LEAD SCORE CASE STUDY

- **Submitted by :** Sanjana Suman

Problem Statement :

- X Education sells online courses to industry professionals. 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%.
- **Business Goal**:
- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned 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%.

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 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

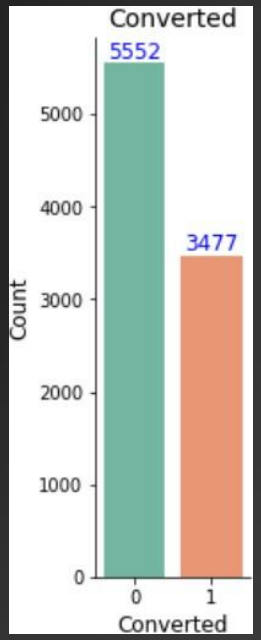
Calculate various metrics like accuracy, sensitivity, specificity, precision and recall and evaluate the model.

Result

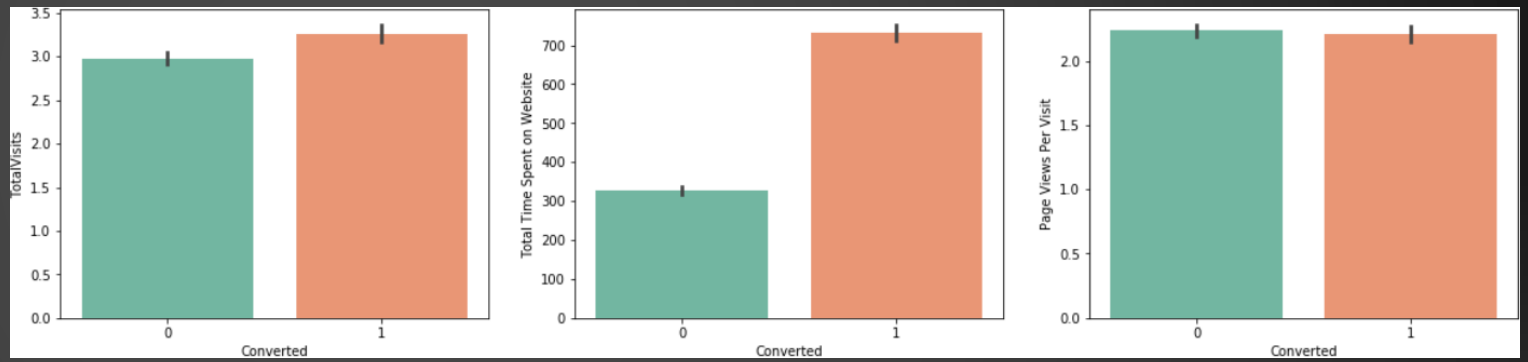
- Determine the lead score and check if target final predictions amounts to 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics

EXPLORATORY DATA ANALYSIS

We have around 39% Conversion rate in Total

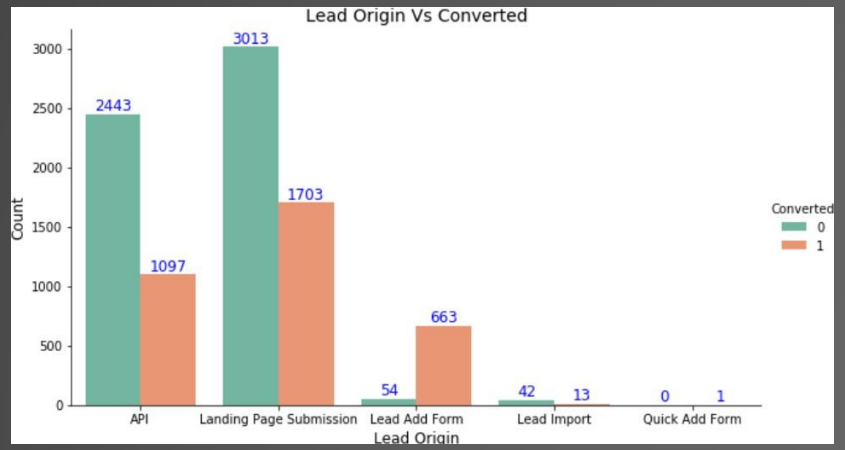


The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit

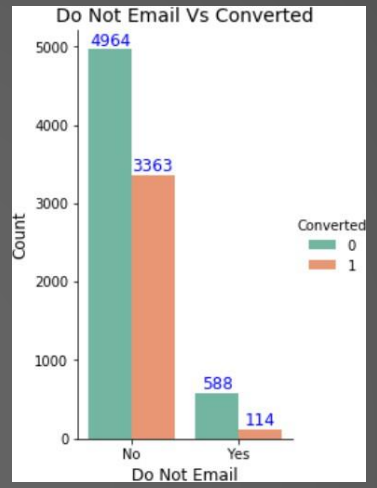


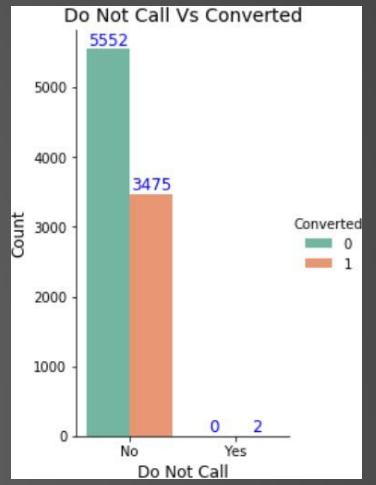
.

In Lead Origin, maximum conversion happened from Landing Page Submission

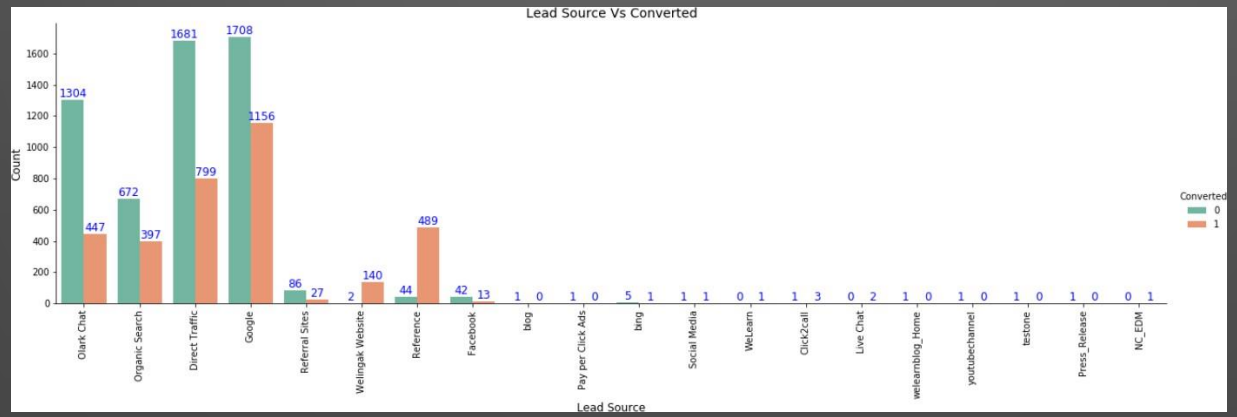


Major conversion has happened from Emails sent and Calls made

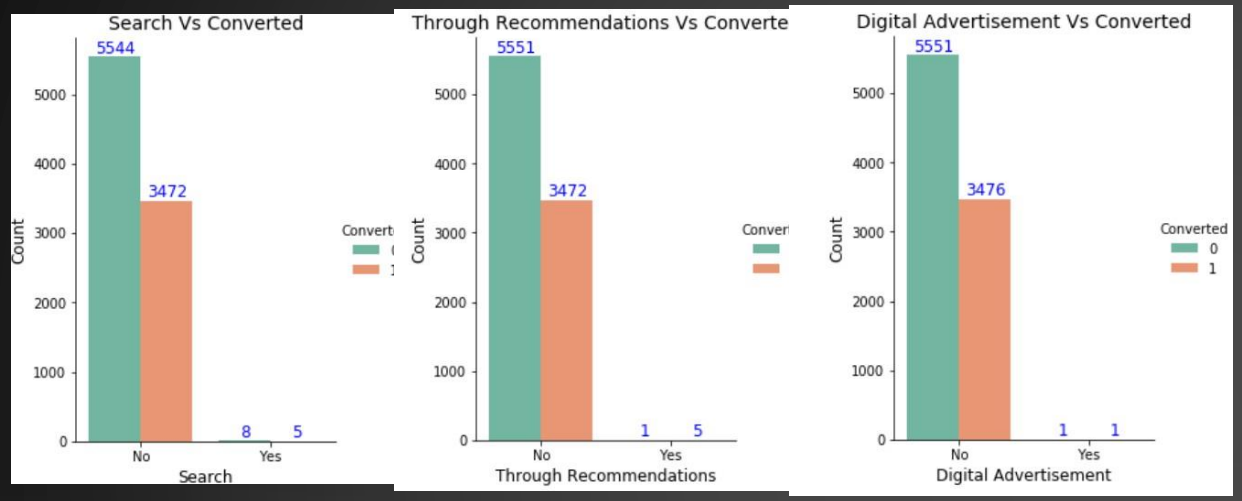




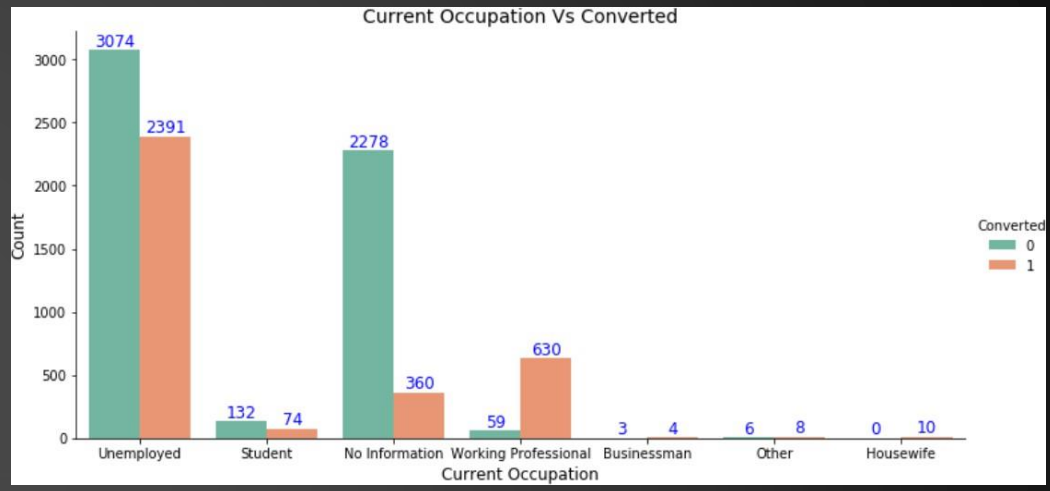
Major conversion in the lead source is from Google



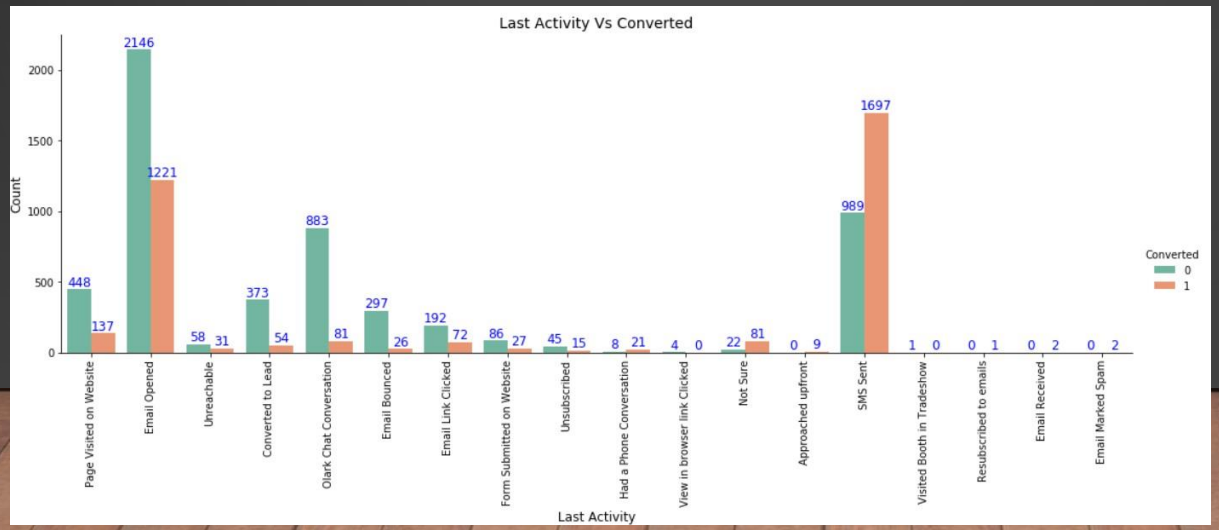
Not much impact on conversion rates through Search, digital advertisements and through recommendations



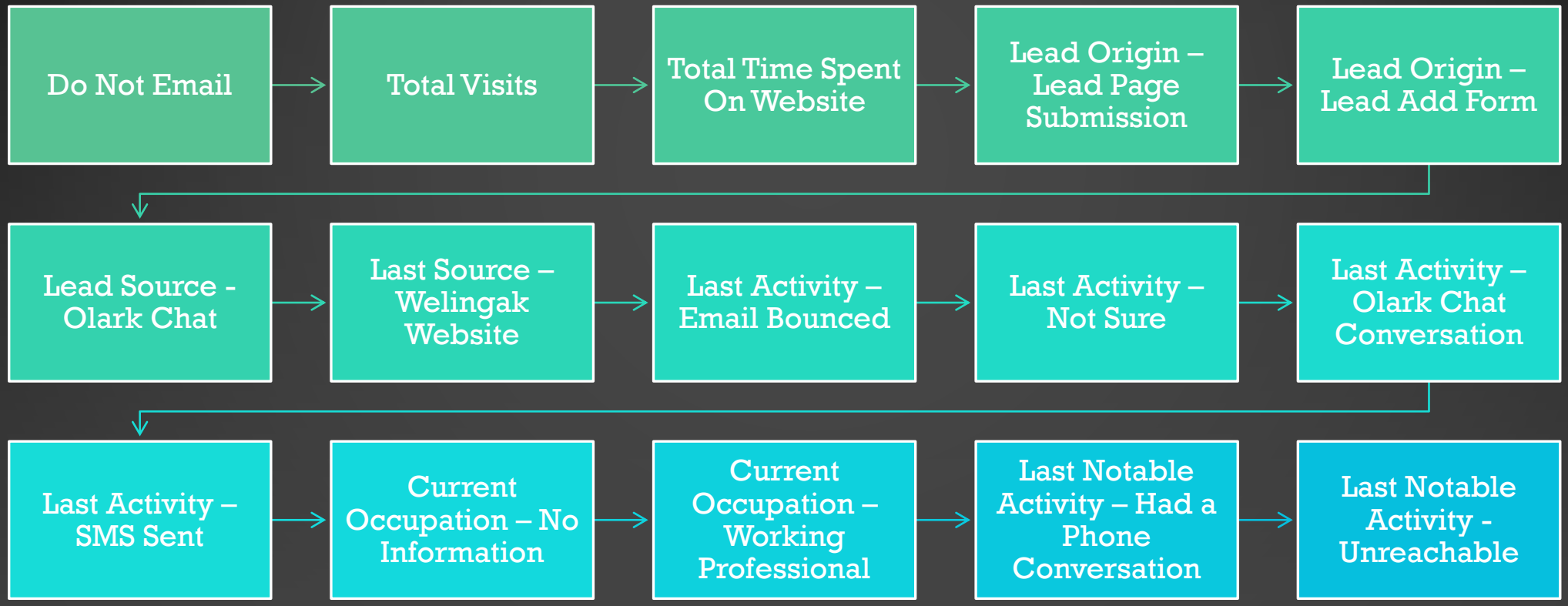
More conversion happened with people who are unemployed



Last Activity value of SMS Sent' had more conversion.

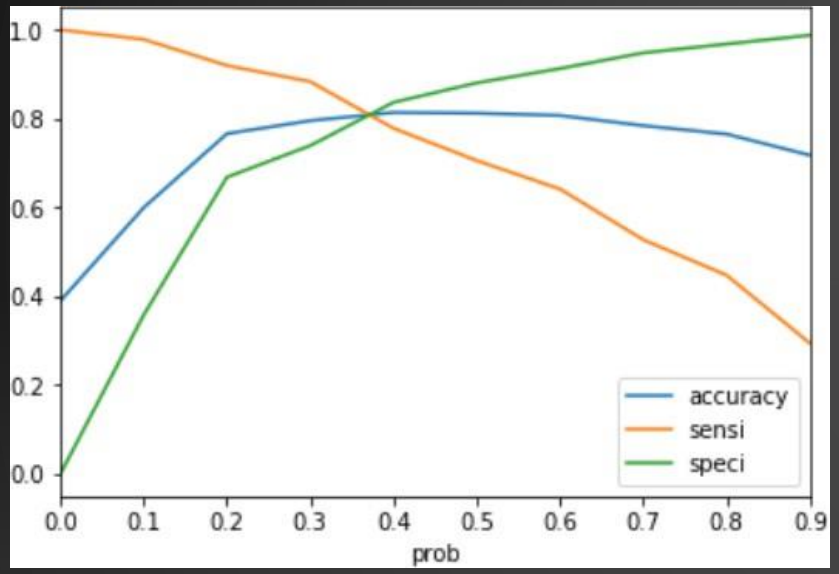


VARIABLES IMPACTING THE CONVERSION RATE



MODEL EVALUATION - SENSITIVITY AND SPECIFICITY ON TRAIN DATA SET

The graph depicts an optimal cut off of 0.37 based on Accuracy, Sensitivity and Specificity



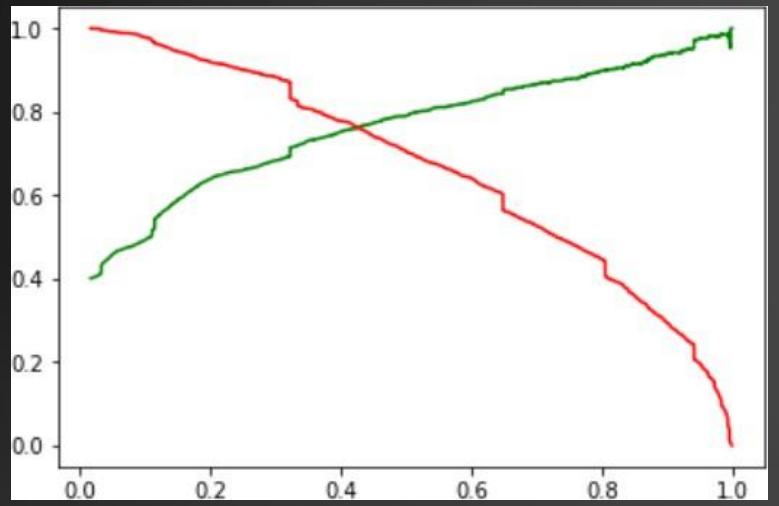
Confusion Matrix

3161	697
974	1965

- Accuracy - 81%
- Sensitivity - 80 %
- Specificity - 82 %
- False Positive Rate - 18 %
- Positive Predictive Value - 74 %
- Positive Predictive Value – 86%

MODEL EVALUATION- PRECISION AND RECALL ON TRAIN DATASET

The graph depicts an optimal cut off of 0.42 based on Precision and Recall



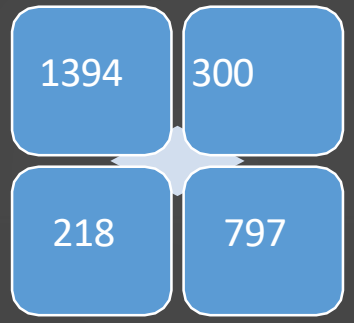
Confusion Matrix

3397	461
725	1737

- Precision - 79 %
- Recall - 71 %

MODEL EVALUATION – SENSITIVITY AND SPECIFICITY ON TEST DATASET

Confusion Matrix



- Accuracy - 81 %
- Sensitivity - 79 %
- Specificity - 82 %

CONCLUSION

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction. –
- Accuracy, Sensitivity and Specificity values of test set are around 81%, 79% and 82% which are approximately closer to the respective values
- calculated using trained set.
- 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
- 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
- Hence overall this model seems to be good.