Lead Scoring Case Study

Submitted by:

Lakshana Govindarajan Vijaya

Mayank Kumar

Amit Kumar

Business Problem

- 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%.
- 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%.

Goal 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.

Business Objective

- Identify the most potential leads known as ‘Hot Leads’
- Build a machine learning model that identifies the hot and promising leads
- Deployment of model for future use

Approach

1. Data Processing
   - Read the data
   - Data sanity checks
   - Missing value treatment
   - Univariate bivariate analysis
   - Handle categorical variables
     - Convert binary to 1/0
     - Grouping values to reduce the number of levels
     - Dummy encoding
   - Handle numerical variables
     - Remove outliers
   - Class imbalance check
   - Correlation analysis

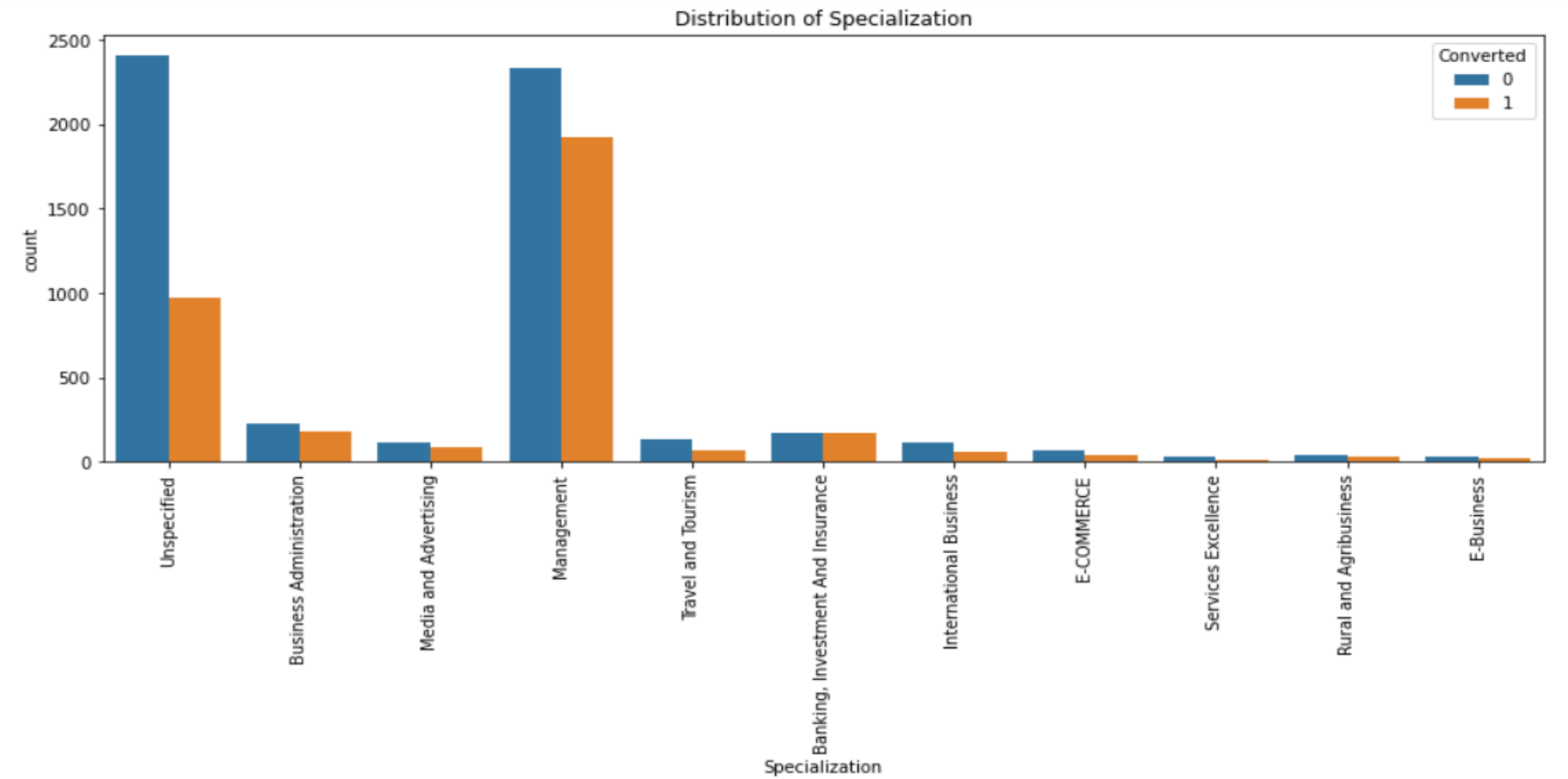
2. Model Building
   - Classification model using logistic regression
   - Feature selection using RFE
3. Model Evaluation
   - Evaluation metrics – Accuracy, Specificity, Sensitivity
4. Recommendation
   - Model Interpretation
   - Conclusion & Recommendations

Data Processing

- No of rows – 9240, No of columns – 37
- Replace Select with null values
- Remove the columns that has missing values of more than 45%
- Replace missing values with median (numerical) and mode(categorical)
- Group different categories to reduce the number of levels within categorical variables, like grouping the low frequency categories into one.
  - Ex- Group Google, Bing → Search Engine
  - Lead Activity, Specialization, Lead Source are the variables where we have implemented this

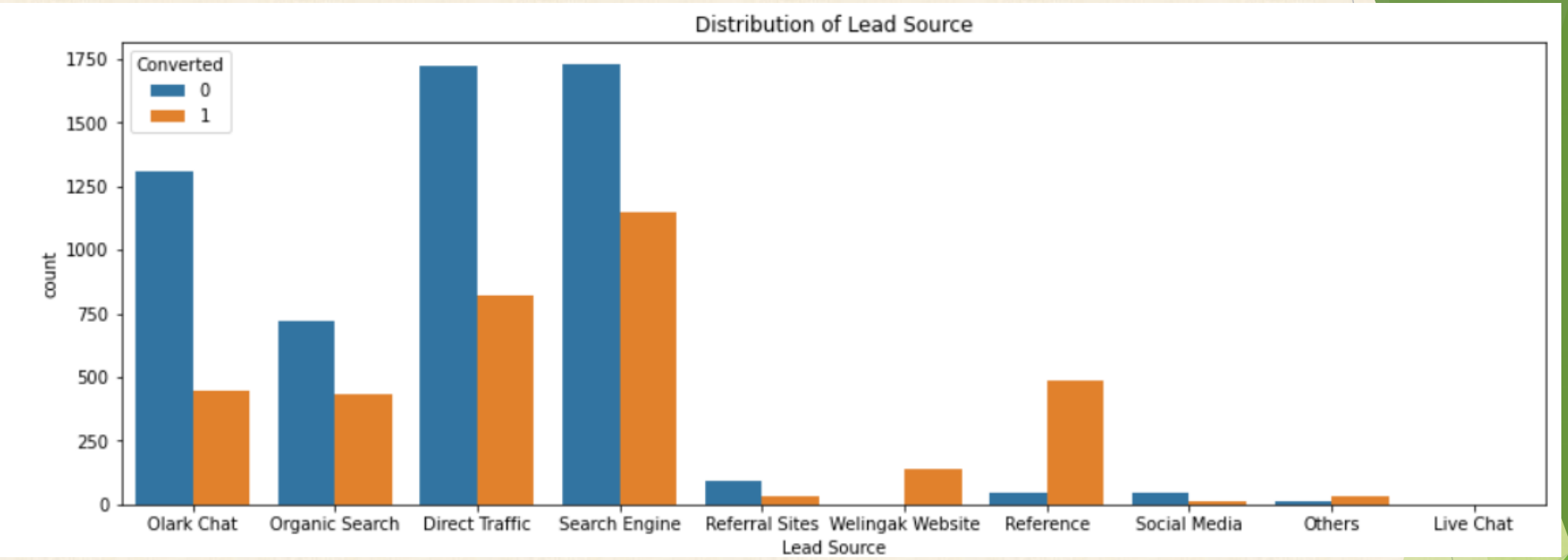
- Drop the variables that do not contribute to the analysis
  - Ex. Country
- Remove the imbalanced variables which do not contribute to analysis
  - News paper article – Yes(2), No(9238), 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'
- Create dummy variables for the categorical variables
  - Convert binary variables with Yes/No to 1/0
  - Perform one hot encoding for the categorical variables and drop one least significant variable each
  - Ex- Lead Origin, Specialization, Lead Source etc

EDA – Impact of Specialization on Target



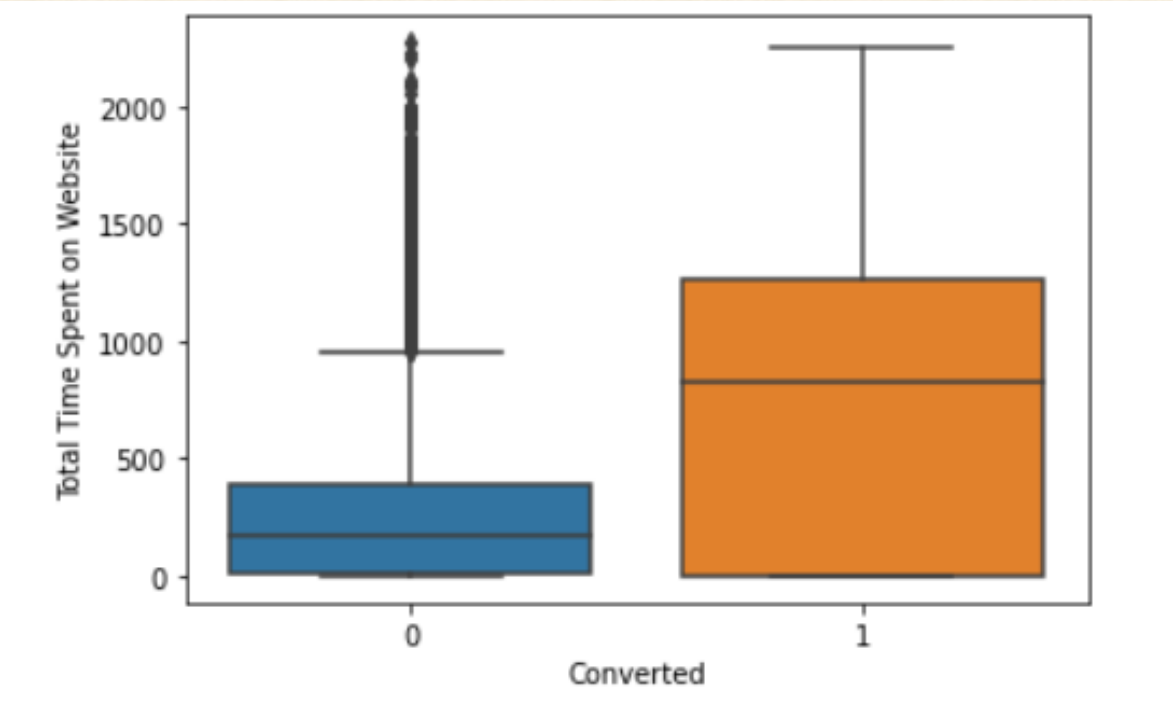
- Most of the conversions happen in the Management specializations
- Leads in the Banking, Investment and Insurance specialization has the higher conversion rate.

EDA - Impact of Lead Source on Target



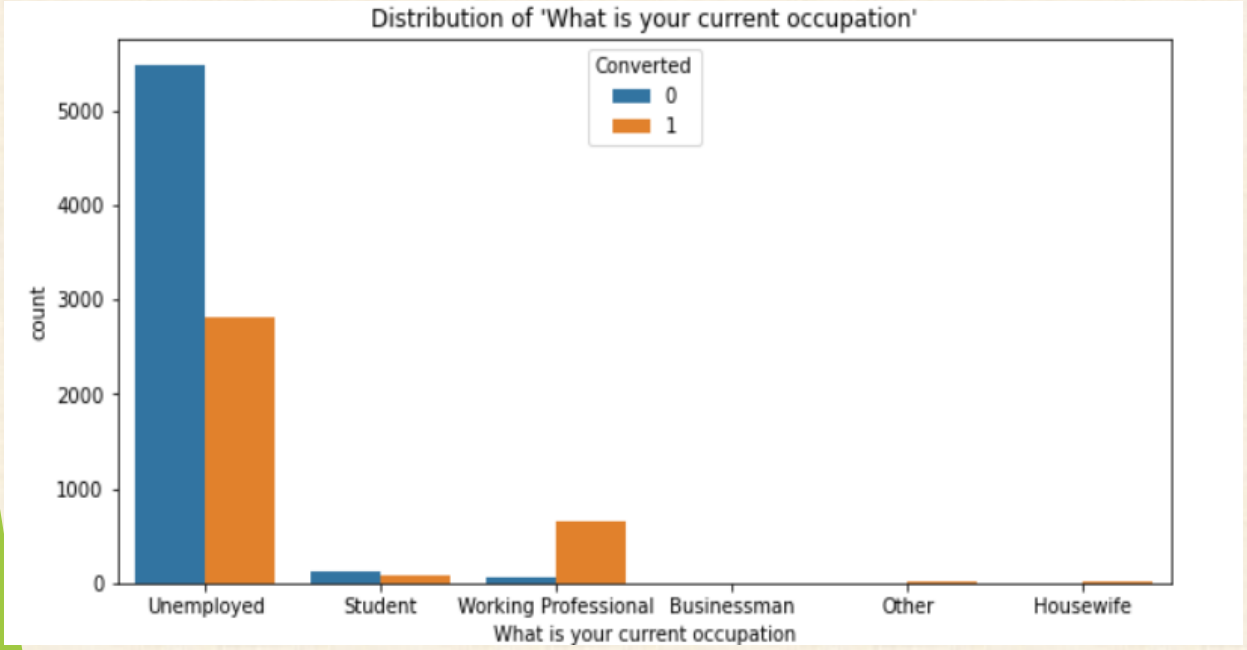
- Search Engine, Direct Traffic and Olark Chat are the most common lead sources.
- Organic Search has the highest conversion rate and its obvious
- Most of the leads who come in through Reference convert and its obvious as they have first hand review from the referring person.

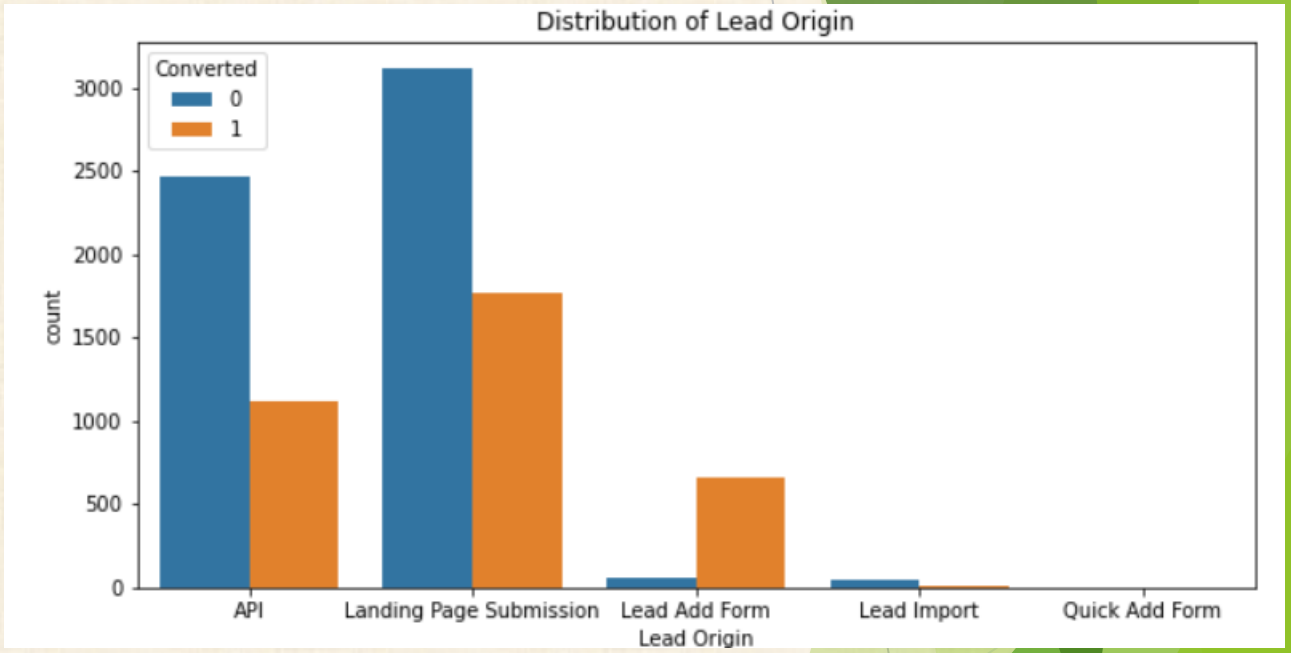
EDA - Impact of Total Time Spent on Website on Target



- We can infer that the leads who spend more time on the website are more likely to convert than the ones who spend less time

EDA - Impact of Current Occupation & Lead Origin on Target



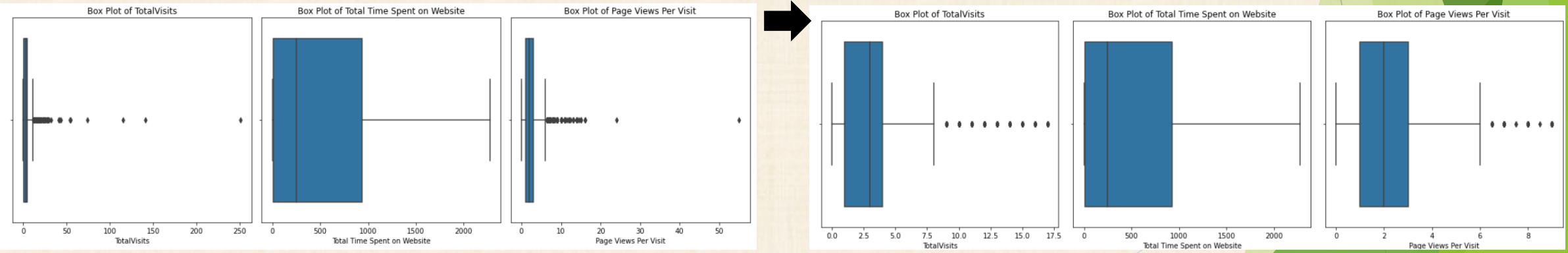


- Most of the leads are Unemployed
- Working Professionals have higher conversion rate

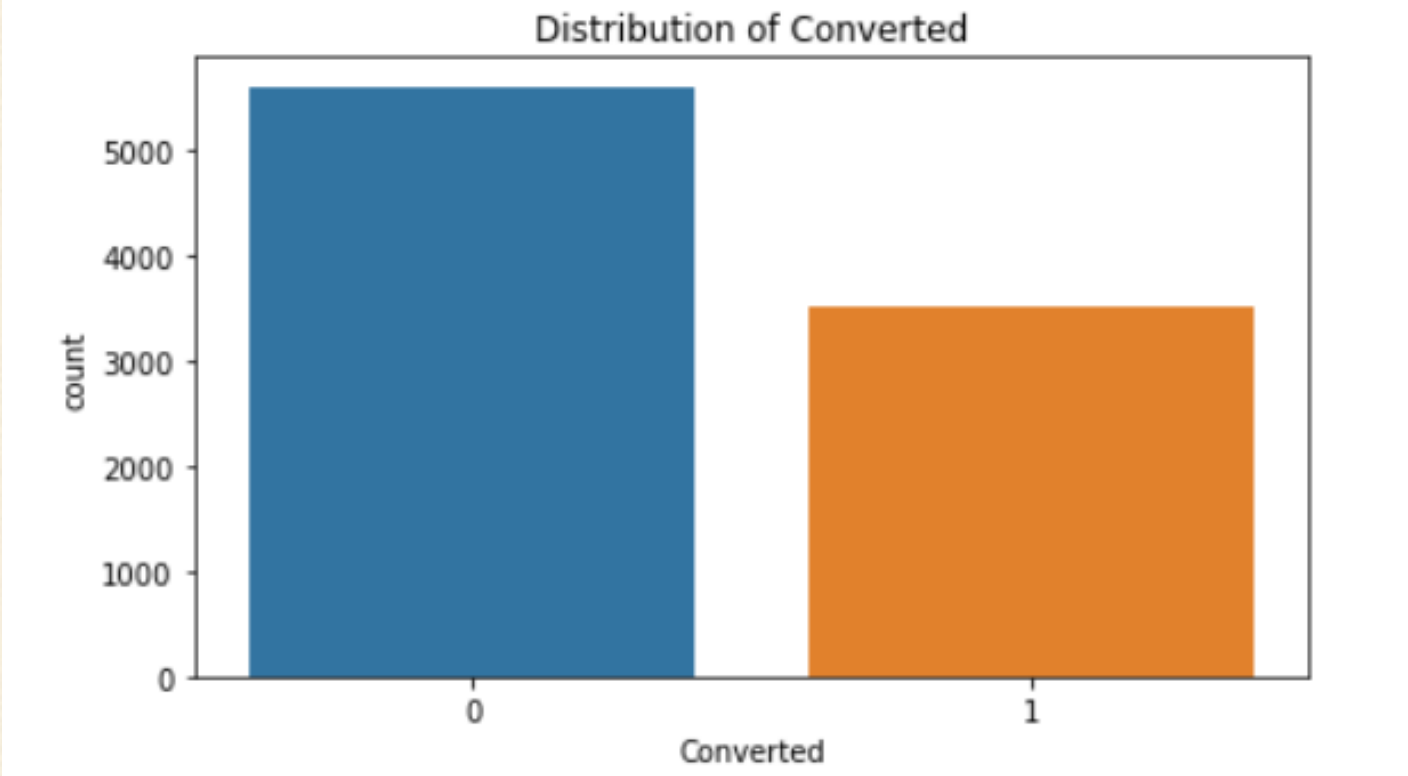
- Those leads who submit their details in landing Page / API have ~50% chance of conversion
- The enquiries through Lead Add Form always convert

Handling Outliers

- Remove top 1% of the outliers from numerical columns
- Total Visits had outliers and are removed
- Page views per visit had outliers and are removed



Class imbalance

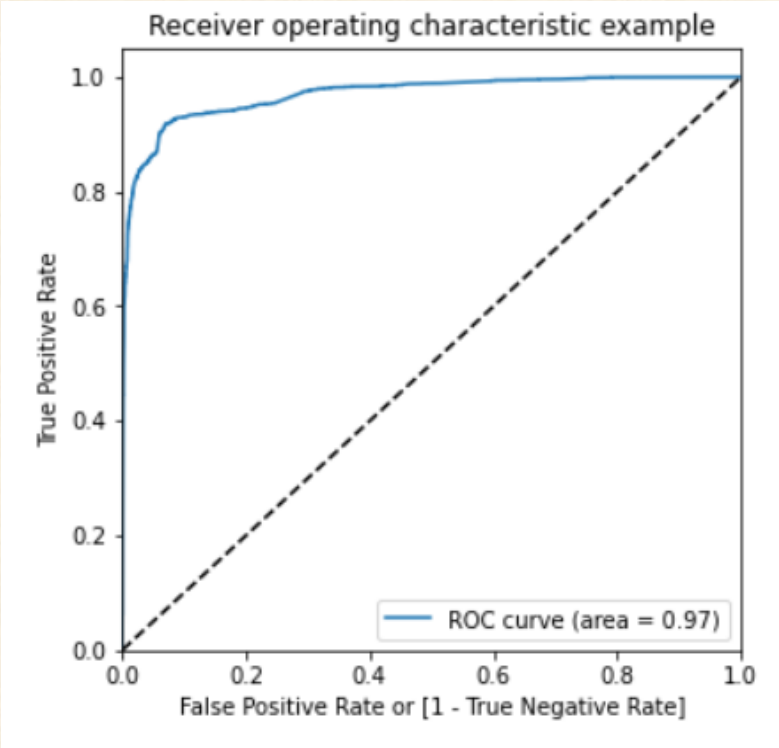


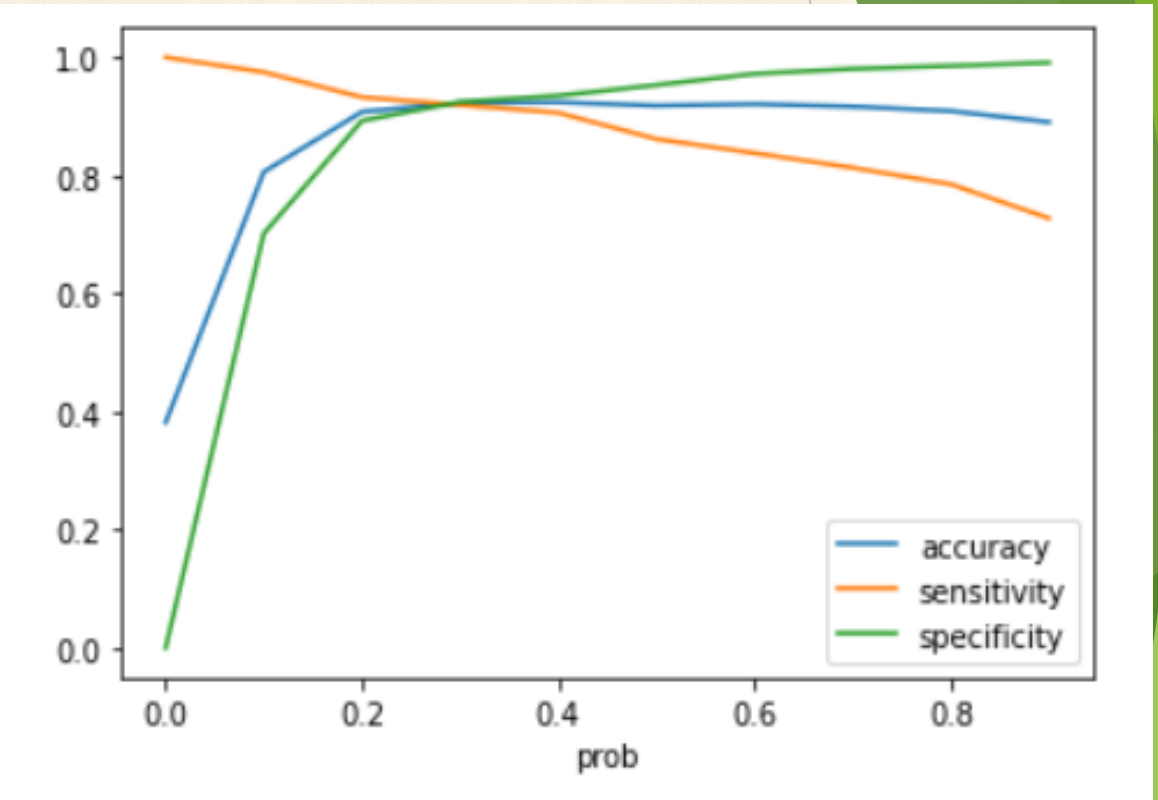
- The imbalance ratio is 1.6
- The data is not heavily imbalanced
- This means for every 1 lead that converts, there are ~1.6 leads that do not convert

Model Building

- Split the data into train and test sets (70:30 ratio).
- Standardize numerical columns.
- Feature selection using RFE with top 15 variables.
- Assess the model using stats model and build the model by removing variables with p-value greater than 0.05 and VIF greater than 5.
- Prediction on the test data set.
- Overall accuracy of the model : 91.95%

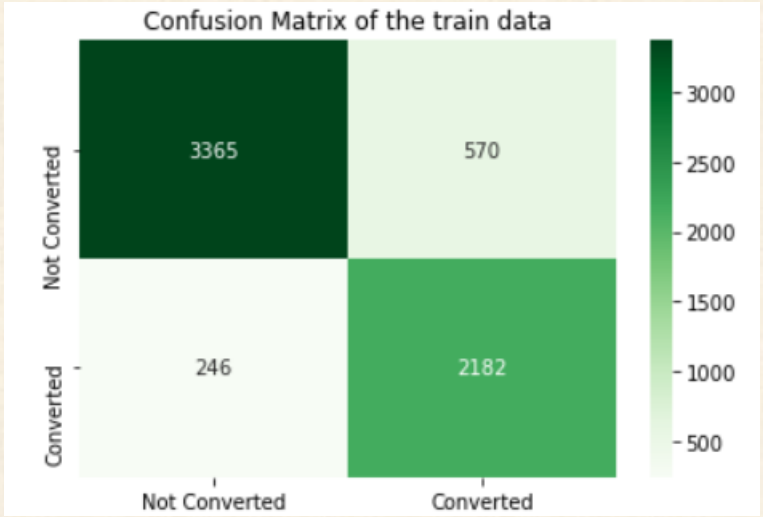
ROC Curve



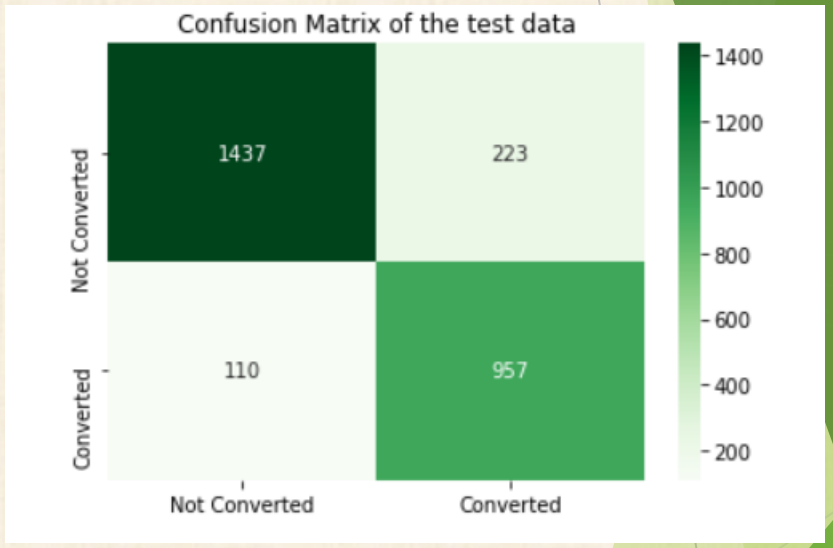


- The Area Under the ROC Curve is 0.97 which is very close to 1 and it's a very good model
- When we check the Sensitivity Specificity trade off, we get 0.25 as the optimal cut off point

Prediction



Train Data Set



Test Data Set

Prediction

Accuracy – 87%

Sensitivity – 89%

Specificity – 85%

Train Data Set

Accuracy – 88%

Sensitivity – 89%

Specificity – 86%

Test Data Set

Recommendations

- **Inference from data & model**
  - If the lead has been added through Lead Add Form, they are more likely to convert
  - Those leads who have responded as ‘Will revert after reading the email’ have a positive impact on conversion
  - Leads who spend more time on the website have a positive impact on conversion
  - Those are Working Professionals also have higher conversion rate
  - Huge number of leads come in through Search Engine and Direct Traffic and they also have higher conversion
  - If the last notable activity was sms sent this also contribute to the lead conversion

Using the above recommendation X-education can convert most of their potential leads into buyers using informed data driven decision making