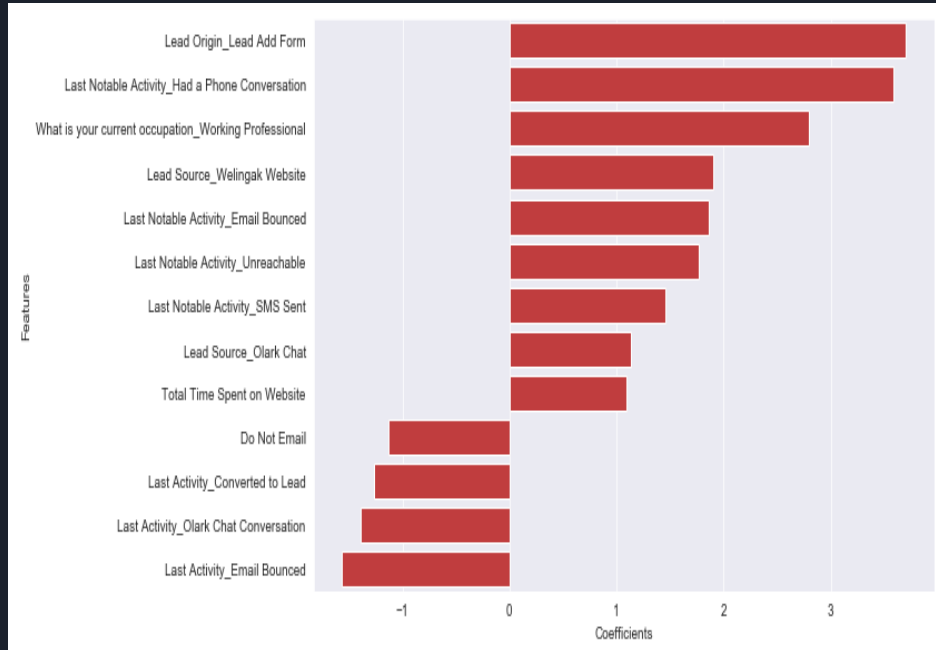
## Train Data:

- - Accuracy: 80.71%
- - Sensitivity: 80.45%
- - Specificity: 80.86%
- - Precision: 78.96%
- - Recall: 69.87%

## Test Data:

- - Accuracy: 81.64%
- - Sensitivity: 80.55%
- - Specificity: 82.35%
- - Precision: 74.87%
- - Recall: 80.55%

# Inferences from Model



Top 3 variables in Model, that contributes towards lead conversion are:

- Lead Origin\_Lead Add Form
- Last Notable Activity\_Had a Phone Conversation
- What is your current occupation\_Woring Professional

Top 3 variables in Model, that should be focused are:

- Last Activity\_Email Bounced
- Last Activity\_Olark Chat Conversation
- Last Activity\_Converted to Lead



# Conclusions

The Model seems to predict the Conversion Rate within our required range and we can give the CEO confidence for making good calls on the basis of this model.

- CEO should focus most on Lead Origin\_Lead Add Form, Last Notable Activity\_Had a Phone Conversation and What is your current occupation\_Woring Professional the conversion rates are high for all the three.
- CEO should focus less on Last Activity\_Email Bounced, Last Activity\_Olark Chat Conversation, Last Activity\_Converted to Lead and Do not Email as the conversion rates are low for all the three



# Recommendations

- By using a low threshold value for Conversion rate, Sensitivity value will be very high, which will make sure that almost all leads who are likely to convert are identified correctly and the sales team of X Education can make phone calls to as much of people as possible and complete their quarter target before the deadline and will have spare time to work on other fields
- When the target is achieved, by using the high threshold value for Conversion rate, Specificity value will be very high which will make sure that all the leads who were at the edge of getting converted or are not selected. By this sales teams will not have to make unnecessary phone calls after the target is achieved.