Lead Score Case Study

Submitted By:

Antara Chatterji

Satvik Yadav

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.

Lead Conversion Process - Demonstrated as a 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, it is required 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.

The company requires to build a model wherein it is required 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%.

Case Study Objective

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.

Solution Methodology

Data Sourcing, Cleaning and Preparation:

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

Feature Scaling and Splitting Train and Test Sets:

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

Model Building:

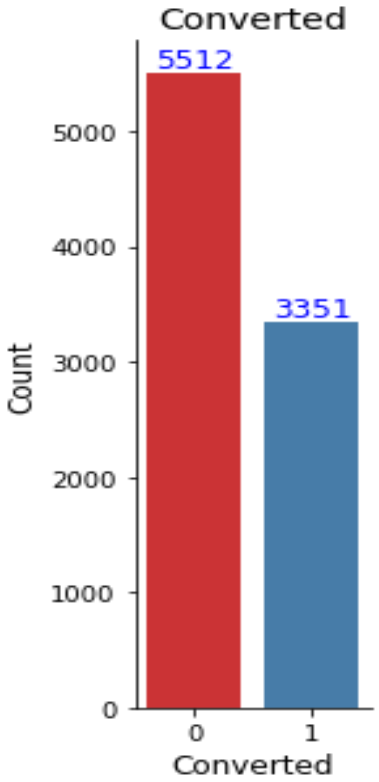
- Feature Selection using RFE.
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity ,precision , recall and evaluate the model

Result

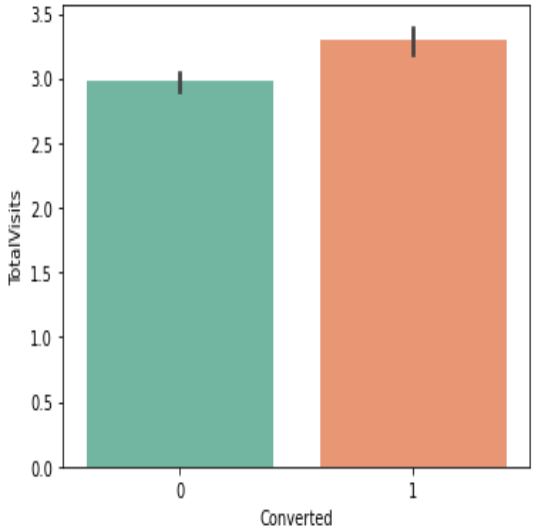
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics.

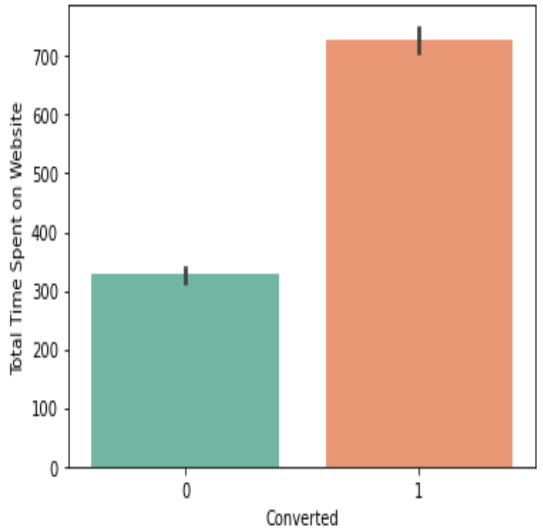
Exploratory Data Analysis

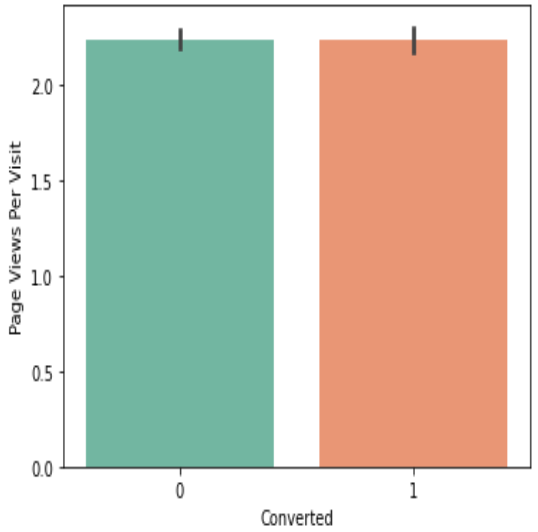
The Conversion rate in total in around 39%



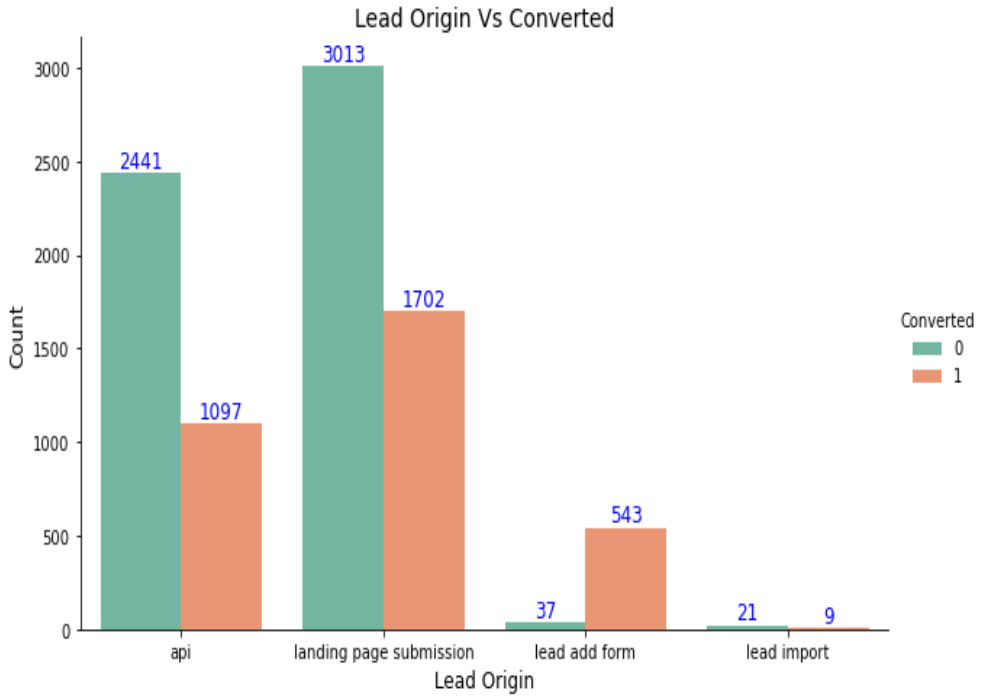
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit



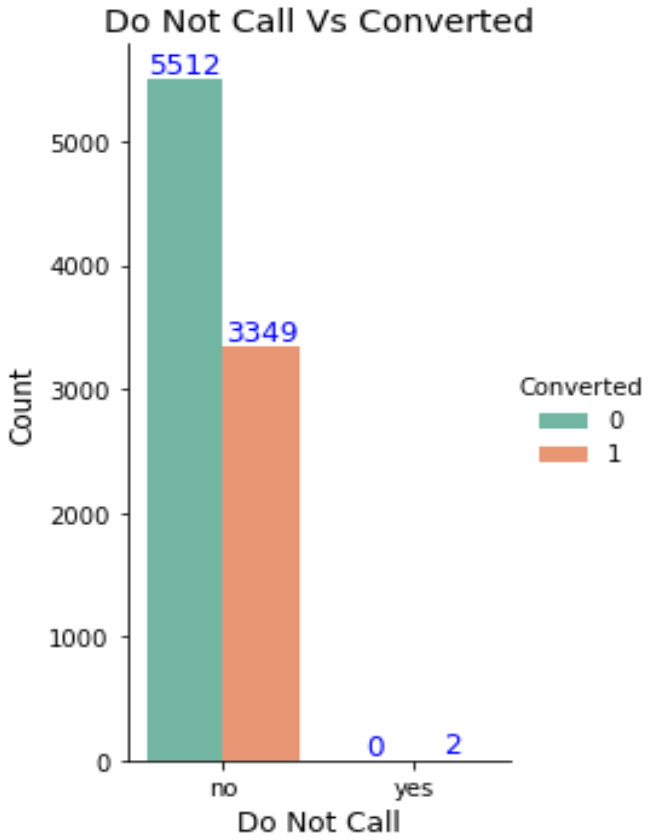


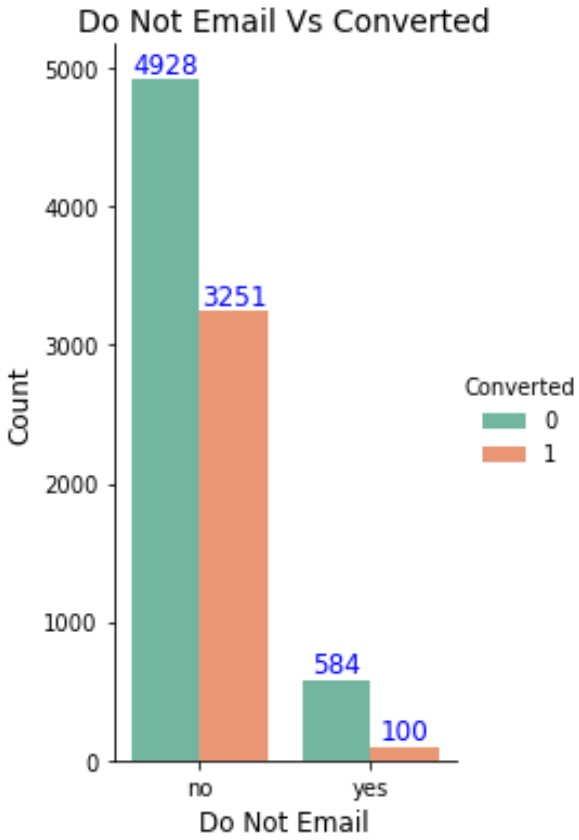


In Lead Origin, maximum conversion happened from Landing Page Submission

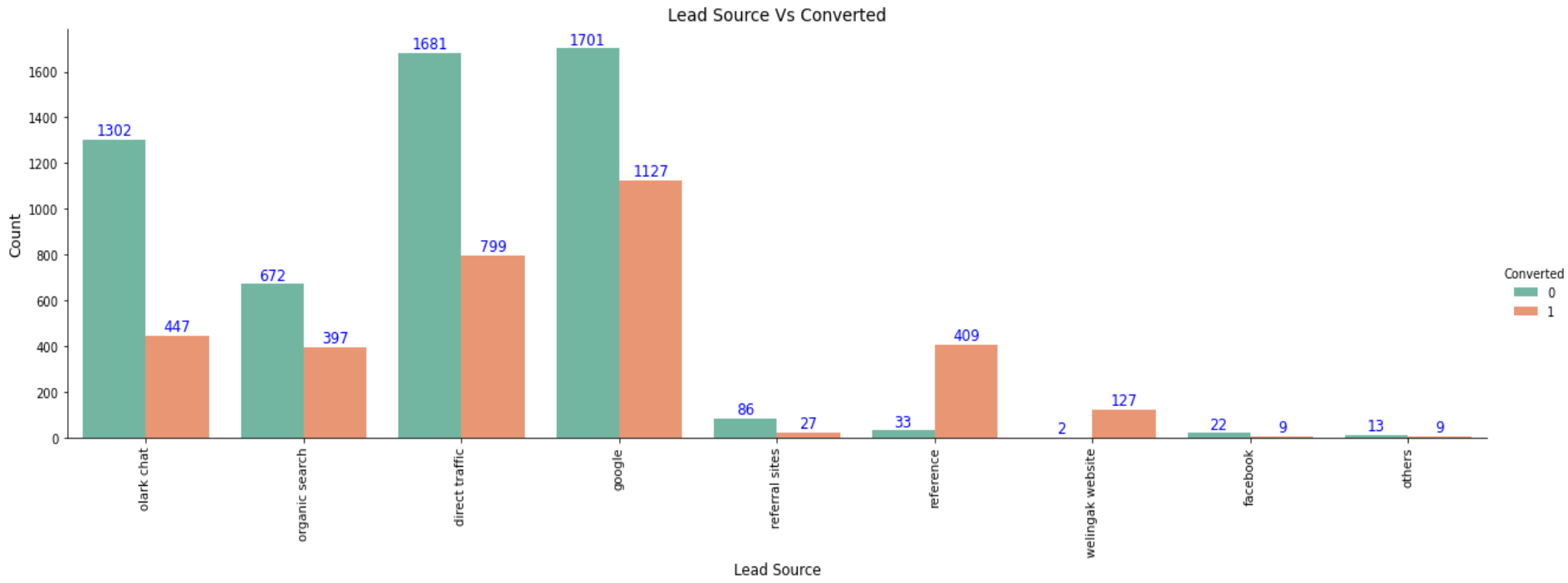


Major conversion has happened from Emails sent and Calls made

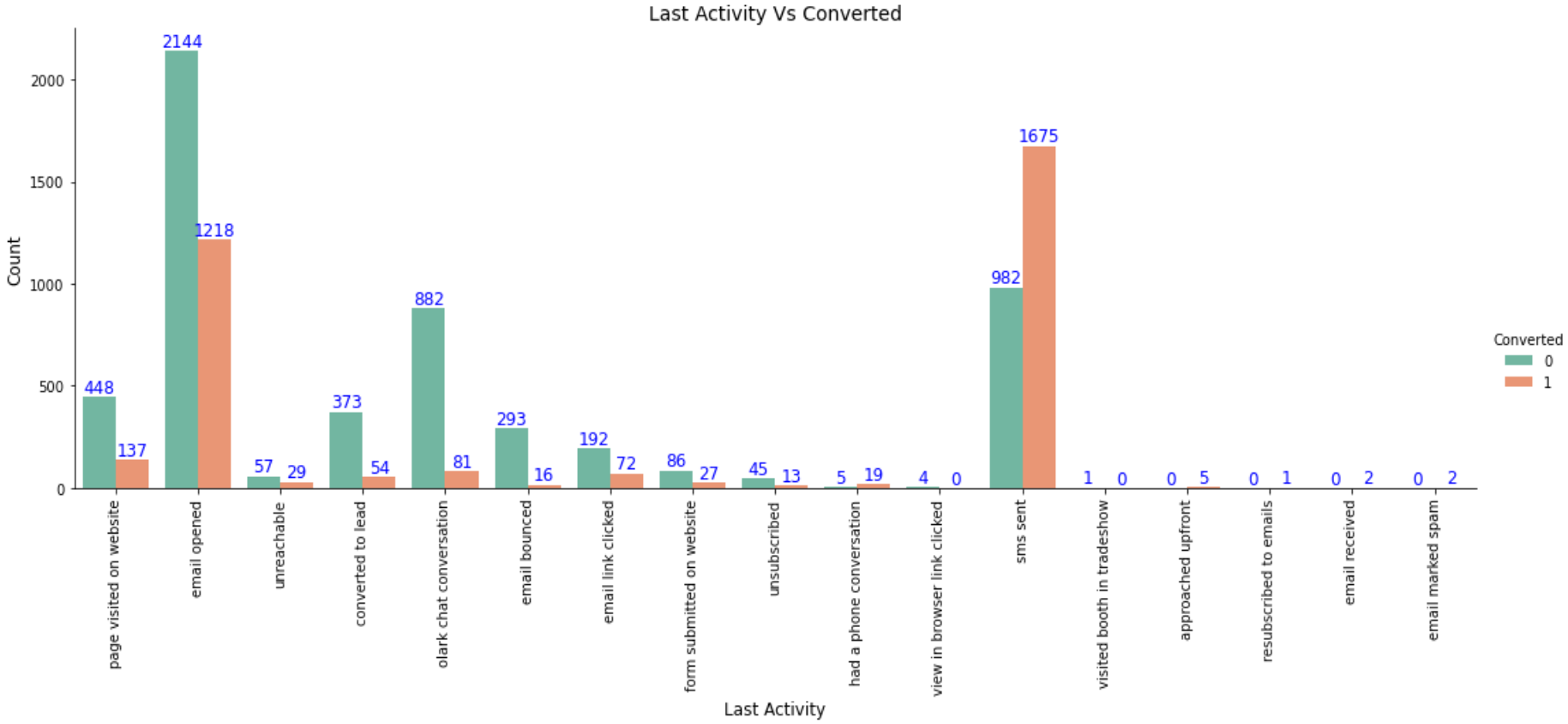




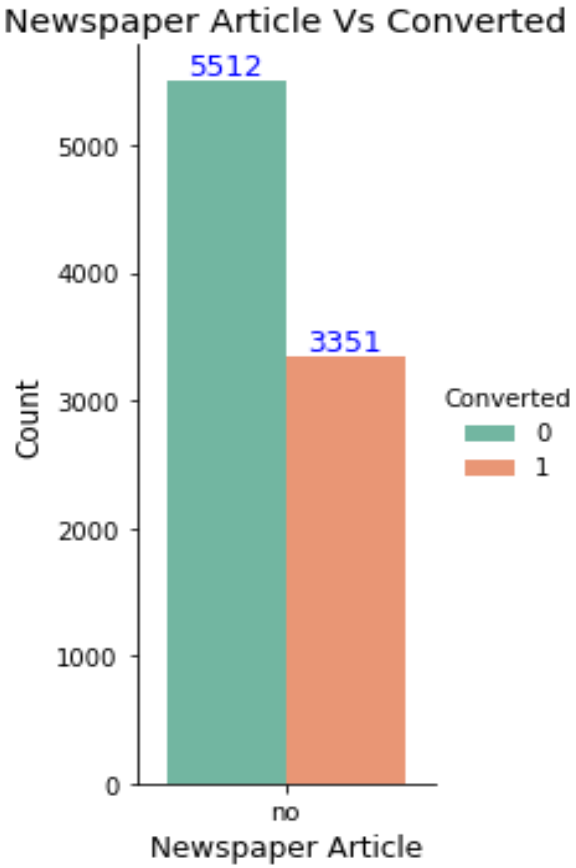
Major conversion in the lead source is from Google

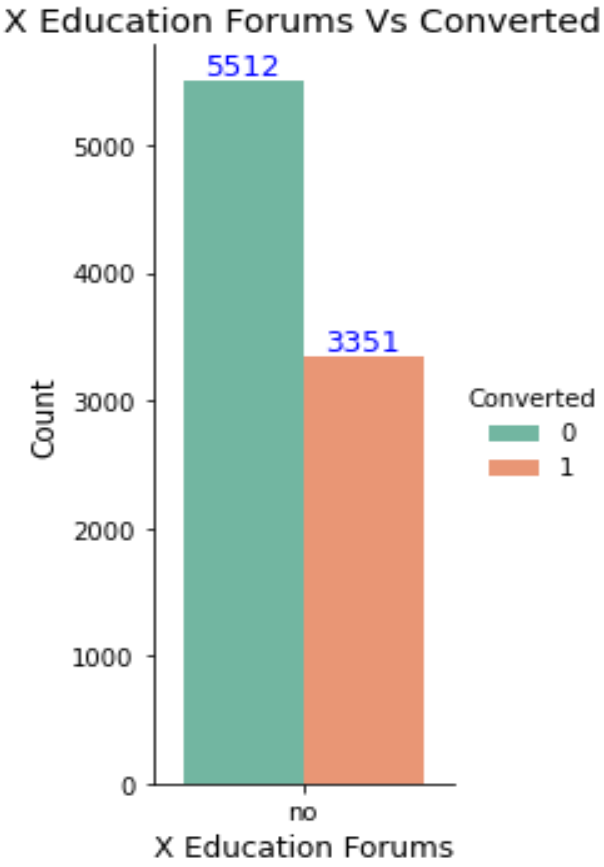


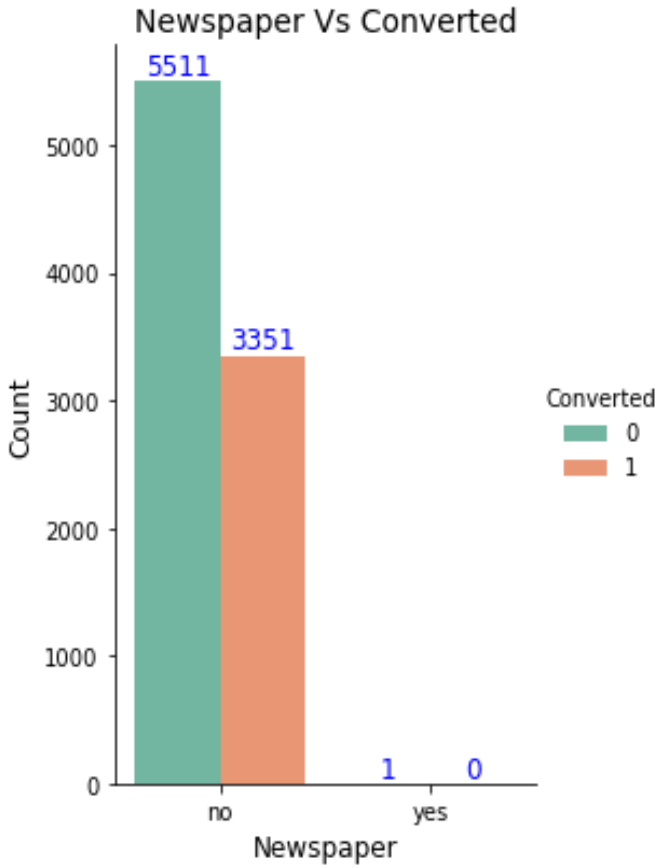
Last Activity value of SMS Sent had more conversion.



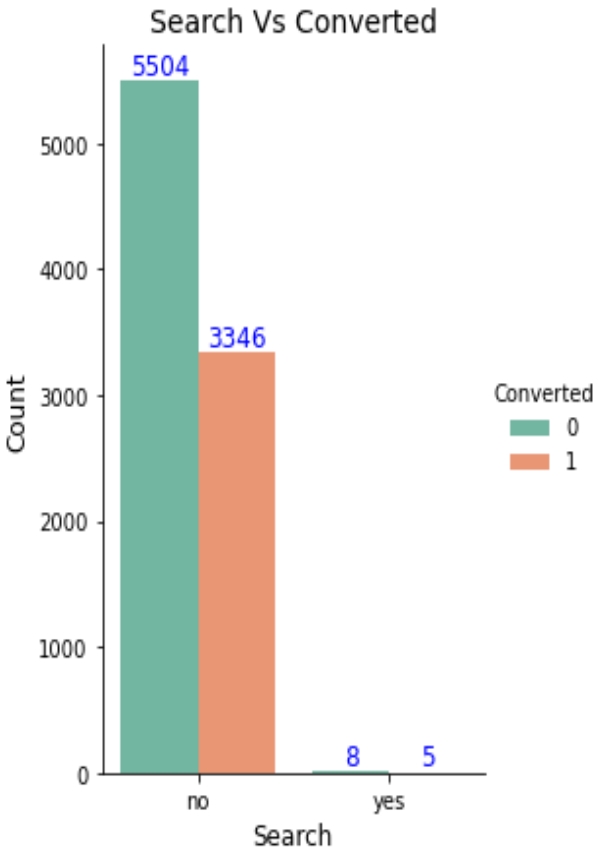
The columns with single values = “No” were dropped

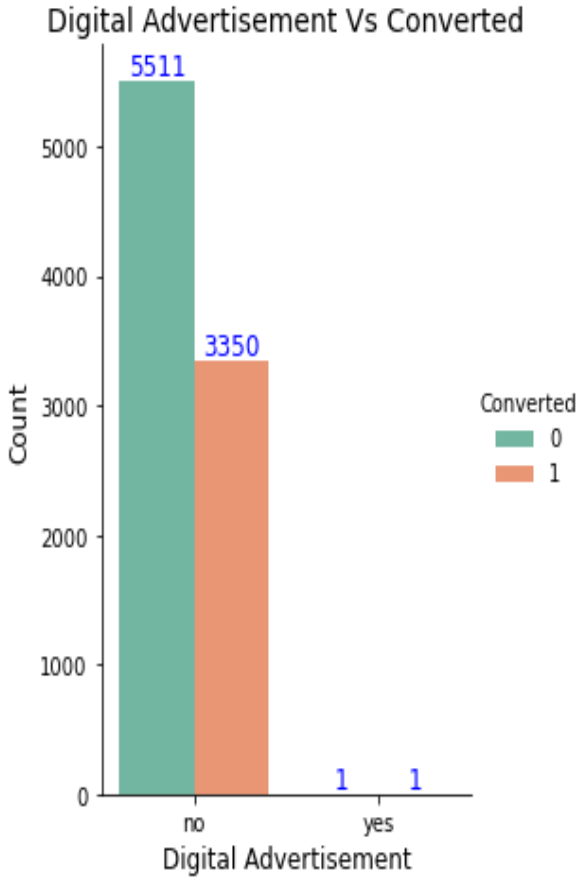


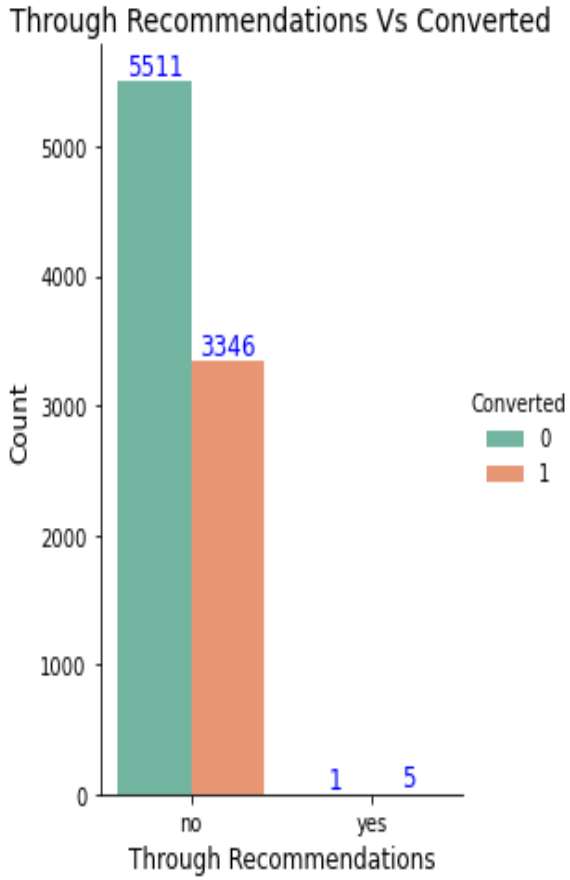




Not much impact was seen on conversion rates through Search, Digital Advertisement, Through Recommendations.





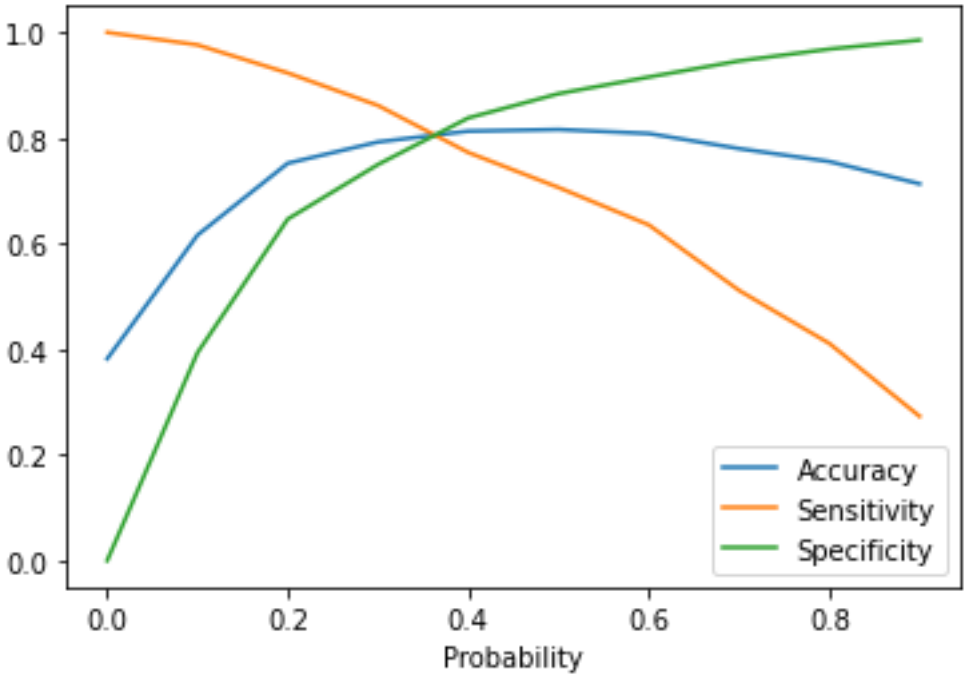


Variables Impacting the Conversion Rate (the topmost)

- Also, after analyzing, the following variables(Top 10) were found to be mostly important for getting potential buyers converted:
  - `TotalVisits` : Total number of people who visited the profile
  - `What is your current occupation_housewife` : Visitors who belong to the housewife category
  - `Last Activity_email marked spam` (from Last Activity)
  - `Last Activity_email received` (from Last Activity)
  - `Lead Source_welingak_website` (from Lead Source)
  - `Total Time spent on the website`
  - `Lead Source_reference` (from Lead Source)

Model Evaluation - Sensitivity and Specificity on Train Data Set

The graph depicts an optimal cut off of 0.35 based on Accuracy, Sensitivity and Specificity



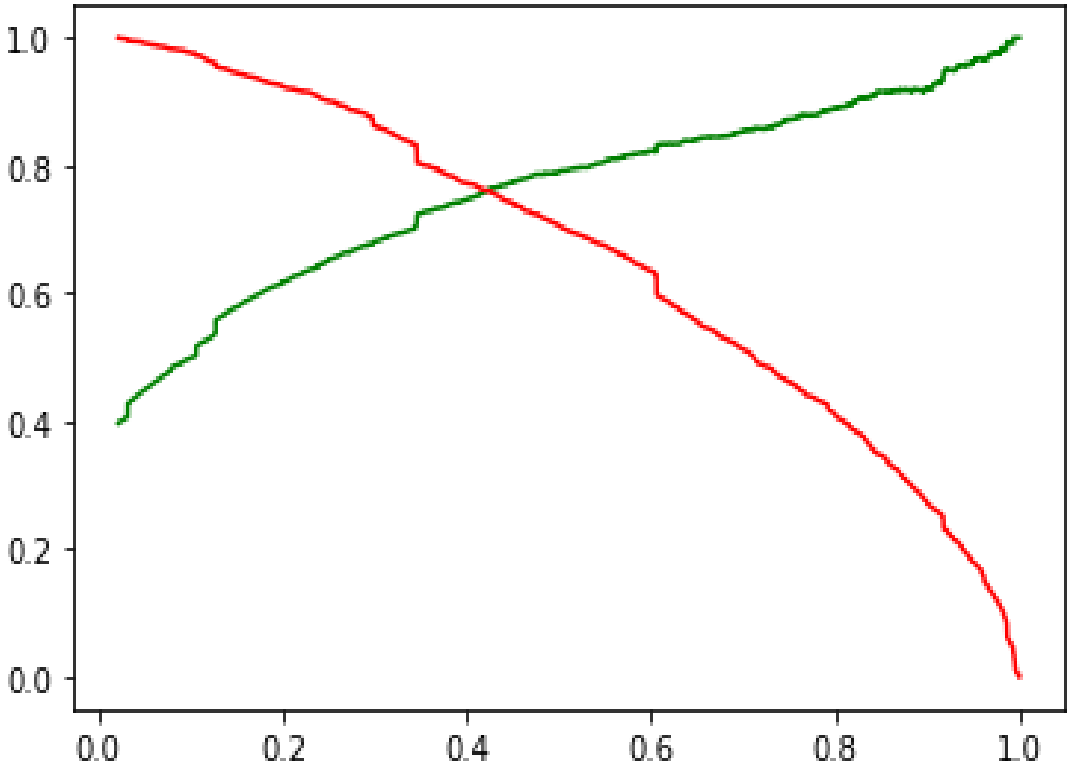
Confusion Matrix

3112	722
468	1902

- Accuracy - 81%
- Sensitivity - 80 %
- Specificity - 81 %

Model Evaluation- Precision and Recall on Train Dataset

The graph depicts an optimal cut off of 0.41 based on Precision and Recall



Confusion Matrix

3231	603
548	1822

- Precision - 75 %
- Recall - 76 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

Confusion Matrix

1426	252
248	733

- Accuracy - 81 %
- Sensitivity - 79 %
- Specificity - 81 %

Conclusion

- `Accuracy`, `Sensitivity` and `Specificity` values of test set are around **`81%`, `80%` and `81%`** which are approximately closer to the respective values calculated using trained set.

- Also the `lead score` calculated in the `trained set` of data shows the conversion rate on the final predicted model is **around 80%. Hence overall this model seems to be good.**

- Also, after analyzing, the following variables were found to be mostly important for getting potential buyers converted:

-`TotalVisits` : Total number of people who visited the profile
- `What is your current occupation_housewife` : Visitors who belong to the housewife category
- `Last Activity_email marked spam` (from Last Activity)
- `Last Activity_email received` (from Last Activity)
- `Lead Source_welingak_website` (from Lead Source)
- `Total Time spent on the website`
- `Lead Source_reference` (from Lead Source)