Lead Scoring Case Study

PRESENTED BY:

ARYA P AND SONALI JOSEPH

Problem Statement:

X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines.

People who browse these website and use the search engines may come across the course and browse through the course or fill up a form for the course or watch a video about it. If any of those people fill up the form providing their email and phone number, those people are classified as ‘Leads’. Company may also get leads through past referrals.

Once these leads are acquired, employees from the sales team reach out to them via calls, emails etc. Through this process some of the leads get converted i.e they join one of the courses. The typical lead conversion rate is around 30%.

Business Goal

X Education needs help in selecting the most promising leads, that is the leads that are most likely to get converted.

The company needs a mode; wherein a lead score is assigned to each lead, such as the leads with higher lead score are more likely to convert compared to those with a lower lead score.

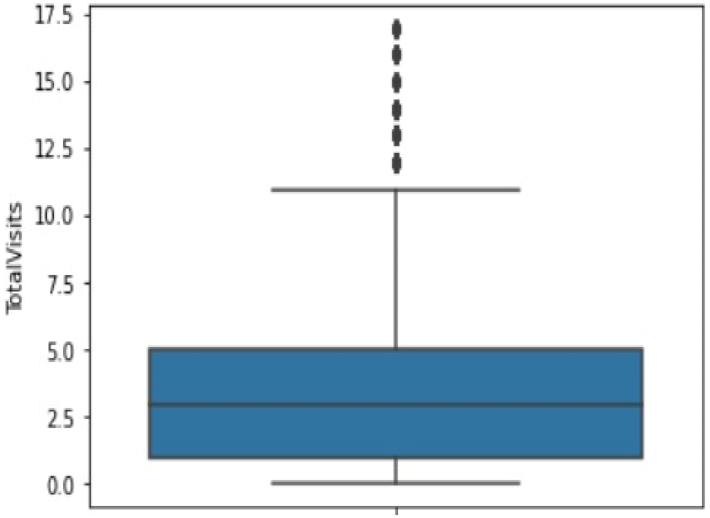
The CEO wants the lead conversion rate to be around 80%.

Steps:

- Understand the data.
- Clean and prepare the data.
- Exploratory Data Analysis.
- Feature Scaling.
- Splitting the data into Train-Test dataset.
- Building a Logistic Regression Model and calculating the Lead Score.
- Evaluating the model using Specificity and Sensitivity metrics.
- Applying the best model on the test data to make predictions.

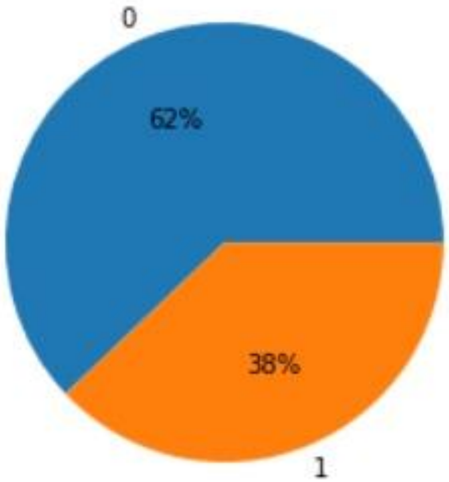
Data Cleaning and Preparation

- First we read and understood the data, the original data had 9240 rows and 37 different variables.
- We then went on the check the missing values in the data, and handled those by dropping the columns, imputing with modes or binning the data.
- The next step involved identifying using box plots and handling the outliers, we did so by capping the outliers for eg the outliers in 'Total Visits' column was capped at 1st and 99th percentile.
- The data was now clean and ready for EDA.



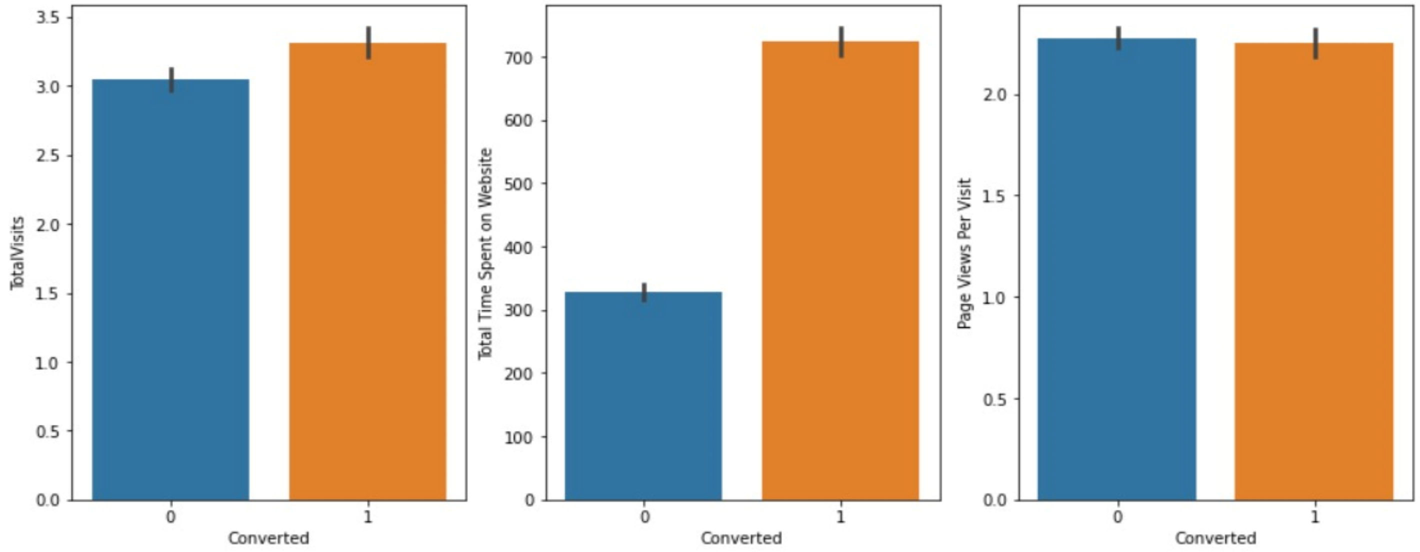
Box Plot showing the presence of outliers in 'Total Visits' column.

Exploratory Data Analysis:

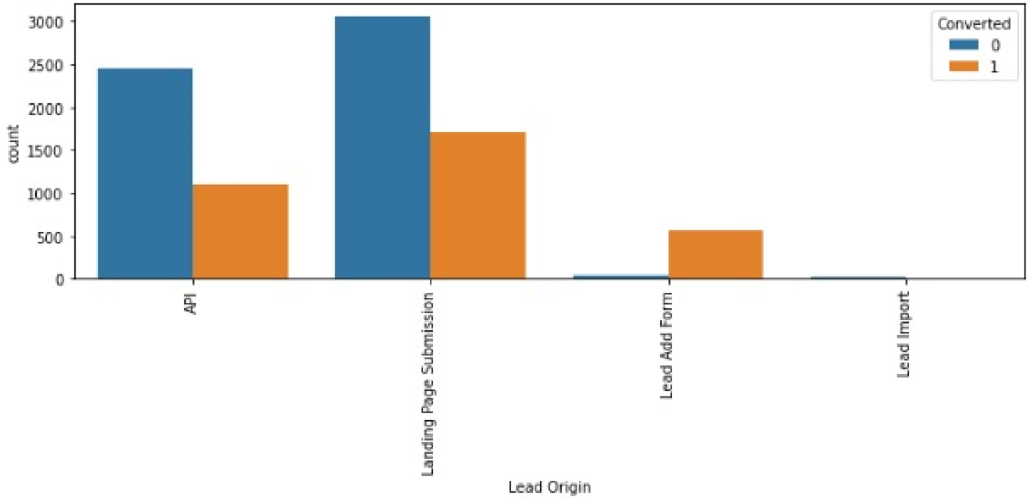


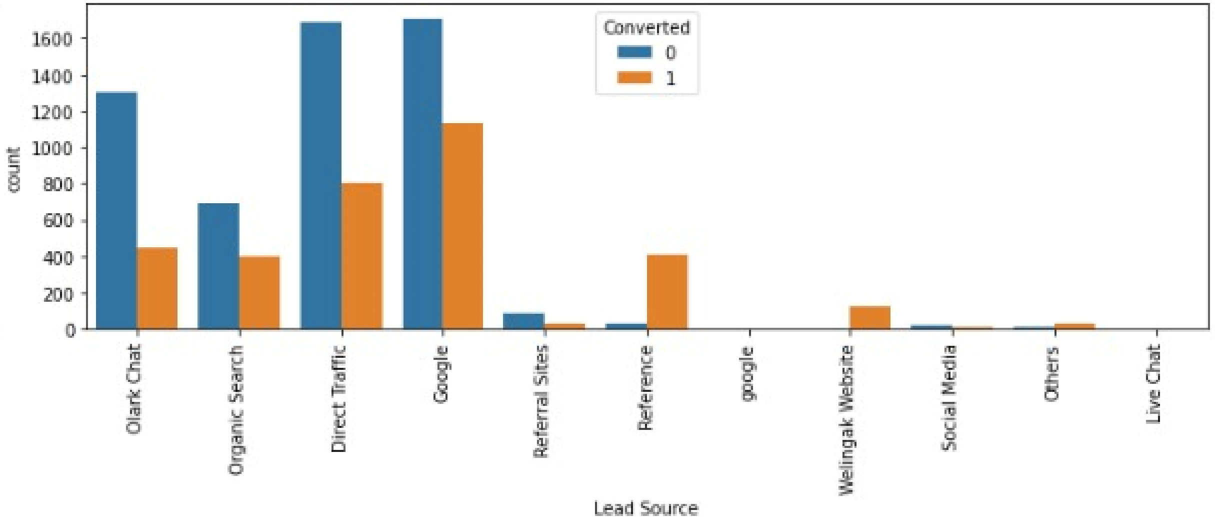
The conversion rate in around 38%

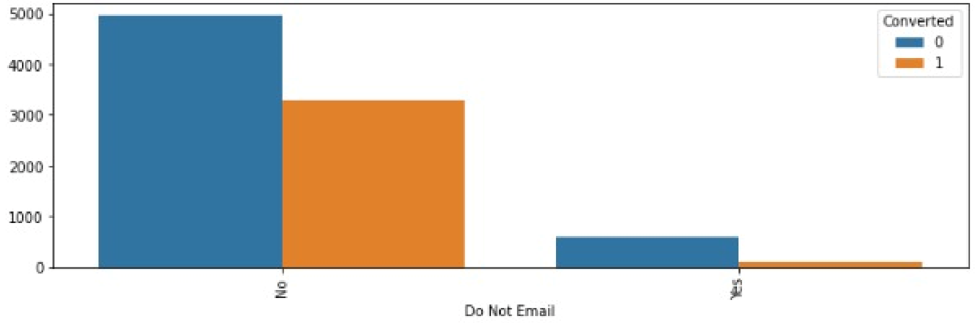
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.



EDA Part 2

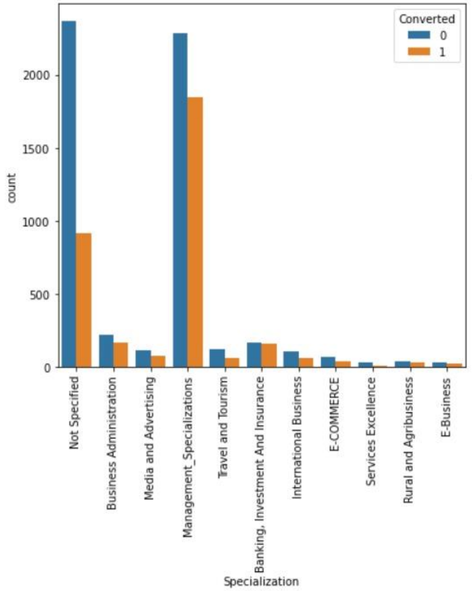


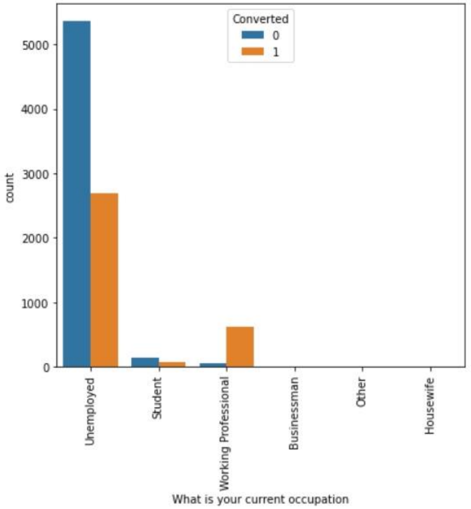




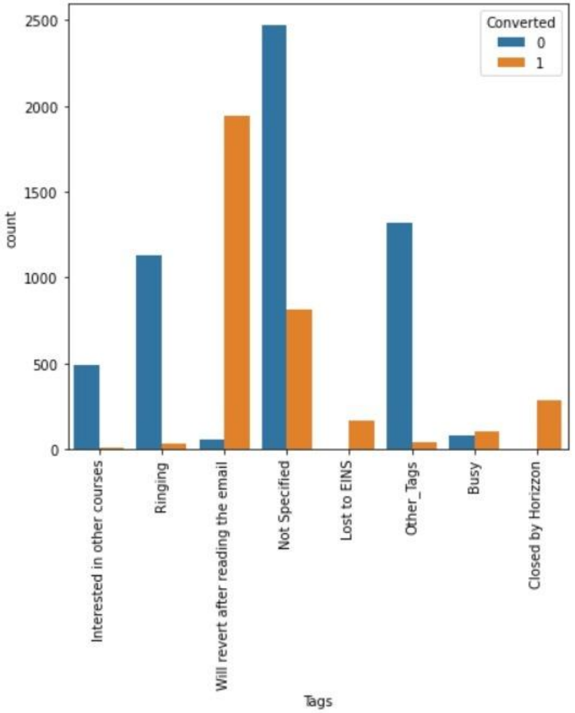
- In Lead Origin, maximum conversion came from Landing Page Submissions.
- In Do Not Email the conversion rate was high for those who opted no than those who opted yes.
- The highest conversion in Lead Source was from Google.

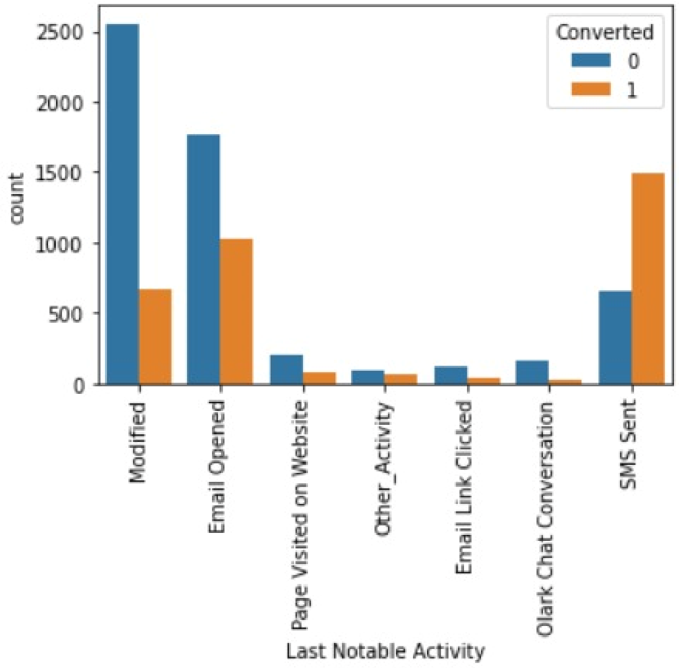
EDA Part 3





- In specialization the people in management specialization have a higher conversion rate.
- In occupation the unemployed people are more likely to get converted.





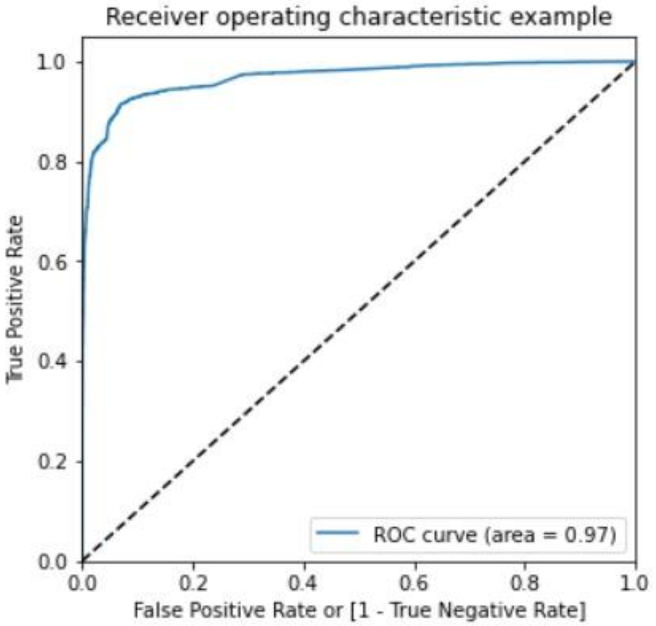
- The leads with tag will revert after reading the email had a higher conversion rate
- Also leads with Last Notable Activity as Email opened and SMS sent have high conversion rate

Assumptions:

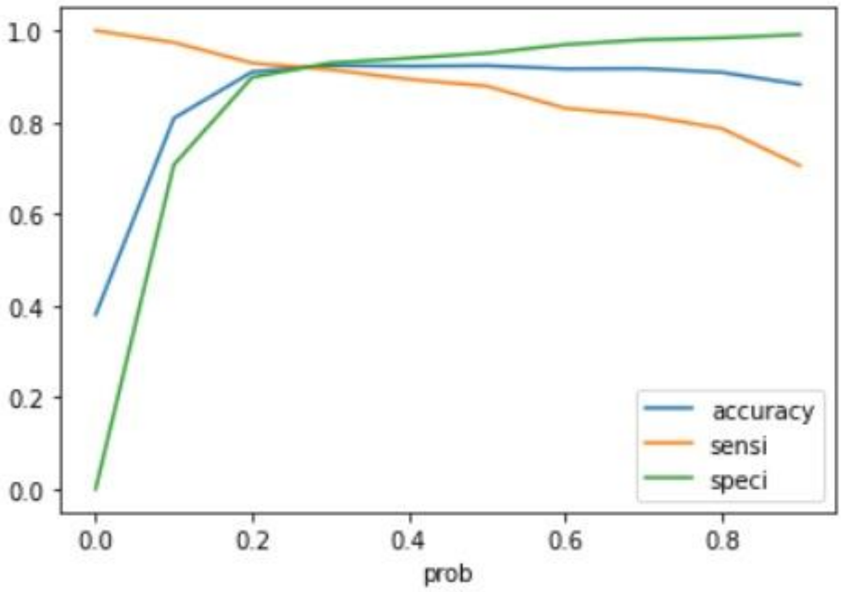
From the EDA we can make some assumptions about the variables that might impact the Conversion Rate:

- Total Visits
- Total Time Spent On Website
- Lead Origin – Landing Page Submission
- Lead Origin- Lead Add Form
- Lead Source- Google
- Lead Source- Olark Chat
- Lead Source- Reference
- Current Occupation – Unemployed
- Current Occupation- Working Professional
- Last Notable Activity- Sent a SMS

Model Evaluation



The Area Under the ROC curve is 0.97 which indicates that the model is good.



The accuracy, sensitivity and specificity curve shows the optimal cut-off as 0.3

3608	274
203	2182

CONFUSION MATRIX

Accuracy : 92.38%
Sensitivity:91.48%
Specificity:92.94%

Prediction on Test Set

1571	105
91	919

CONFUSION MATRIX

Accuracy: 92.70%
Sensitivity: 90.99%
Specificity:92.94%

CONCLUSION:

- The accuracy, sensitivity and specificity for the train and test sets are quite similar indicating a good predictive power of the model.
- The lead score calculated shows that the leads with lead score greater than 30 are more likely to convert as compared to those with lead score of less than 30.
- The top three variables that are likely to contribut to the lead getting converted according to our model are:

  - Tags
  - Lead Source
  - Last Notable Activity

- The Area Under the Curve of the ROC and the overall metrics of the test data set suggest that this is a good model.