Lead Score Case study

Submitted by Gagan Shukla

Lead score study for x education

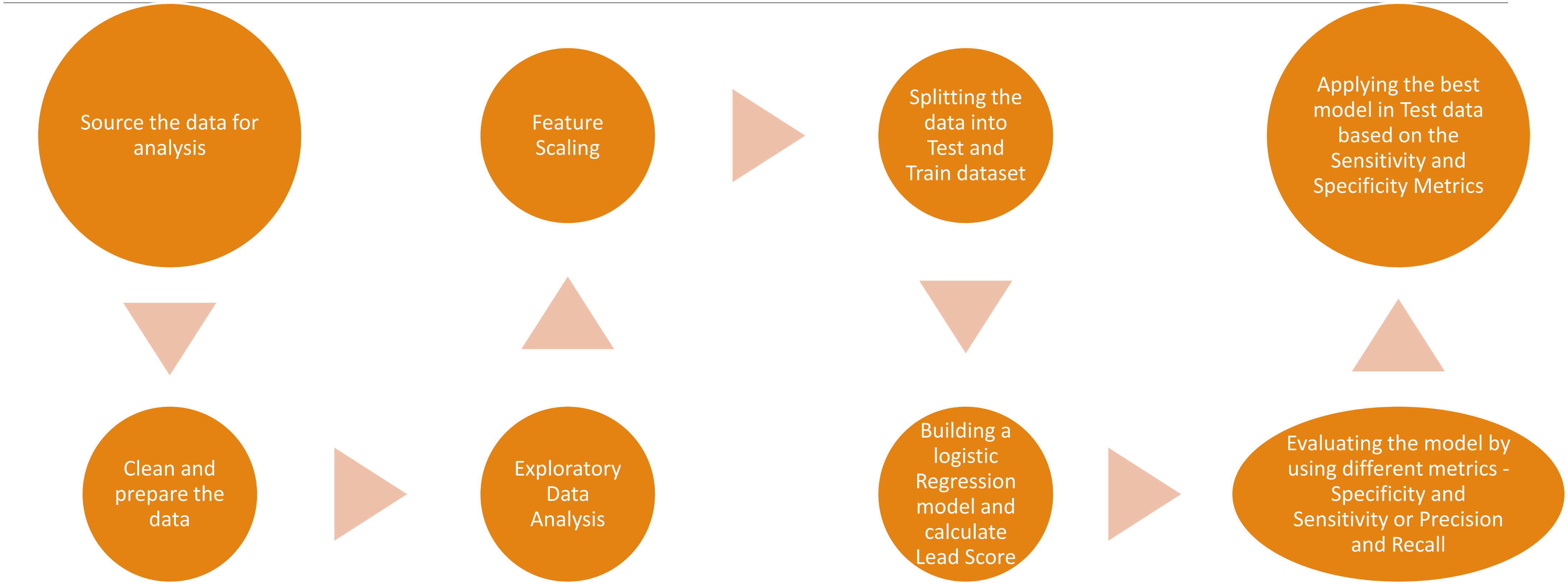
Problem Statement:

An education company named X Education sells online courses to industry professionals. Many professionals who are interested in the courses of land on their website and browse for courses. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. The company markets its courses on several websites and search engines like Google platform. When these people fill up a form providing their email address or phone number, they marked to be a lead. The company also gets leads through past referrals. Once these leads are potential, employees from the sales team start making calls, writing emails & connected with them via different channels etc. This process of leads get converted while most do not. The typical lead conversion rate at X education is around 30%. There are a lot of leads generated in the initial stage, but only a few of them come out as paying customers. In the middle stage, you need to nurture the potential leads well (i.e. constantly communicating, educating the leads about the product 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 has given a ballpark of the target lead conversion rate to be around 80%.

Goals of the Case Study:

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. 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. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Strategy



Source the data for analysis

Clean and prepare the data

Exploratory Data Analysis

Feature Scaling

Splitting the data into Test and Train dataset

Building a logistic Regression model and calculate Lead Score

Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall

Applying the best model in Test data based on the Sensitivity and Specificity Metrics

Problem Solving Methodology

Data Sourcing, Cleaning and Preparation

- Read the Data from Source
- Convert data into clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization

Feature Scalingand Splitting Train and Test Sets

- Feature Scaling of Numeric data
- Splitting data in to train and test set.

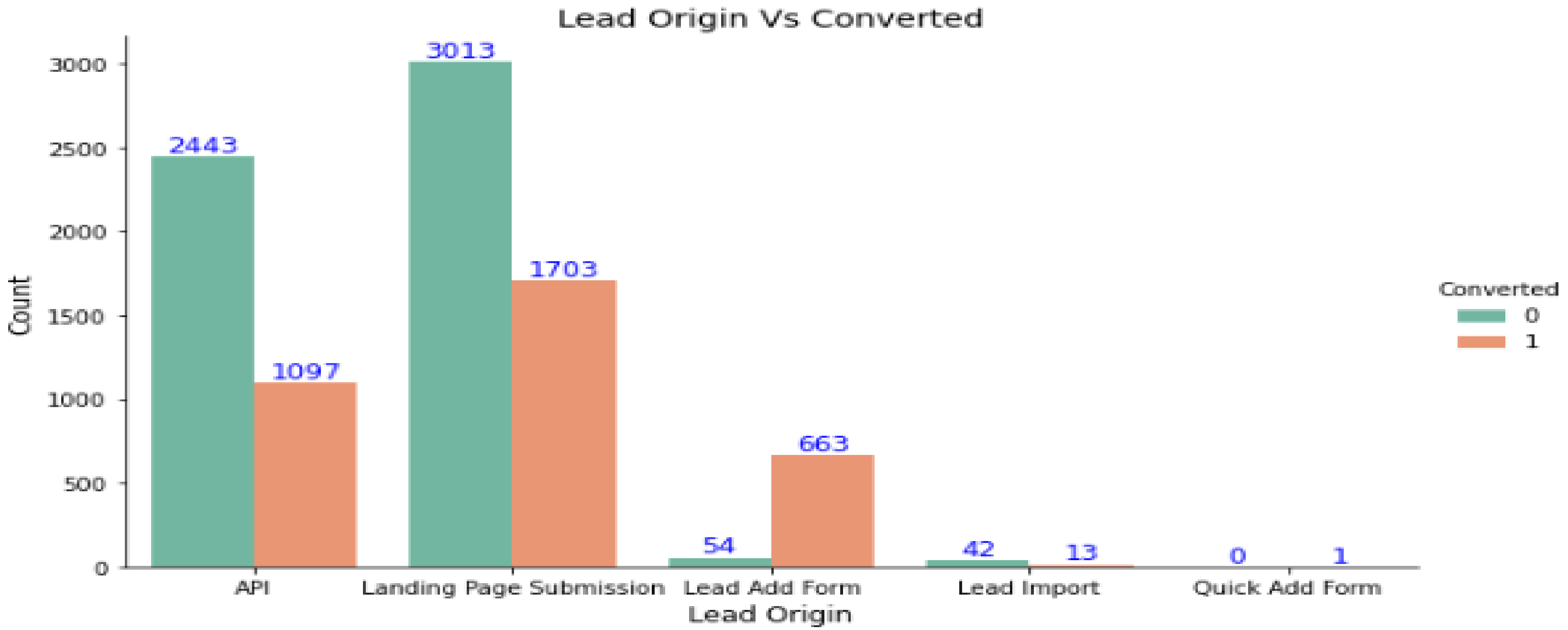
Model Building

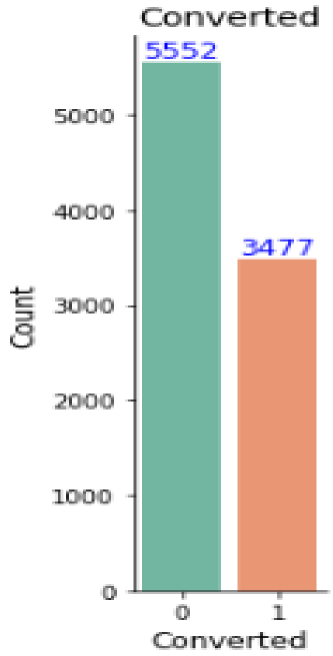
- Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model

Outcome

- Determine the lead score and check if target final predictions amounts to 80% conversion rate
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

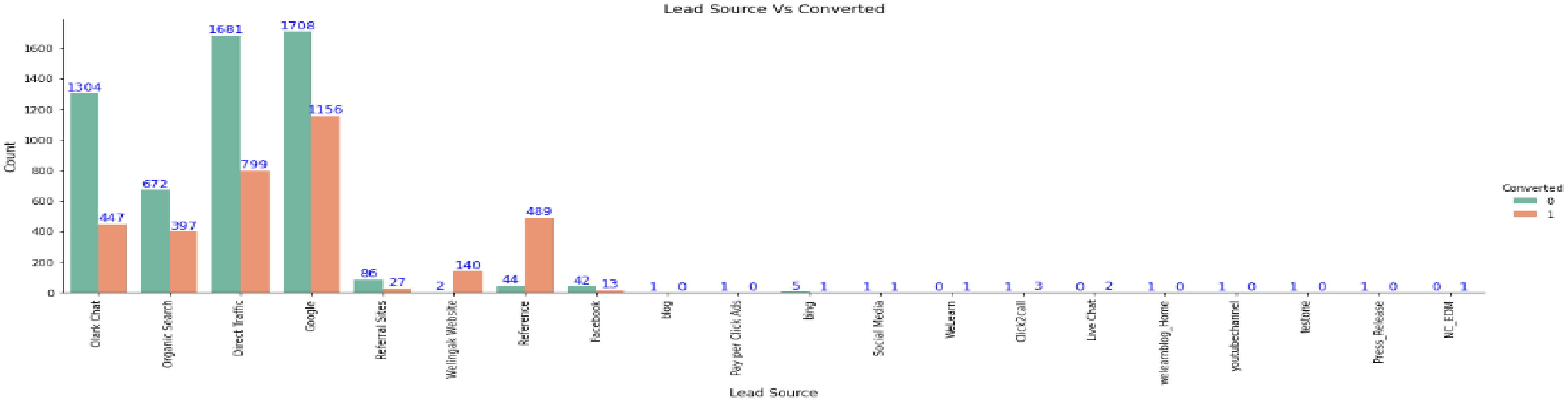
Exploratory Data Analysis

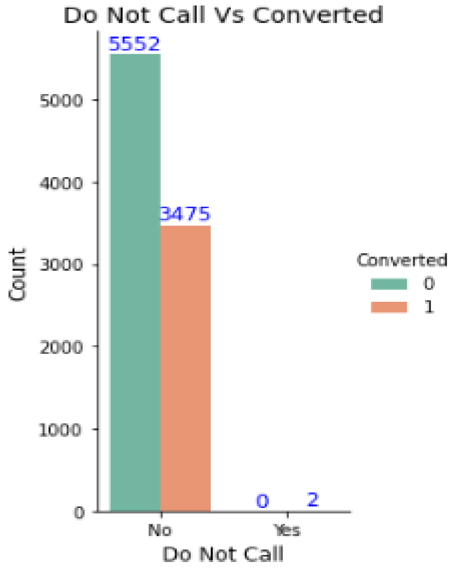


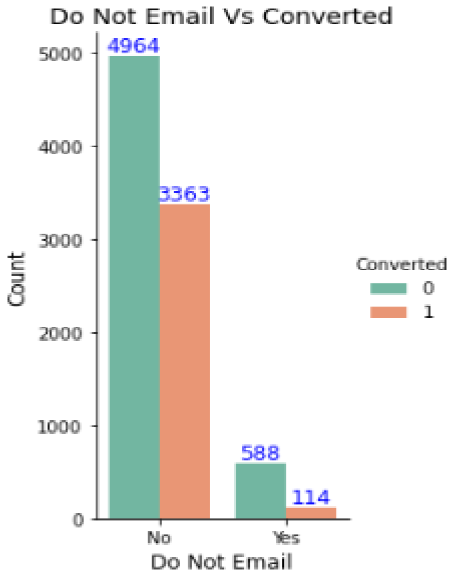


Commentary

- Landing page submission having maximum conversion from Lead origin
- Total Conversion rate is 39%

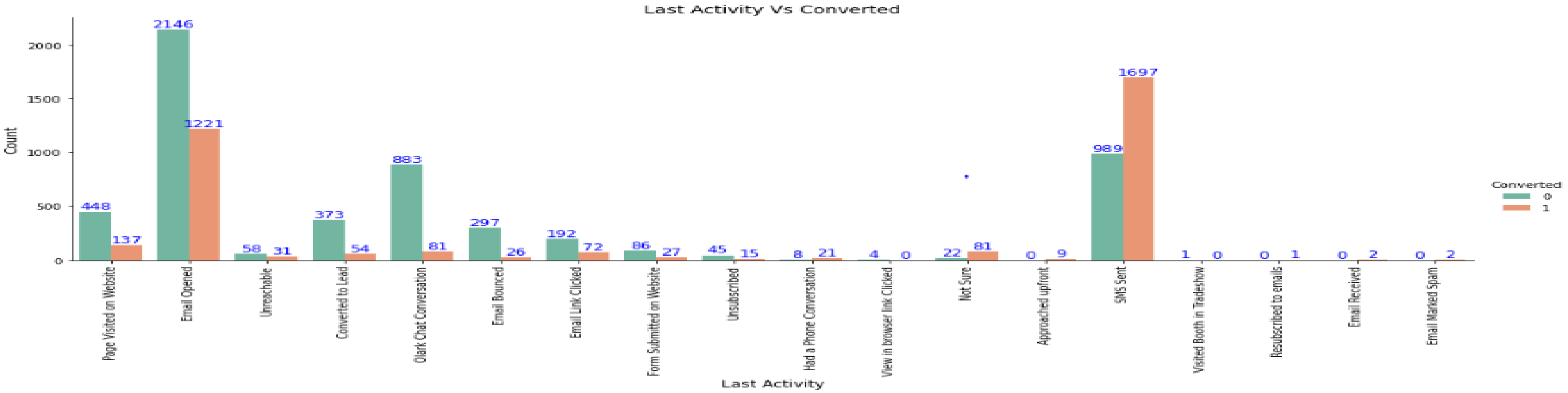


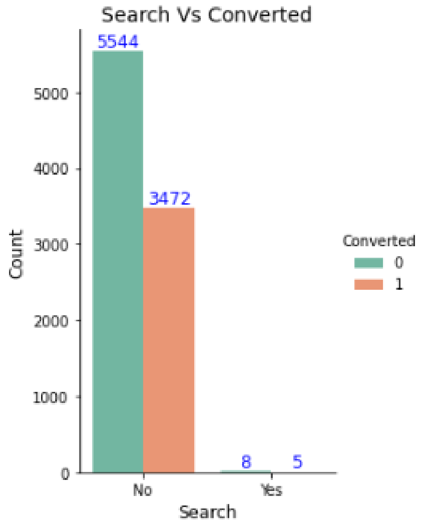


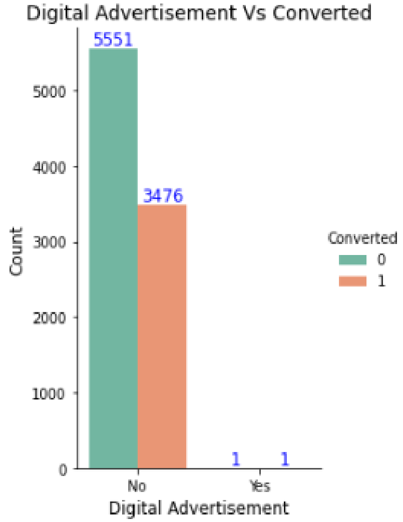


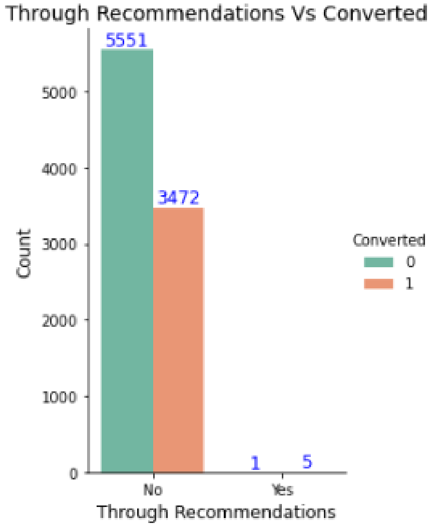
Commentary

- Major conversion in lead source from Google
- Major conversion from Emails sent & calls made



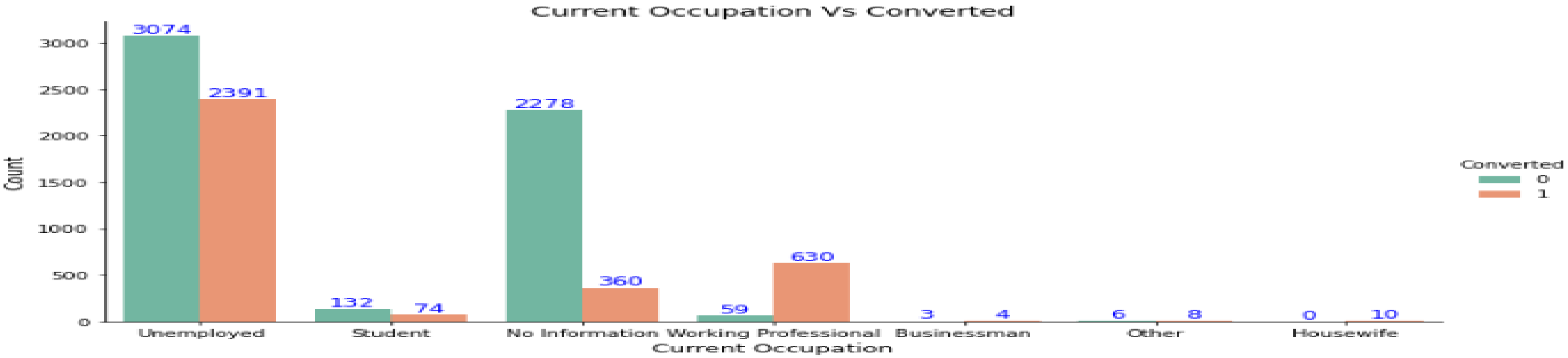


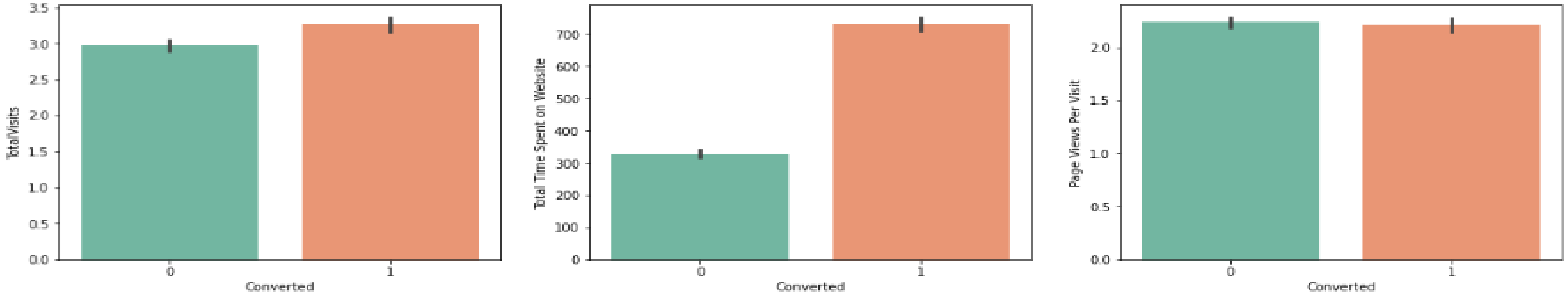




Commentary

- Major conversion in Last activity from SMS Sent
- No much conversion rates seen in Search, Digital Advertisement & Through recommendations

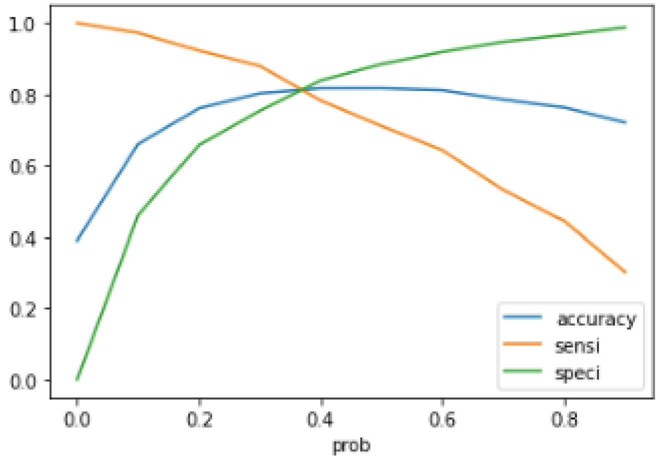


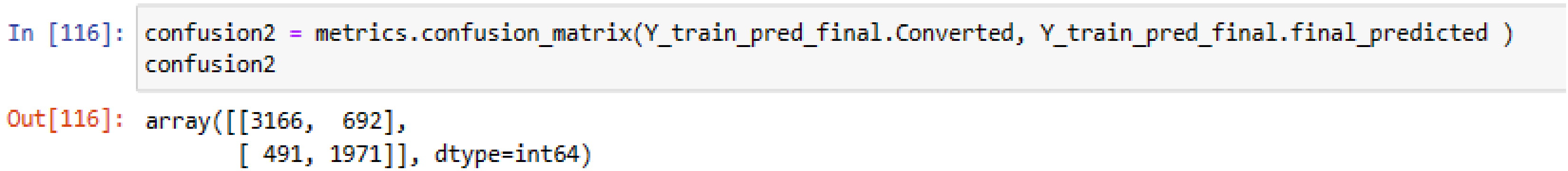


Commentary

- Major conversion in current occupation from unemployed people
- High conversion rates for Total Visits, Total Time Spent on Website, Page Views Per Visit

Sensitivity, Specificity & Accuracy on Train Data Set





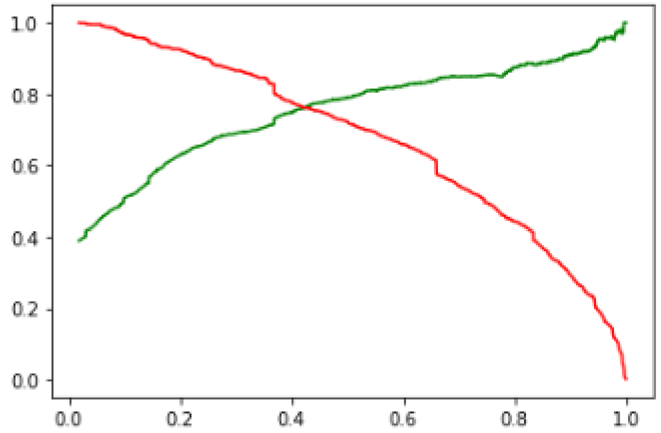
```
In [116]: confusion2 = metrics.confusion_matrix(Y_train_pred_final.Converted, Y_train_pred_final.final_predicted )
          confusion2

Out[116]: array([[3166,  692],
                 [ 491, 1971]], dtype=int64)
```

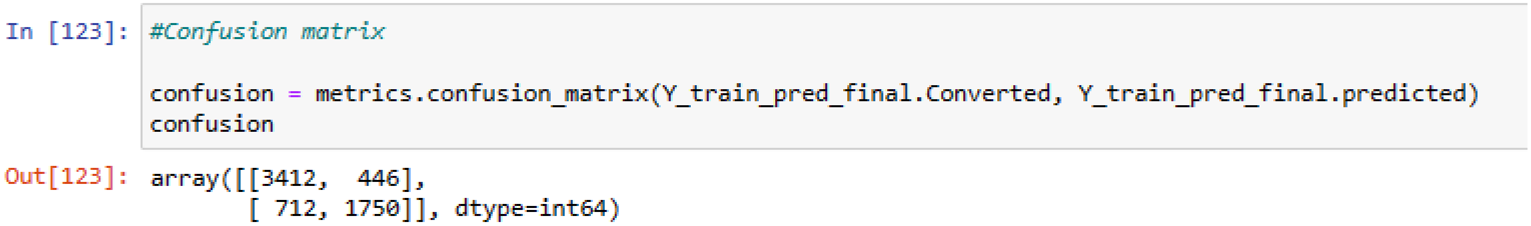
Confusion Matrix	
3166	692
491	1971

Commentary

- 0.37 optimal cut off based on Sensitivity, Specificity & Accuracy
- Sensitivity – 80%
- Specificity – 82%
- Accuracy – 80%

Precision and Recall on Train Data set





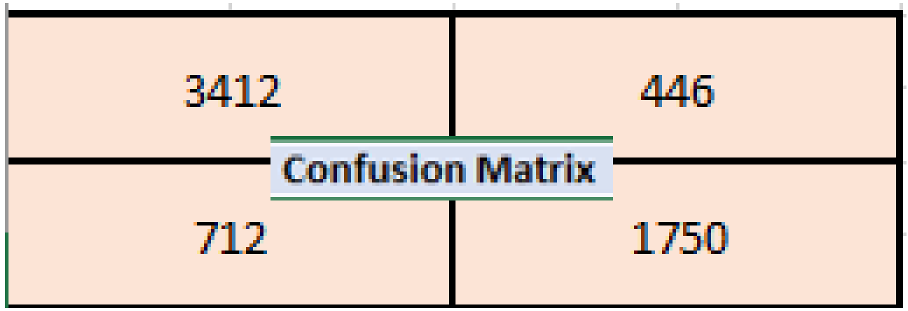
```
In [123]: #Confusion matrix

          confusion = metrics.confusion_matrix(Y_train_pred_final.Converted, Y_train_pred_final.predicted)
          confusion

Out[123]: array([[3412,  446],
                 [ 712, 1750]], dtype=int64)
```

Confusion Matrix	
3412	446
712	1750

Commentary

- 0.42 optimal cut off based on Precision & Recall
- Precision – 79%
- Recall – 71%

Sensitivity, Specificity & Accuracy on Test Data Set



Confusion Matrix

3412	446
712	1750

Commentary

- Sensitivity – 80%
- Specificity – 82%
- Accuracy – 81%

Variables Impacting the Conversion Rate

- Do Not Email
- Total Visits
- Total Time Spent On Website
- Lead Origin–Lead Page Submission
- Lead Origin–Lead Add Form
- Lead Source-Olark Chat
- Last Source–Welingak Website
- Last Activity–Email Bounced
- Last Activity–Not Sure
- Last Activity–Olark Chat Conversation
- Last Activity–SMS Sent
- Current Occupation–No Information
- Current Occupation–Working Professional
- Last Notable Activity–Had a Phone Conversation
- Last Notable Activity-Unreachable

Conclusion

The top 3 variables that contribute for lead getting converted in the model are

- ✓ Total time spent on website
- ✓ Lead Add Form from Lead Origin
- ✓ Had a Phone Conversation from Last Notable Activity

From business perspective:

- Focus: X education needs to focus on running campaigns that direct user to landing page for submission
- Target: X education’s target audience should be ‘unemployed’ people who are either looking for job change or career change with learning as an edge.

Model: Consultation selling model i.e. engaging phone conversations is the go-to-market strategy for X education.

1. **Testing**: We have checked both Sensitivity-Specificity as well as Precision and Recall Metrics.
2. **Calculation**: The optimal cut off has been calculated for the final prediction based on Sensitivity and Specificity. Also, the lead score calculated shows the conversion rate on the final predicted model is around 80% (in train set) and 79% in test set.
3. **Calculated using trained set**: The approximately closer values of test to the respective values set are around-
   - ✓ Accuracy: 81%
   - ✓ Sensitivity: 79%
   - ✓ Specificity: 82%

Conclusion: Overall model seems to be good.