Lead Scoring Assignment

Pallavi Vippagunta | Shalini Dixit

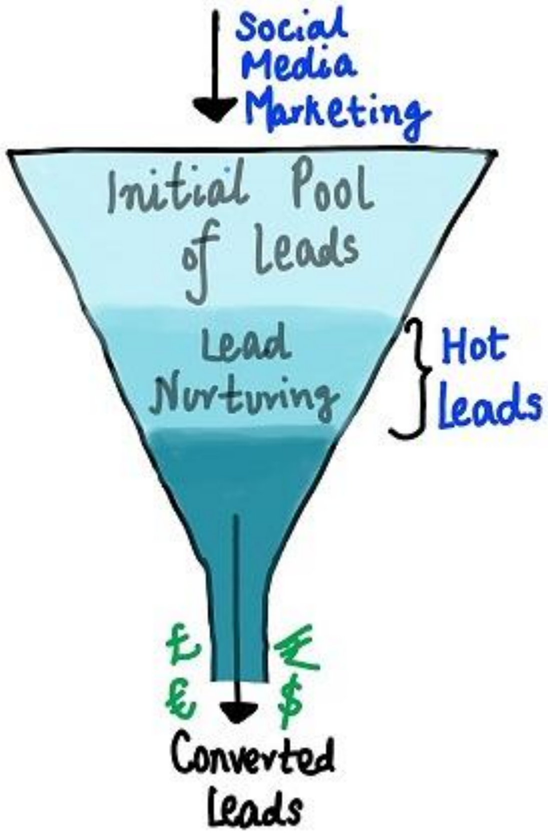
14th June , 2021

Problem Statement

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. Moreover, 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%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone. A typical lead conversion process can be represented using the following funnel:



As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as 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.

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Approach to analysis

1. **Understand what is Lead Scoring and its benefits**
2. **Reading & Understanding the Leads dataset**

- Import the Leads.csv data using pandas
- Study the 1st five and last five rows of the complete data
- Check the shape , datatypes and statistical summary of the data using shape , info and describe
- Check for null value columns and their percentage

3. **Data Cleaning and Preparation**

- Treating null values
- Dropping columns with only 1 unique/distinct value
- Dropping irrelevant columns that will not add any value to the data analysis
- Replacing values reflecting “Select” with “nan” in 4 columns – Specialization , Lead Profile , City & How did you head about X Education
- Dropping columns with null values >50%
- Dropping Prospect Id and Lead Number as they are unique for each lead
- Grouping values wherever needed to reduce the number of unique values

4. **Exploratory Data Analysis (EDA)**

- Check existing conversion rate (it comes out to be 38%)
- Encoding binary variables
- Create Dummy variables for categorical columns

5. **Split the data into Test- train sets (Test – 30%, Train-70%)**
6. **Feature Scaling**
7. **Model Building**
8. **Feature Selection using RFE (Recursive Feature Extraction)**

- Running RFE with 15 variables as output

9. **Plotting the ROC Curve**
10. **Finding the Optimal Cut-Off Point**

- Plotting Accuracy, Senstivity and Specificity to see the tradeoff
- Confusion Matrix
- Calculate False Positivity Rate , Positive & Negative Predictive value ,Precision & Recall value

11. **Making predictions on the Test Set**
12. **Compare the outputs of Train & Test sets & make inferences**
13. **Calculate Lead Scores for each Lead**

What is Lead Scoring

Lead scoring is a way of ranking prospects to determine their sales-readiness. Each lead is assigned a point value. These points are based on particular attributes and behaviors that show that the lead is ready to become a customer in the near future. The higher the lead score, the more likely they'll buy.

For instance, a lead that subscribed to your email list then followed you on social media AND visited the pricing page on your website would have a higher lead score than a lead who hasn't done any of those things.
This process can help sales and marketing teams identify which leads they should be focusing on. By focusing on the leads that are most ready to buy, businesses can easily increase conversions by turning leads into opportunities

Benefits of Lead Scoring

Fewer Lost Opportunities

Increased Conversions

Improved Productivity

Better Marketing & Sales Alignment

Higher ROI

Data Understanding

Target Variable " Converted" - 38% of the data is Converted Leads

Shape of the Data – Leads Data has 9240 rows and 37 columns

Column Datatypes – 30 columns are categorical , 4 float and 3 int types

Statistical Summary – From the describe function we can see that The column "Total Visits" & "Total Time Spent on website" seem to have outliers.

Null Values – Following columns have null values – Total Visits, Page Views per visit ,Last Activity , Country, Specialization ,How did you hear about X Education ,What is your current occupation ,What matters most to you in choose a course ,Tags , Lead Quality ,Lead Profile , City, Asymmetrique Activity Index , Asymmetrique Profile Index, Asymmetrique Activity Score, Asymmetrique Profile Score

Columns with "Select" as values - Specialization, How did you hear about X Education, Lead Profile,City .

Data Cleaning & Preparation

Columns with “Select” as values have been replaced with nan - Specialization, How did you hear about X Education, Lead Profile, City

Dropping Columns :

- Dropped columns with only 1 unique/ distinct value – Magazine’, 'Receive More Updates About Our Courses’, 'I agree to pay the amount through cheque’, 'Get updates on DM Content’, 'Update me on Supply Chain Content'
- Dropped columns with >50% null values - 'Lead Profile','Lead Quality','How did you hear about X Education’
- Dropped columns that have been inouted by Sales & will not add any value to the analysis – “Tags”, “Notable Activity”, “Last Notable Activity” ,'Asymmetrique Activity Index’, 'Asymmetrique Activity Score’, 'Asymmetrique Profile Index’, 'Asymmetrique Profile Score’
- Dropped “ Prospect ID and Lead number as they are unique to each lead.
- Dropped “City”

Merging all values =<7 into one as “Outside India”

Imputing columns with null values – Filling null values with “not provided” in the following columns: Specialization , What matters most to you in choosing a course , Country , Tags ,What is your current occupation

Data Cleaning & Preparation

Checking the value counts for certain columns and imputing:

- Lead Source – Converted all the values <6% to “Other Sources “ so as have better analysis by reducing the number of dummy variables
- Country – Imputed the Country column so as to have just 3 distinct values in it - India , not provided and outside India

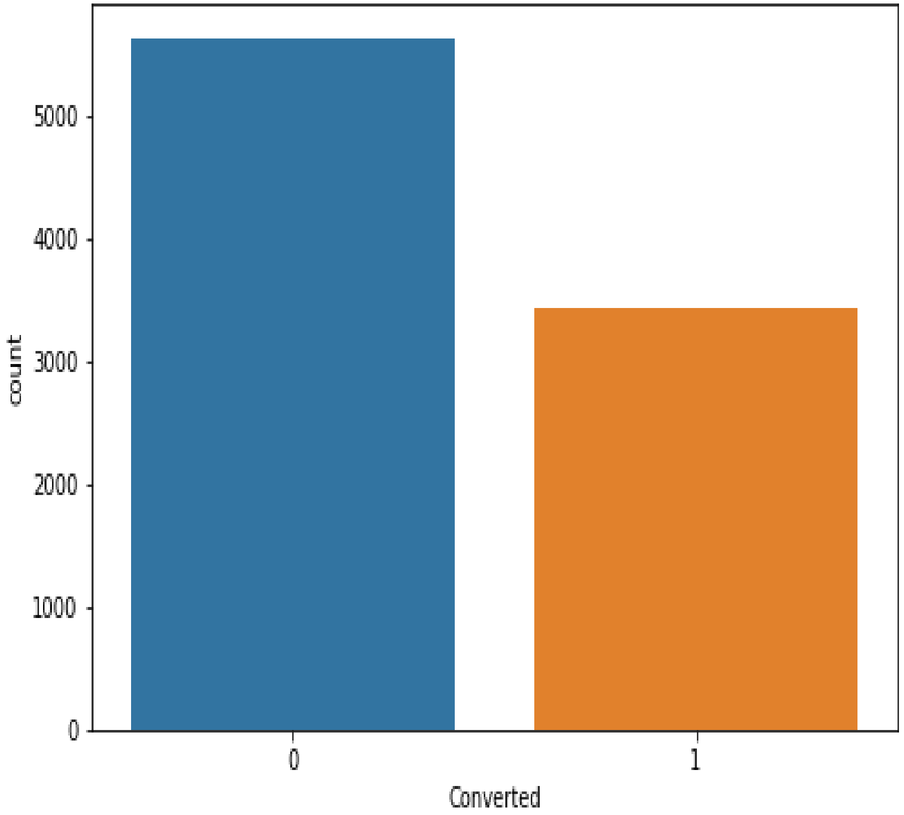
Outlier detection and treatment : TotalVisits & Page Views Per Visit have outliers. We have capped them at 99th percentile , post which imputed the null values with mean value

Converting some binary variables(Yes/No) to 1/0 – The following columns were subjected to encoding : "Do Not Email", "Do Not Call", "Search", "Newspaper Article", "X Education Forums", "Newspaper", "Digital Advertisement", "Through Recommendations", "A free copy of Mastering The Interview"

Dummy Value creation for categorical columns – The following columns with 3 or more distinct values were subjected to dummy value creation - 'What matters most to you in choosing a course’, 'Specialization’, 'What is your current occupation’, 'Country’, 'Lead Source’, 'Lead Origin']

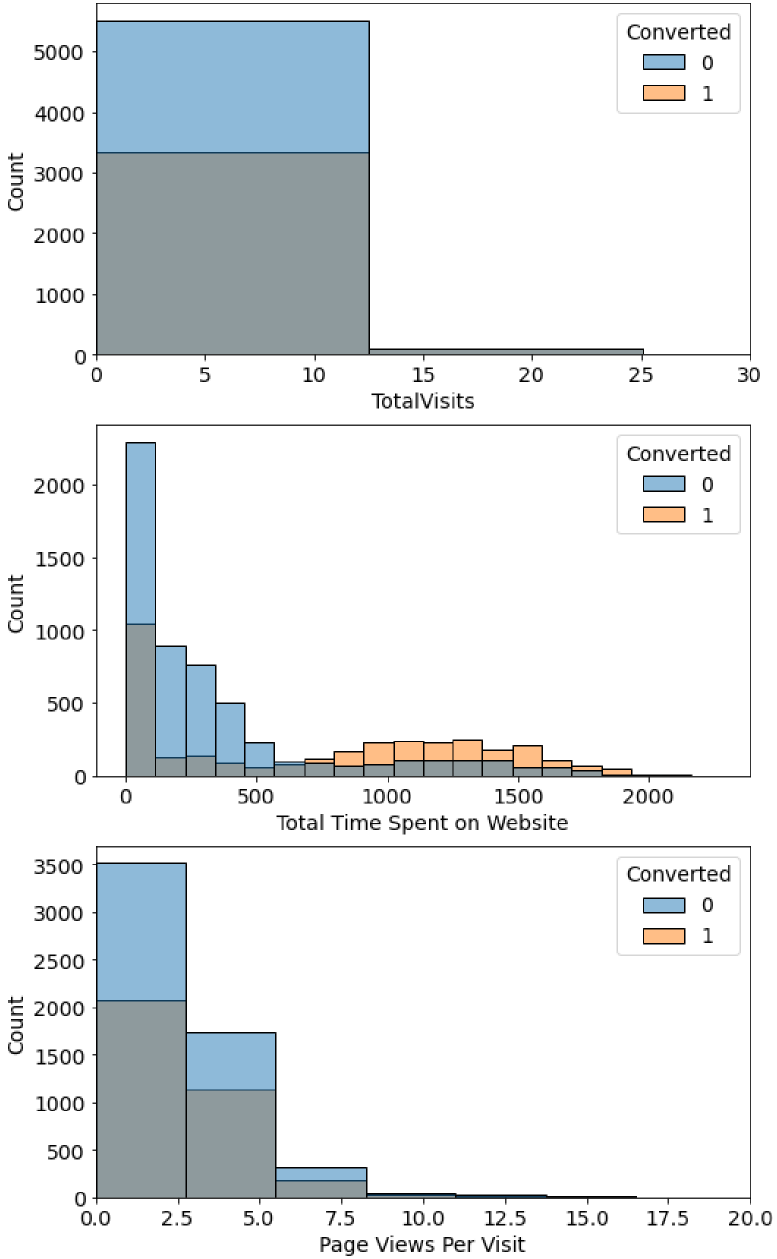
Exploratory Data Analysis

OVERALL CONVERSION RATE

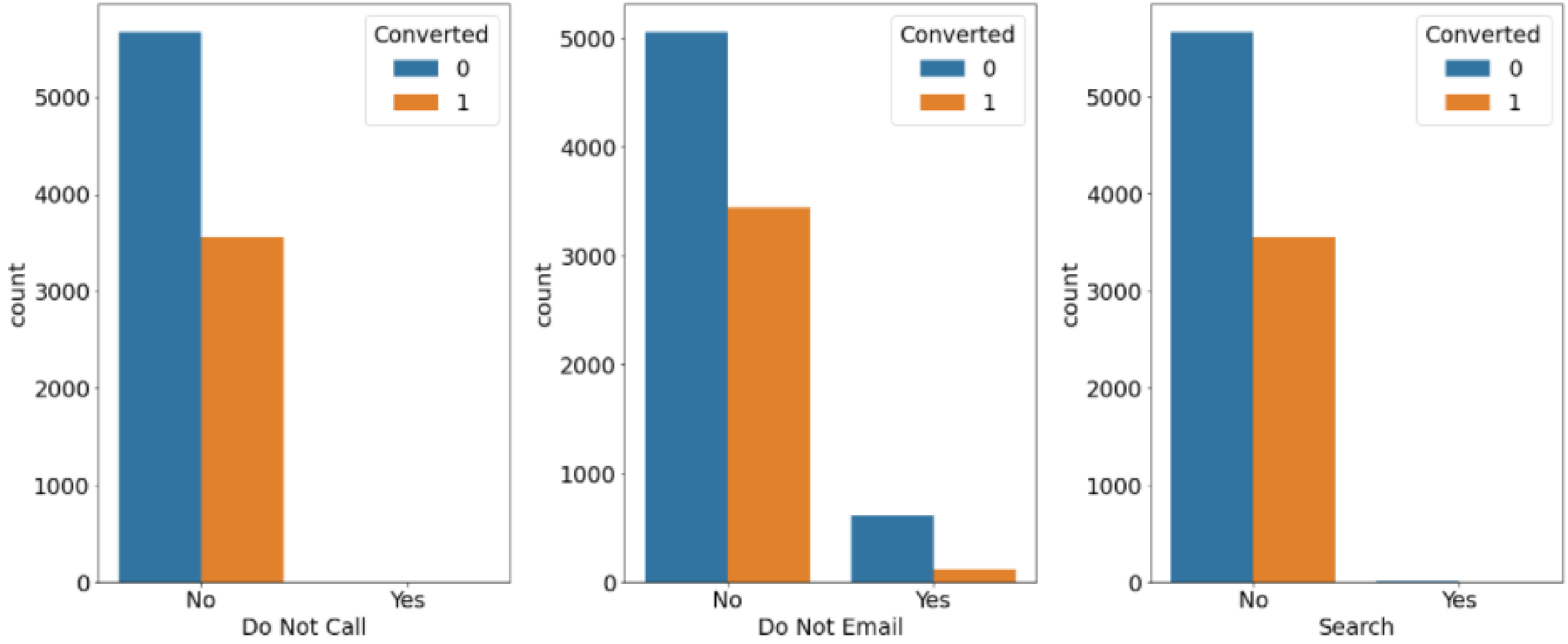


Overall Conversion Rate of the Dataset is around 38% as can be seen from the Countplot.
Leads with higher number of Total Visits , Total Time Spent on Website & Page Views Per Visit are having higher conversion

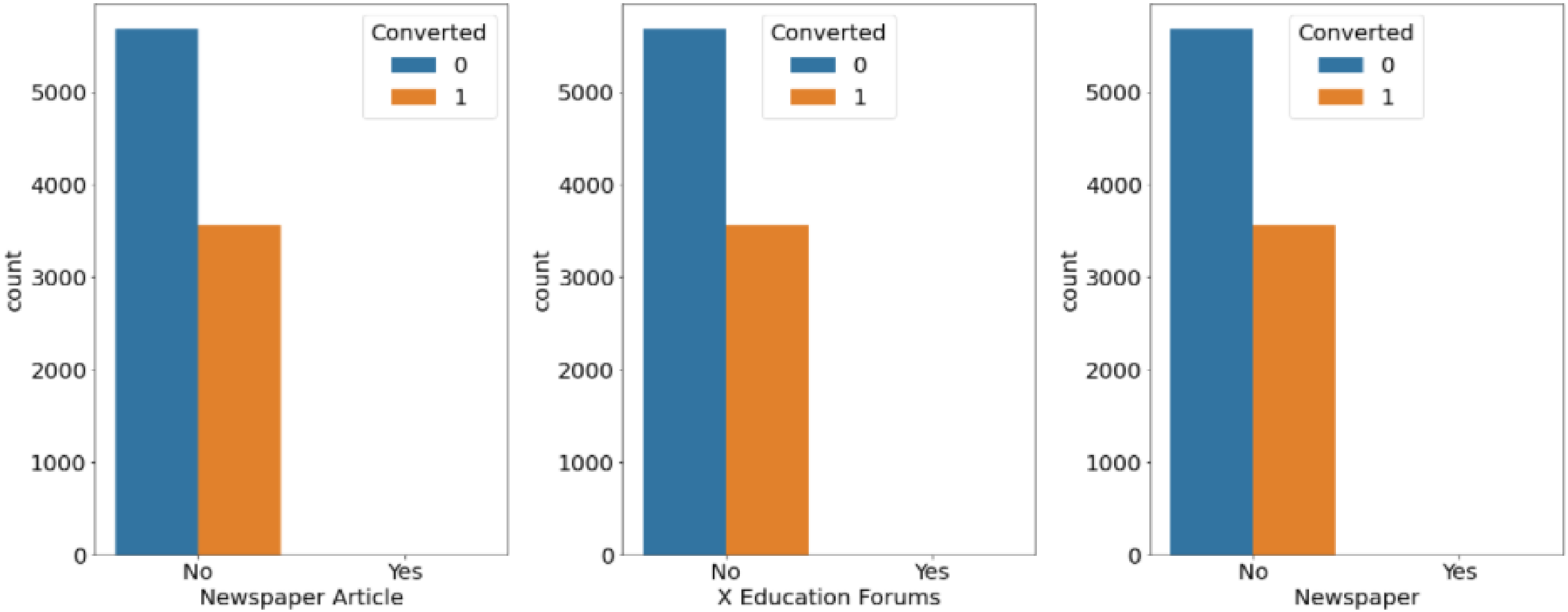
VARIABLES WITH HIGHER CONVERSION RATE



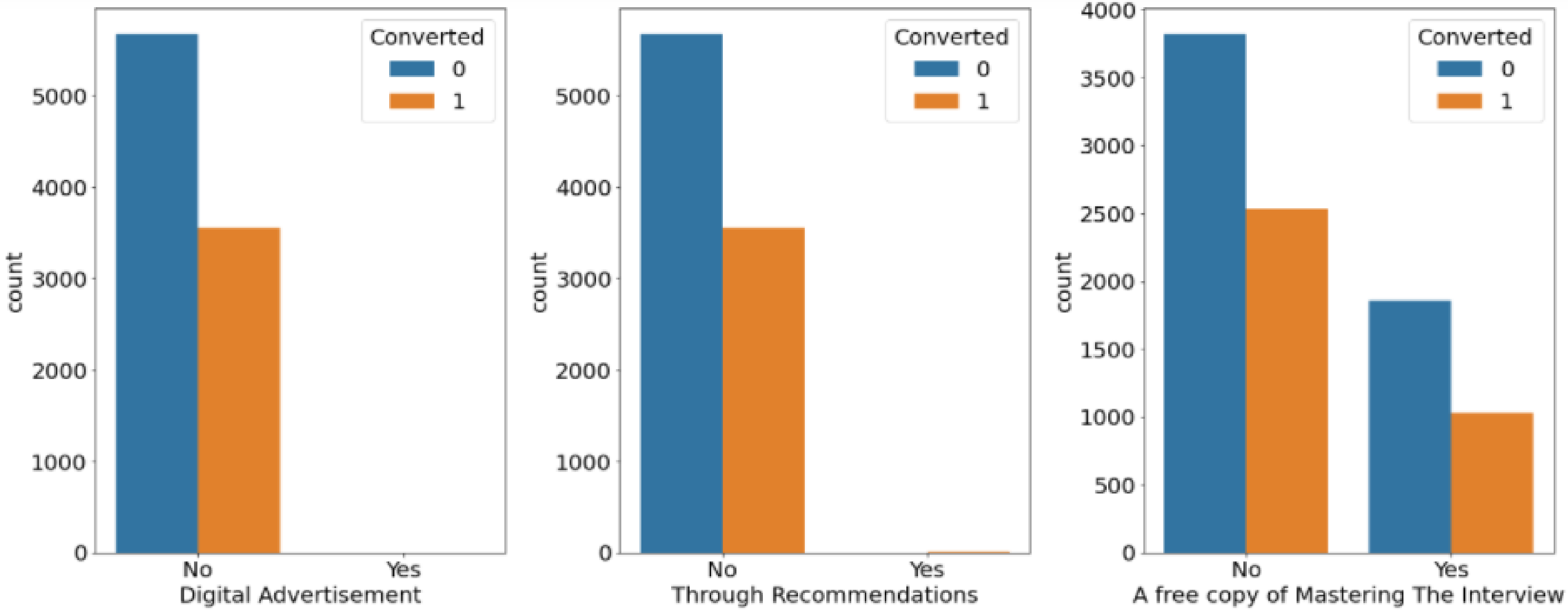
Plots showing target variable – converted against the variables : Do Not Call, Do Not Email and Search



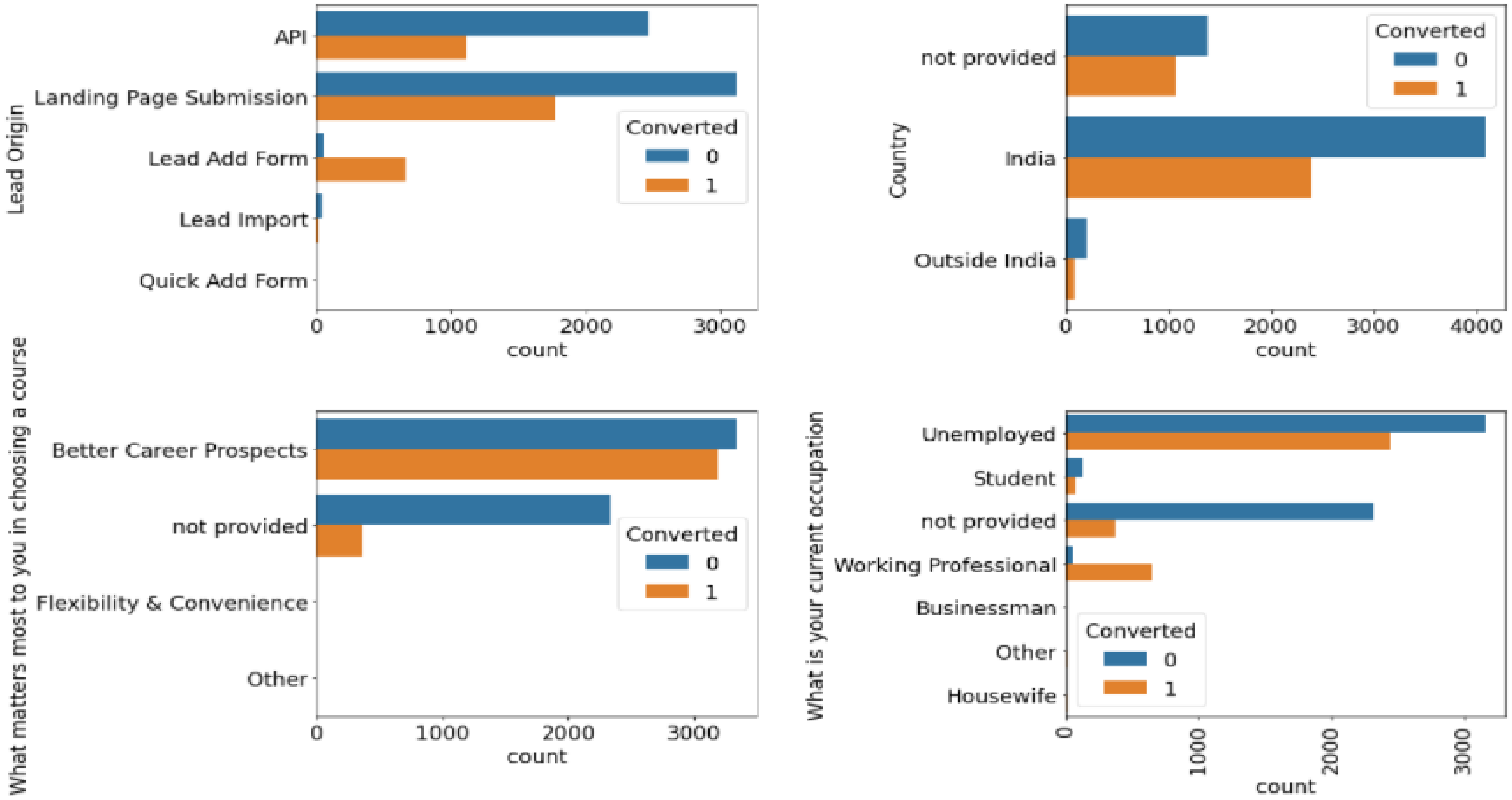
Plots showing target variable – converted against the variables : Newspaper Article, X Education Forums, Newspaper



Plots showing target variable – converted against the variables : Digital Advertisement, Through Recommendations,
A free copy of Mastering The Interview

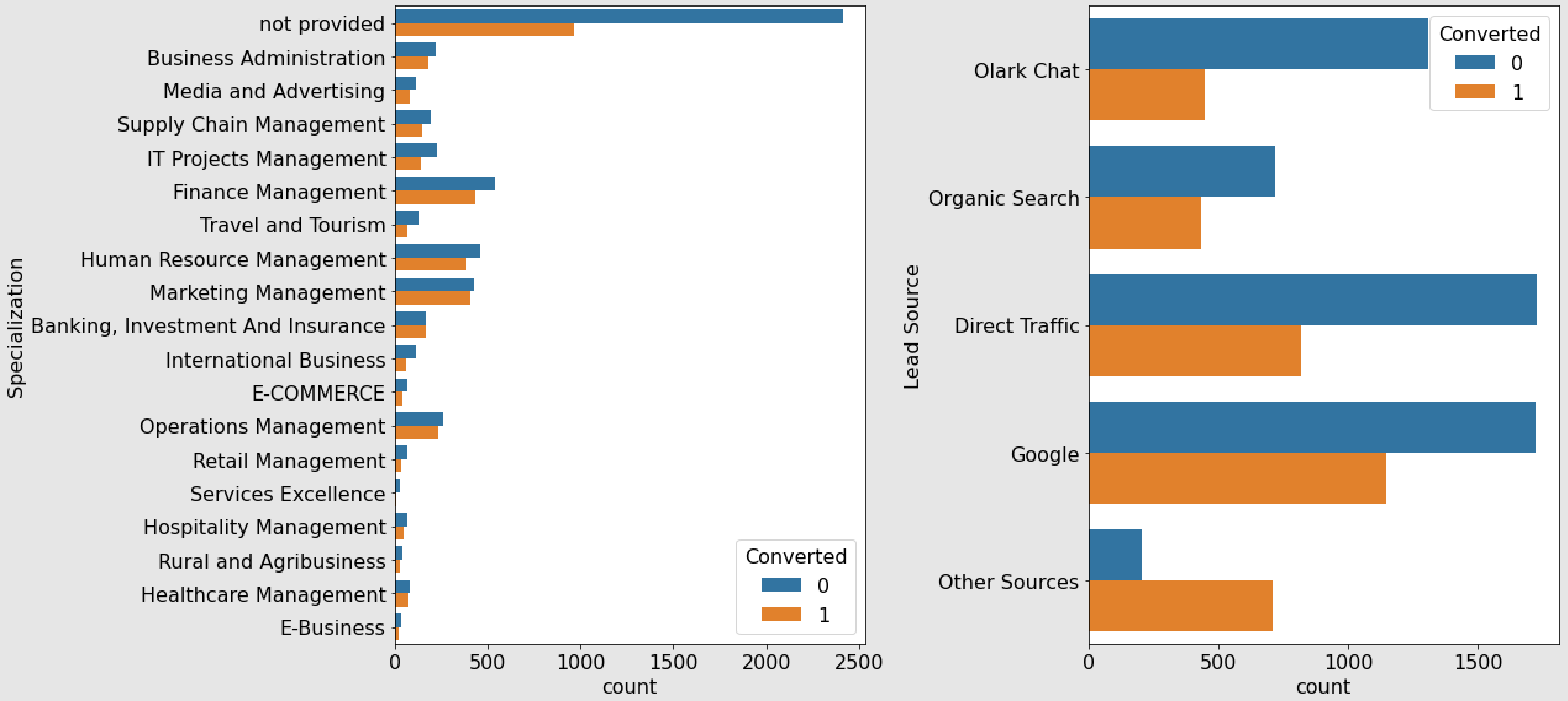


Plots showing target variable – converted against the variables : Lead Origin, What matters most to you in choosing a course,
Country, What is your current Occupation



Leads who are working professionals, had filled up Lead Add Form and looking for better career prospects contribute most of the converted leads

Plots showing target variable – converted against the variables : Specialization and Lead Source



Feature Selection with RFE

Step 1:
The entire data was split into train and test in the ration 7:3.

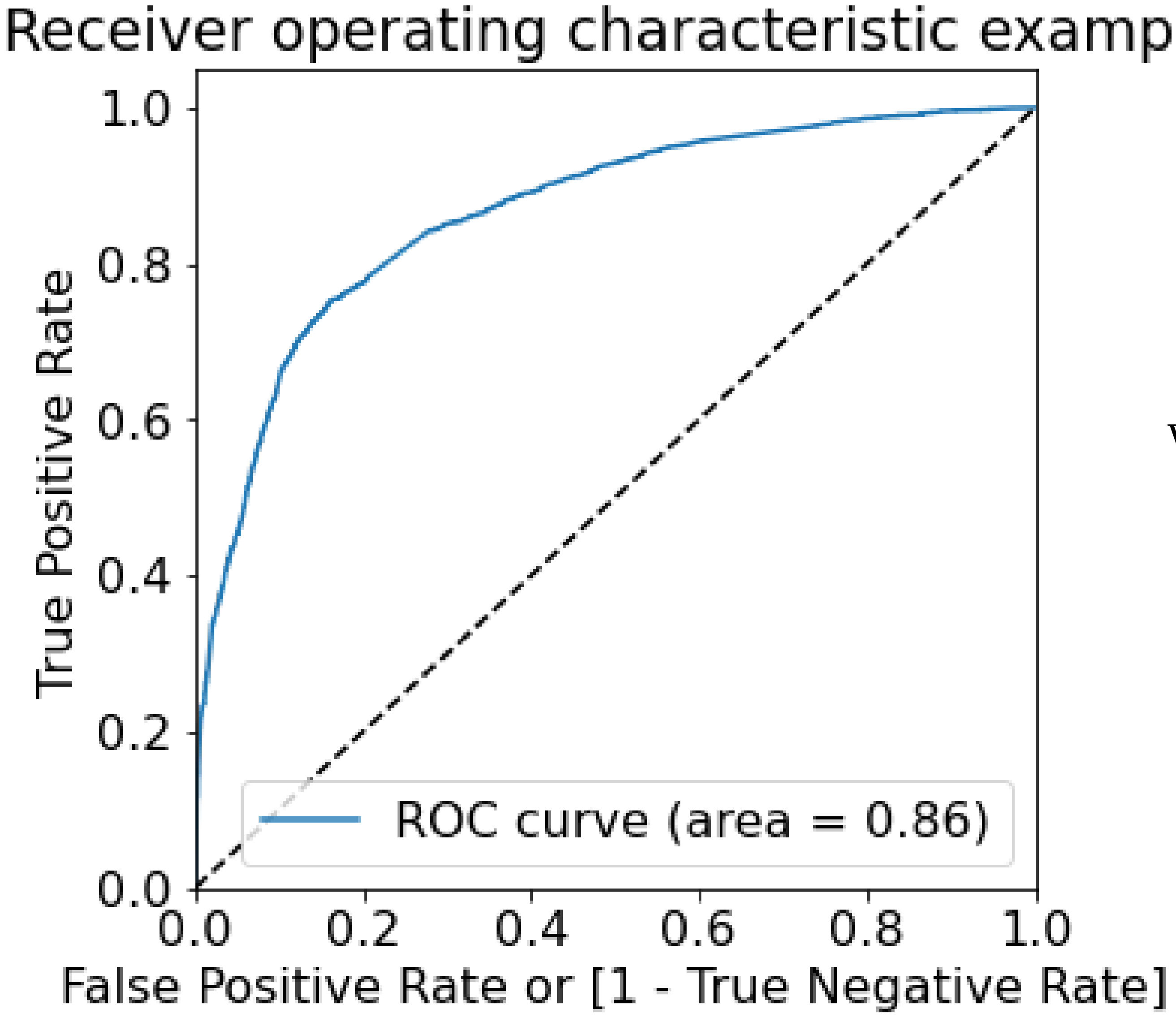
Step2:
Feature scaling is done by StandardScaler

Step3:
15 columns were selected through the automated RFE method.

Step4:
Model building with automated RFE and manual elimination based on p-value and VIF

Do Not Email	What is your current occupation_Housewife
Total Time Spent on Webiste	What is your current occupation_Working Professional
Newspaper	Country_not provided
Specialization_Hospitality Management	What is your current occupation_Other
What is your current occupation_not provided	Lead Origin_Landing Page Submission
What matters most to you in choosing a course_not provided	Lead Origin_Lead Add Form
Specialization_not provided	What is your current occupation_Student
Lead Origin_Lead Import	

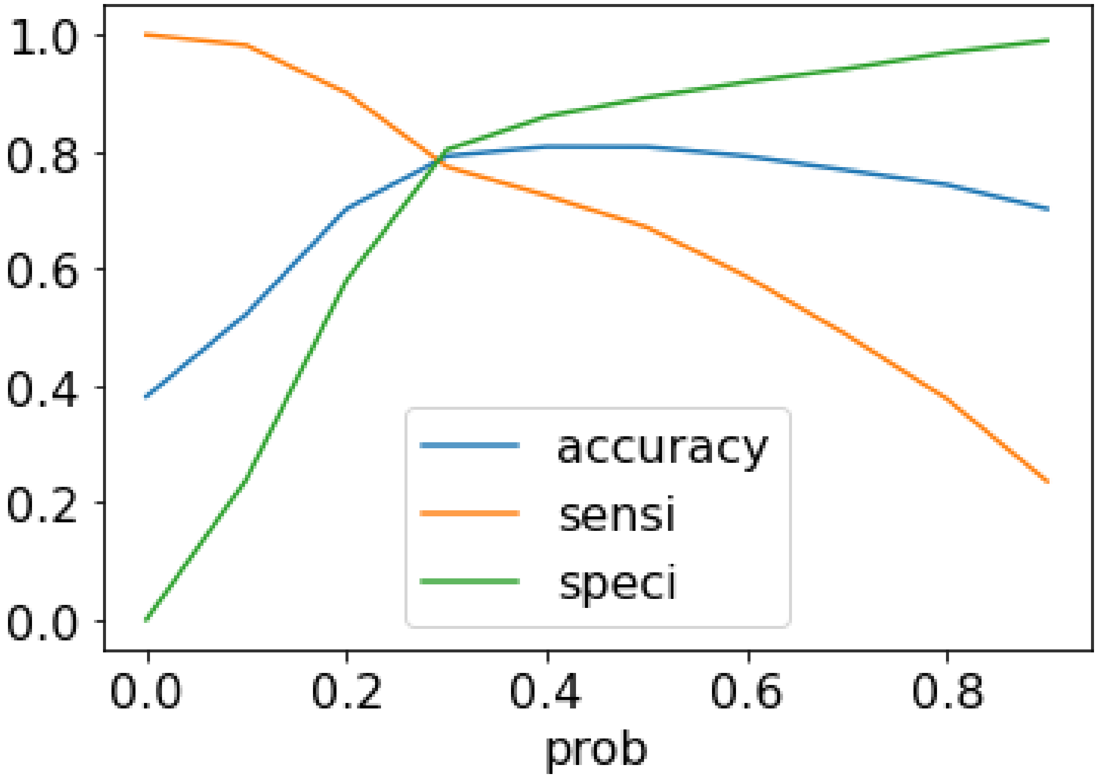
ROC Curve



With ROC Area Under The Curve as .86 , our model seems to be doing a good job

Model Evaluation – Train Set

OPTIMAL CUT-OFF



Optimal cutoff probability is that prob where we get balanced sensitivity and specificity.The Optimal Cut-Off Probability is .285 as we can see from the above chart

CONFUSION MATRIX

Confusion Matrix

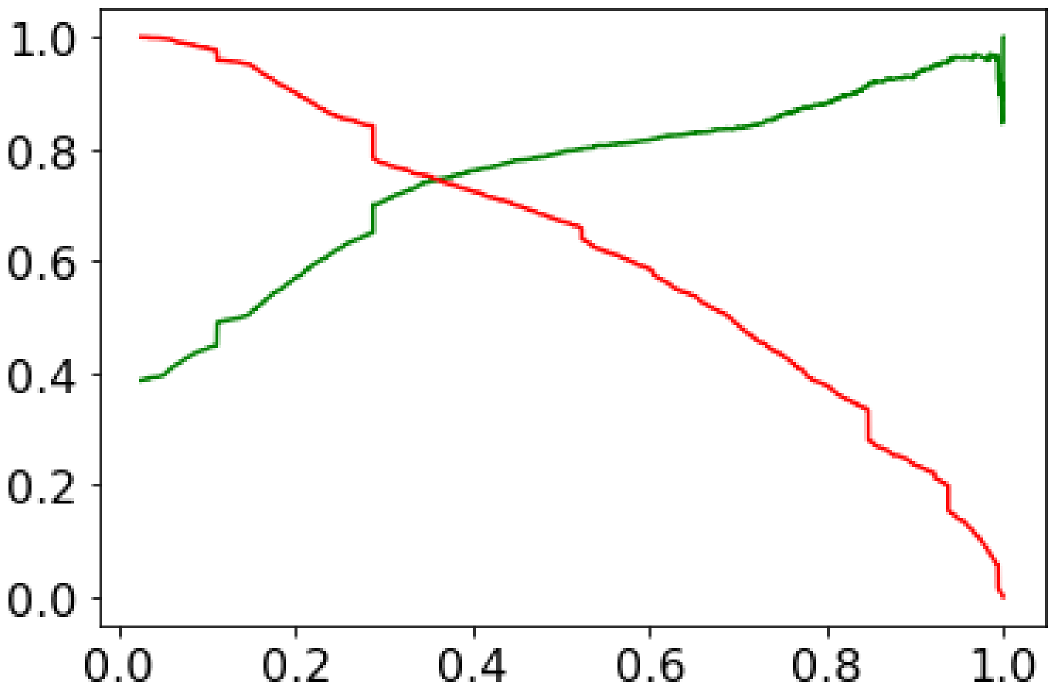
2881	1121
389	2077

Accuracy – **76.65%**
Senstivity – **84 %**
Specificity – **72 %**
False Positive Rate – **28 %**
Positive Predictive Value – **65 %**
Negative Predictive Value – **88 %**

Sensitivity metrics of 84% on the train set is a very good indicator

Model Evaluation –
Precision & Recall (Train Set)

PRECISION & RECALL TRADEOFF CURVE



The Green line reflects Precision and the Red line reflects the recall for the model

CONFUSION MATRIX

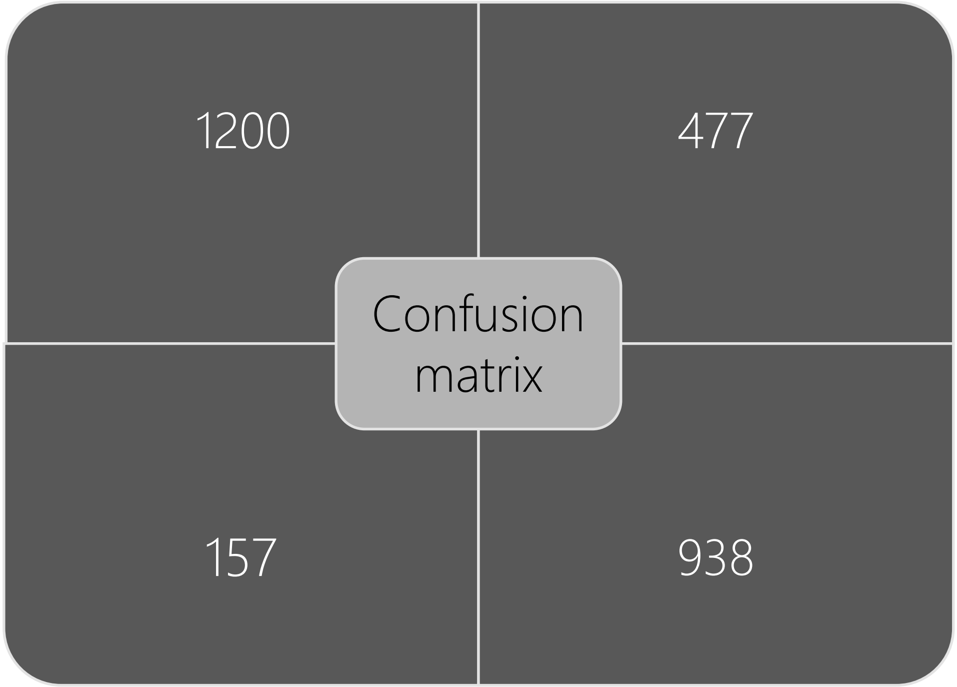
Confusion Matrix

2881	1121
389	2077

Precision Score – **65%**
Recall Score – **84%**

Model Evaluation –
(Test Set)

CONFUSION MATRIX FOR THE TEST SET



Confusion matrix

1200	477
157	938

Accuracy – **77 %**
Senstivity – **85.66 %**
Specificity – **71.55 %**
Precision **- 66 %**
Recall **- 85.66 %**

The test set has managed to manitain a similar senstivity% as the train set which is very good

Metrics comparison between Train & Test datasets

The metrics of Accuracy , senstivity and specificity seem to hold well on the test dataset as well

	Train Set	Test Set
Senstivity	84%	85.66%
Accuracy	76.65%	77%
Specificity	72%	71.55%
Precision	65%	66%
Recall	84%	85.66

Inferences form the Case Study

The following are the top three variables that contribute most towards the probability of the lead getting converted:

- Lead origin with Lead add form which has coefficient of 2.6194.

- What is your current occupation with working professional has a coefficient of 2.3834

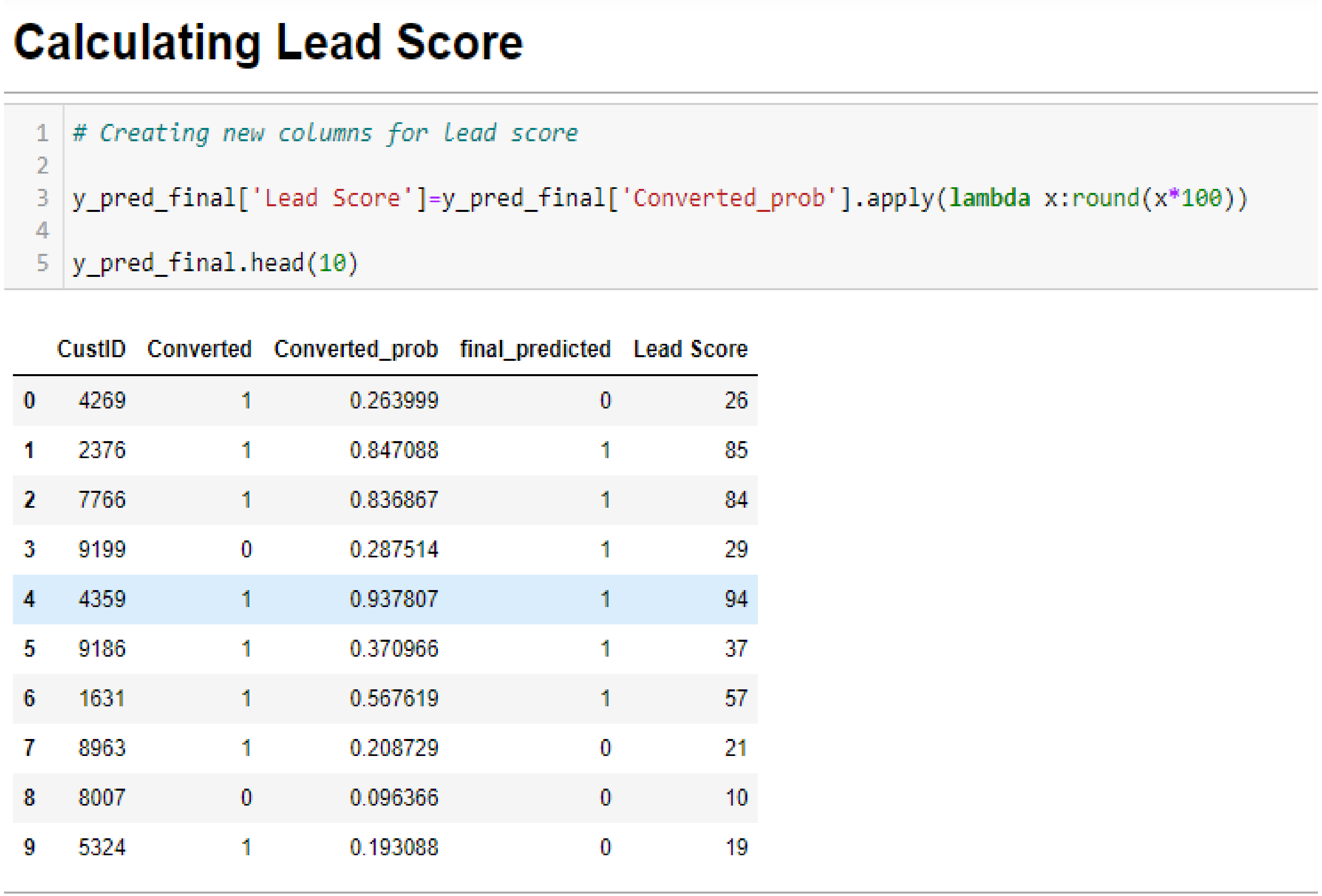
- Total Time Spent on Website has a coefficient of 1.0906.

The above clearly shows that the conversion rate is high when the customer fills the lead add form, is a working professional and has spent high amount of time on company's website.

Goals Of the Case Study & their Outcome

1^{st} 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.

Outcome : As we can see , the leads have been allocated scores between 0-100. Leads with higher score are the ones that are also predicted as converted. And sales team needs to focus their energies on these high scoring leads . Marketing can further nurture leads these leads so that conversion % goes up



Calculating Lead Score

```
# Creating new columns for lead score

y_pred_final['Lead Score']=y_pred_final['Converted_prob'].apply(lambda x:round(x*100))

y_pred_final.head(10)
```

	CustID	Converted	Converted_prob	final_predicted	Lead Score
0	4269	1	0.263999	0	26
1	2376	1	0.847088	1	85
2	7766	1	0.836867	1	84
3	9199	0	0.287514	1	29
4	4359	1	0.937807	1	94
5	9186	1	0.370966	1	37
6	1631	1	0.567619	1	57
7	8963	1	0.208729	0	21
8	8007	0	0.096366	0	10
9	5324	1	0.193088	0	19

Recommendations to X Education

2nd Goal : There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

1. **X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So, during this phase, they wish to make the lead conversion more aggressive. So, they want almost all the potential leads (i.e., the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. Suggest a good strategy they should employ at this stage.**

Ans: The strategy that can be applied to get maximum conversions with the help of sales team and the 10 short term hired interns is as follows:

- Segregate the entire Leads list with high scores (closer to 100) The minimum cut off lead score can be decided by the marketing head basis the Lead that scores high on critical parameters such as: Time Spent on Website , Lead being a working professional and client has filled the Lead form
- Allocate the segregated high score leads to Sales.
- Training to be imparted to the newly hired interns on the product/service being sold (in this case several education courses provided by X Education.
- Sales Head to allocate leads per salesperson and intern giving them a daily, weekly, monthly target and monitor performance.
- Sales team and interns to utilize multiple channels to connect with the customer and nurture them at every step – telephonically, through email, SMS, conduct online webinar for a group of these high conversion probability leads etc.

Recommendations to X Education

2. Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So, during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e., they want to minimize the rate of useless phone calls. Suggest a strategy they should employ at this stage.

Ans: The sales team and the marketing team jointly can address this scenario by taking the following steps:

- Describe customer personas and identify the stage of the customer in the customer life cycle. This will enable create customized messaging.
- Automated SMS's and Emails can be sent to customers with links and updates on existing courses and upcoming ones.
- Company should create a mobile responsive website if not there already.
- Website should be interactive and user friendly for the customer to navigate through the pages.
- Website can have a click to call back tab so that the salesperson/ intern can call back the customer.

Thank You