Lead Scoring Case Study

BY:

KANISHK JAIN

KUMAR SARANSH

SHRINIVAS KASHEENATH SUREBAN

Table of Contents

Background of X Education Company

- X Education sells online courses to industry professionals.
- Professionals frequently visit the website to browse courses.
- The firm promotes its courses on many websites and search engines, including Google.
- Visitors to the website can browse courses, complete forms, and view videos.
- Individuals that provide their email address or phone number on a form are considered leads.
- Once these leads are collected, salespeople begin making calls, sending emails, and so on.
- Through this procedure, some leads are converted, but the majority are not.
- The average lead conversion rate at X Education is roughly 30%.

Problem Statement & Objective of the Study

Problem Statement:

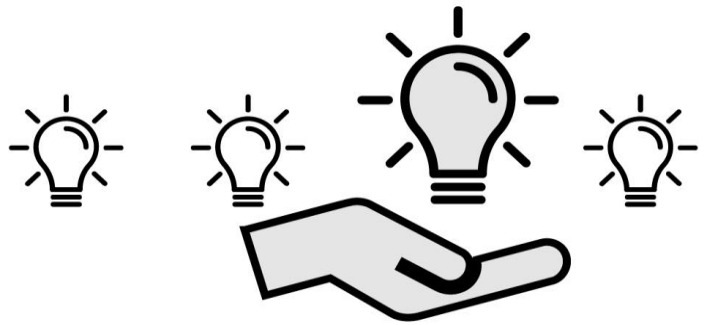
- Despite receiving a high volume of leads, X Education's conversion rate is just about 30%.
- X Education aims to improve lead conversion efficiency by identifying high-potential prospects, commonly known as Hot prospects.
- The sales staff will prioritize talking with these leads instead of calling everyone.

Objective of the Study:

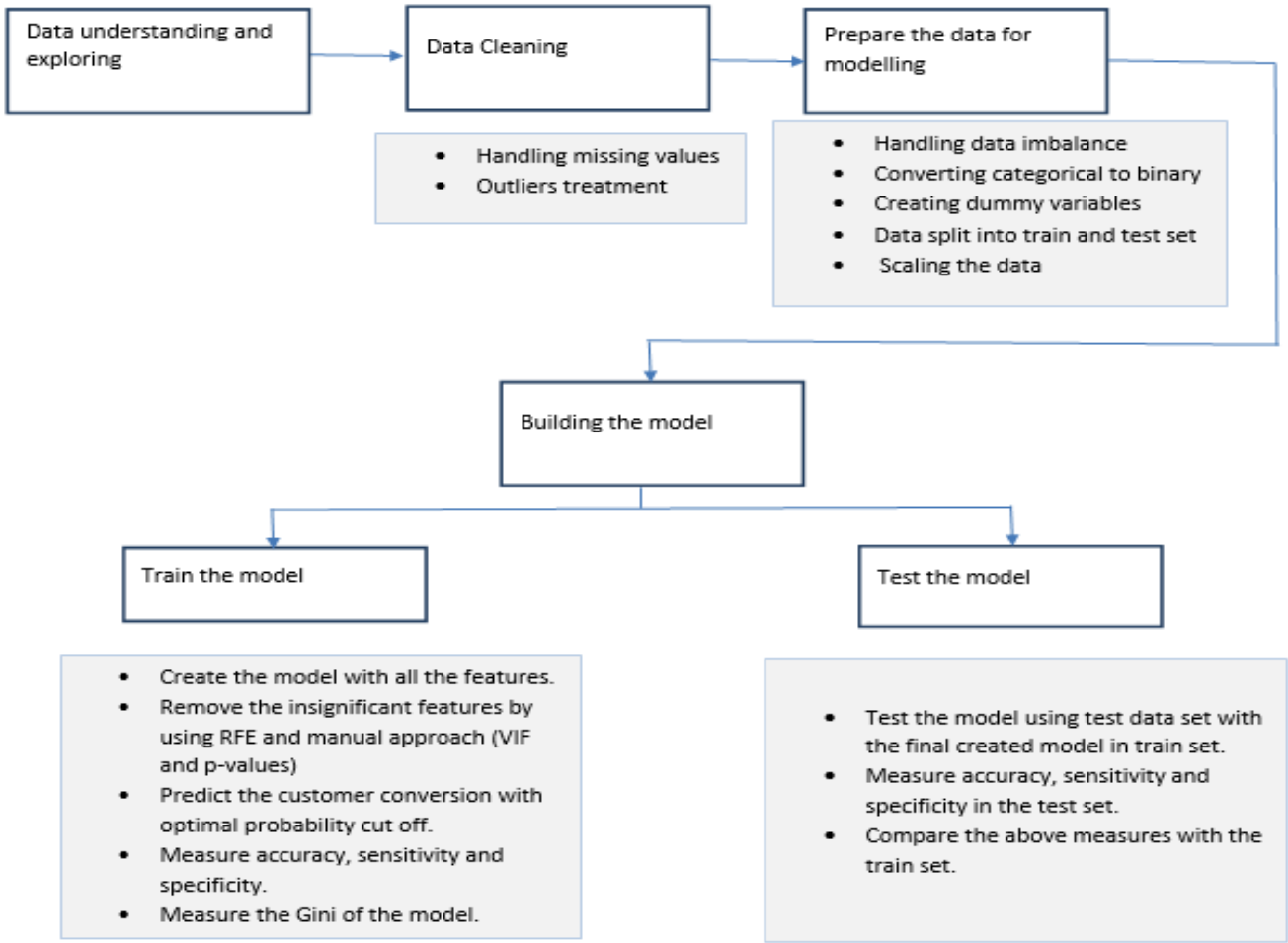
- The study aims to assist X Education in identifying the most promising leads, which are likely to become paying clients.
- The company wants us to create a model that assigns a lead score to each lead.
- Customers with a higher lead score have a higher conversion chance, while those with a lower lead score have a lower conversion chance.
- The CEO has set a target lead conversion rate of around 80%.

Suggested Ideas for Lead Conversion

- Leads are categorized according to their propensity or potential to convert. This produces a targeted set of hot leads.
- We may have a smaller pool of leads to communicate with, allowing us to have a bigger effect.
- We would have a higher conversion rate and be able to meet the 80% goal since we focused on hot leads who were more likely to convert.
- Because we want to achieve an 80% conversion rate, we need to be very sensitive while getting hot leads.



Analysis Approach



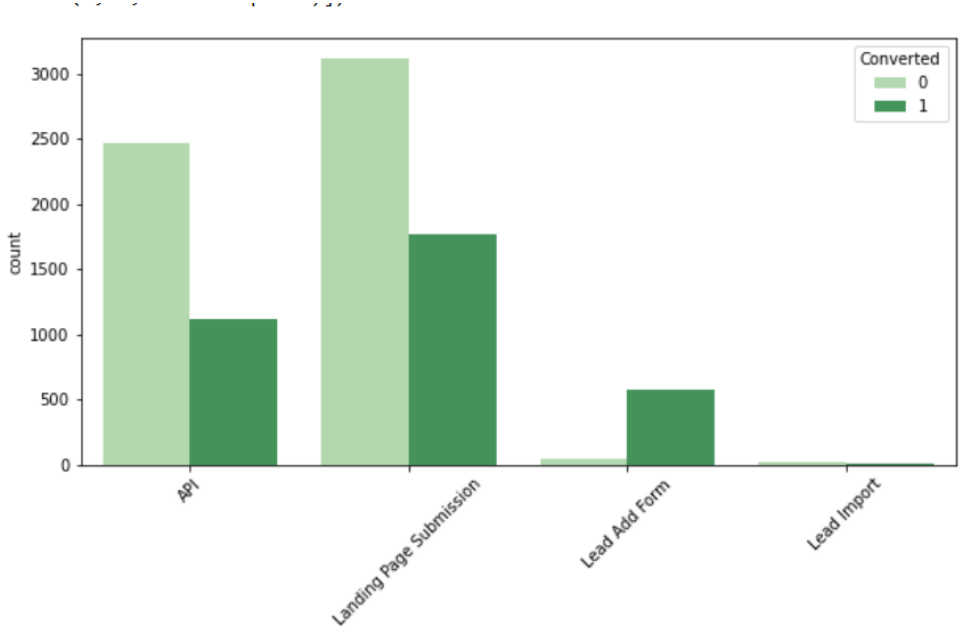
Data Cleaning

- The "Select" level indicates null values for some category variables when consumers did not select any options from the list.
- Columns with more than 40% null values were removed.
- Missing values in category columns were handled using value counts and specific considerations.
- Remove columns that do not contribute to the research aim (e.g. tags, countries).
- Imputation was employed for some category variables.
- New categories were developed for some variables.
- Removed columns such as Prospect ID and Lead Number that were not useful for modeling or had just one answer category.
- Mode was used to impute numerical data following distribution checks.
- Skewed category columns were identified and removed to prevent bias in logistic regression models.
- Outliers in Total visits and views per Visit were identified and limited.
- Fixed invalid values and normalized data in various columns, including lead source.
- Low-frequency values were categorized as "Others".
- Binary category variables were mapped.
- Additional cleaning steps were taken to assure data quality and accuracy.
- Fixed invalid values and standardized data in columns by verifying casing styles, etc. (lead source has Google, google)

EDA

1. Plotting a count plot of Lead origin w.r.t the target variable converted

- Although landing pages submission has nearly 1000 conversion rate,and API has 1700, they nonetheless generate a sizable number of leads.
- Although the Lead Add Form has a conversion rate of above 500, the number of leads is not very large in comparison.
- There are extremely little lead imports

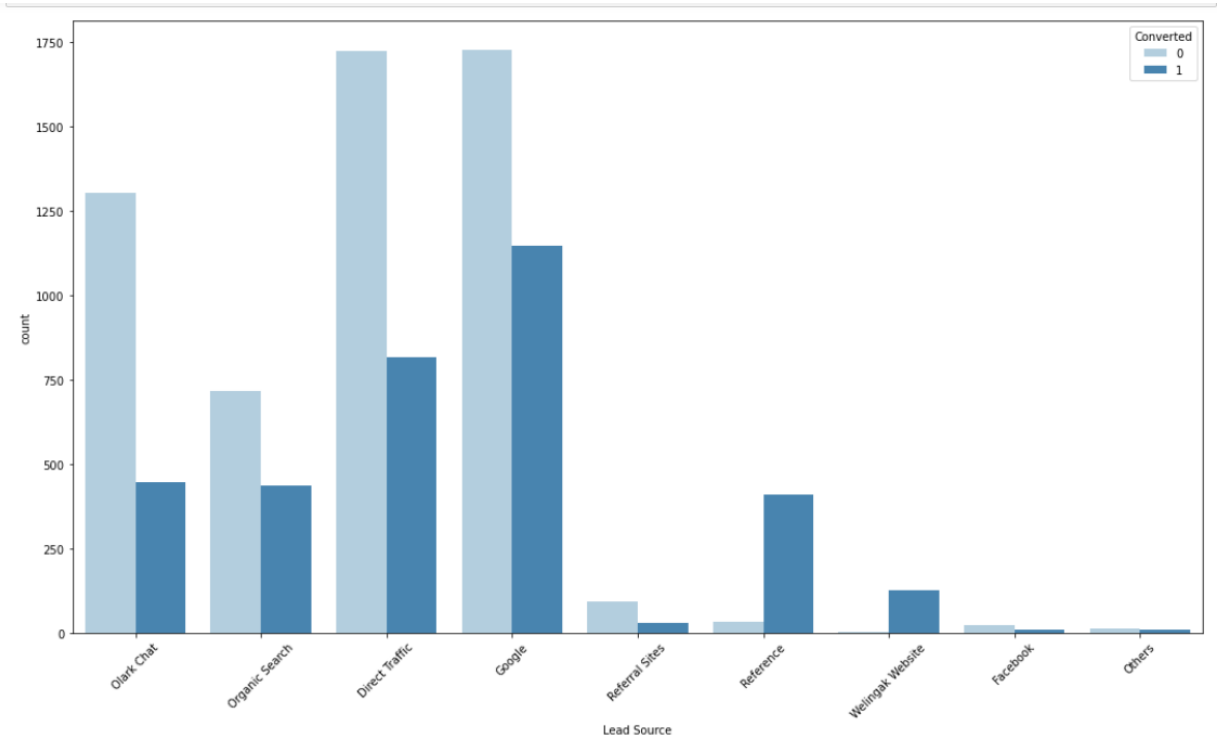


2. Plotting count plot for lead source

- The greatest amount of leads are generated by Google and Direct traffic.

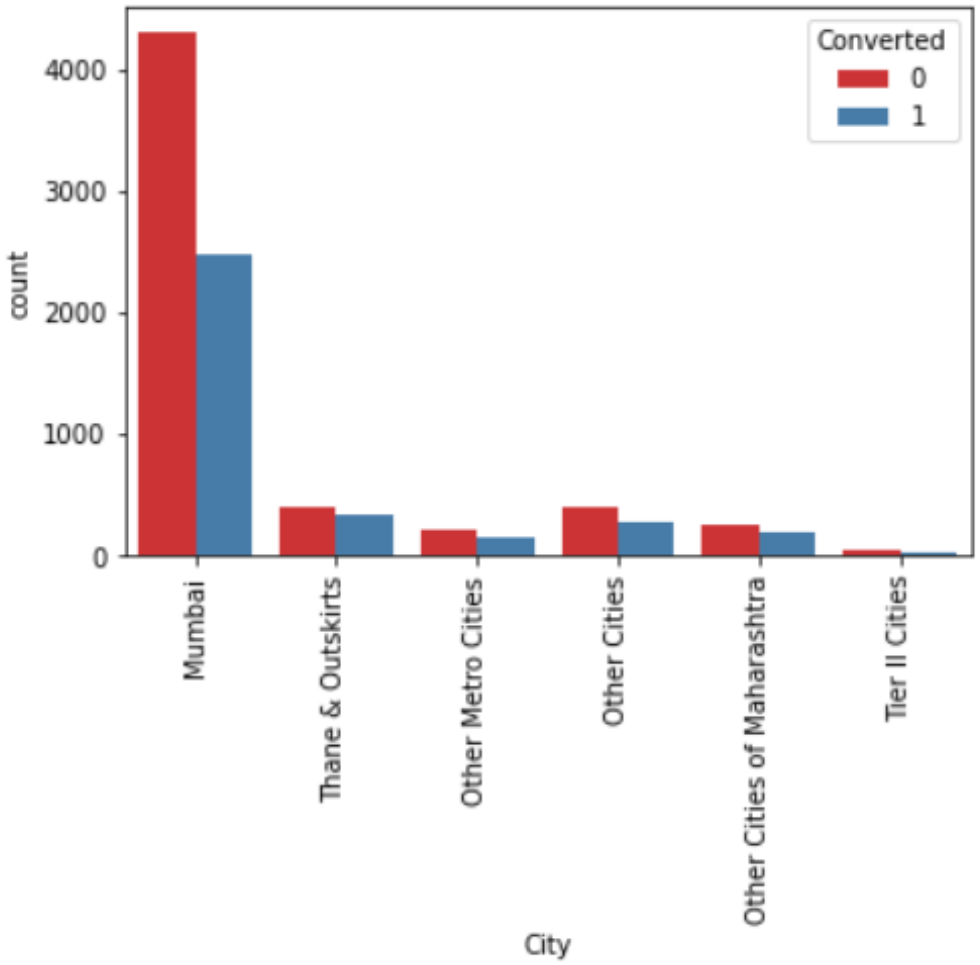
- Reference leads and leads obtained from the Welingak website have a good conversion rate.

- Focus should be placed on increasing lead conversion from Google leads, olark chat, organic search, direct traffic, and Google leads. More leads should be generated via the Welingak website and references.



3. Plotting a count plot of City w.r.t the target variable converted

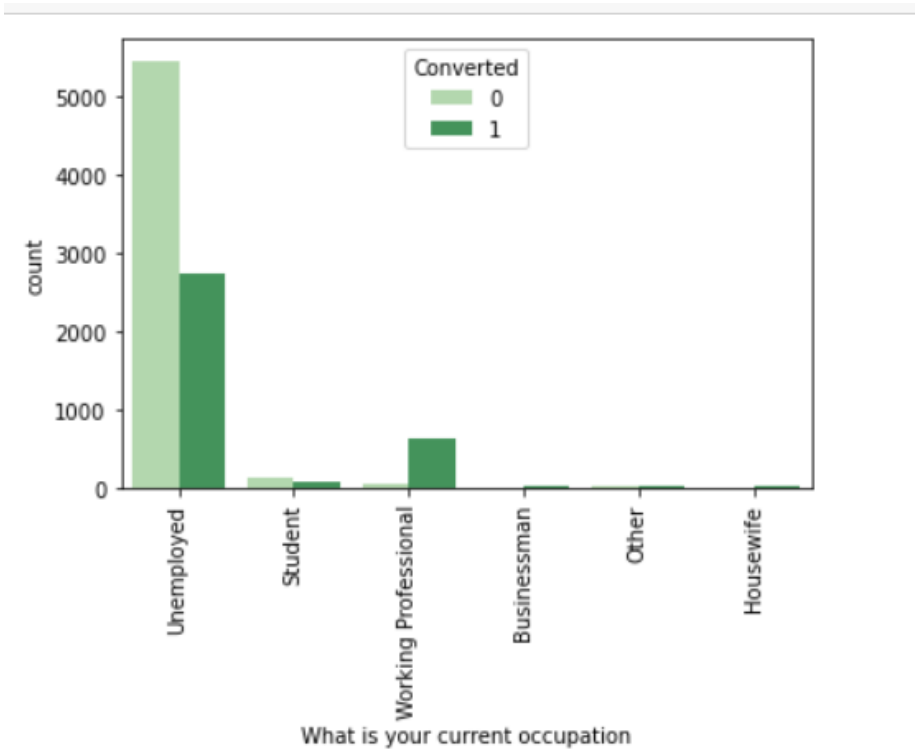
- Most of the conversions are 0 i.e. No.
- Mumbai is the only city with highest conversion rate among all



4. Plotting a count plot of What is your current occupation w.r.t the target variable converted

- Maximum visitors are unemployed but the conversion rate is nearly half.

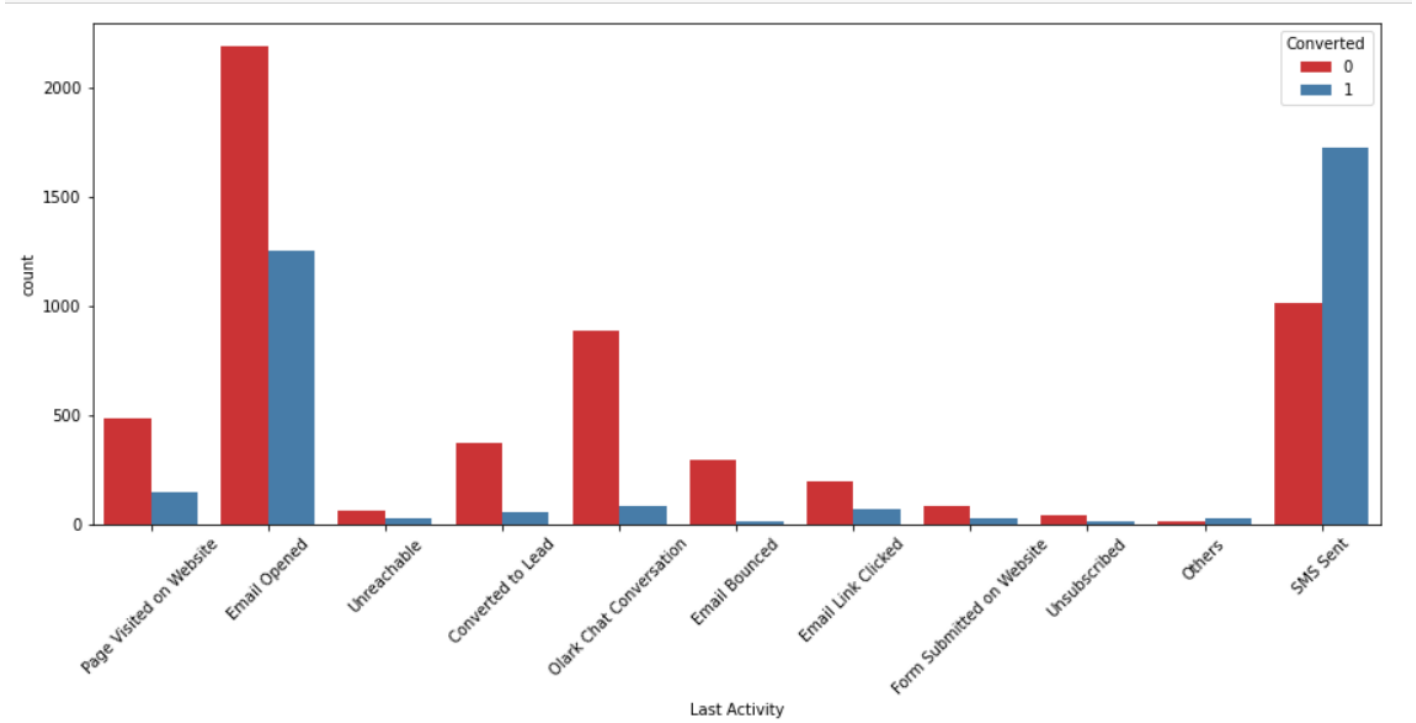
- The highest number of conversions are from the working professionals.



5. Plotting a count plot of Last Activity w.r.t the target variable converted

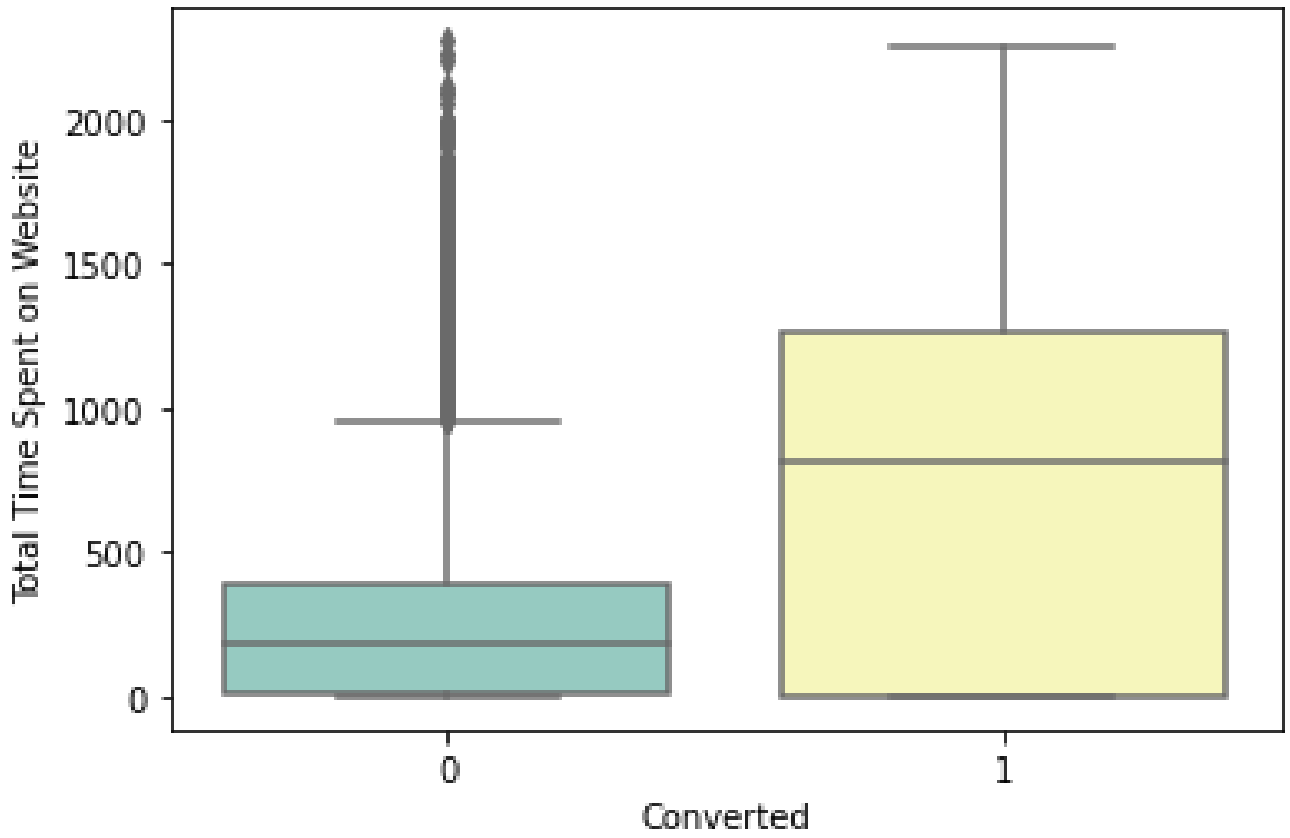
- The majority of leads have opened their emails as their most recent action, but showed maximum as not converted.

- Leads with the most recent activity being an SMS sent have a nearly 1700 conversion rate.



6. plotting a box plot of Total Time Spend on Website w.r.t the target variable converted

Longer sessions on the website increase the likelihood of conversion for leads.



Data Preparation before Model building

- In earlier phases, binary-level category columns were mapped to 1 / 0.
- Additionally, dummy features (one-hot encoded) were created for categorical variables such as Lead Origin, Lead Source, Last Activity, Specialization, and Current_occupation.
- To split the train and test sets, a 70:30% ratio was used.
- Features were scaled using the standardization method.
- Correlations were checked, and predictor variables with high correlations (Lead Origin_Lead Import and Lead Origin_Lead Add Form) were removed.

Model Building

- The data set includes several dimensions and a huge number of characteristics.
- To improve model performance and minimize calculation time, consider using Recursive Feature Elimination (RFE) to choose just significant columns.
- Manually fine-tune the model.
- RFE outcome: 48 columns pre-RFE and 15 columns post-RFE.
- The Manual Feature Reduction procedure was used to generate models by removing variables with p-values larger than 0.05.
- Model 9 is stable after four iterations, with significant p-values (p-values < 0.05) and no indication of multicollinearity (VIFs < 5).
- We will use logm9 as the final model for evaluation and prediction.

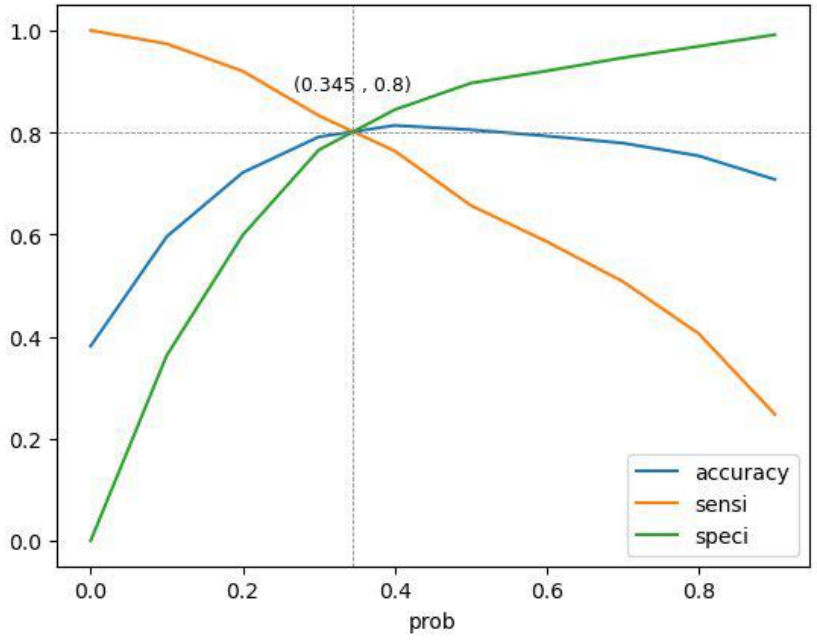
Model Evaluation

Train Data Set

Confusion Matrix & Evaluation Metrics with 0.345 as cutoff:

array([[3144, 809],
       [447, 1972]])

- Accuracy : 0.802887633396108 = 80.3%
- Sensitivity : 0.8152128978916908 = 81.5%
- Specificity : 0.7953453073614976 = 79.5%
- False Positive rate : 0.2046546926385024 = 20.5%
- Positive Predictive Value : 0.7090974469615247 = 70.9%
- Negative Predictive Value : 0.8755221386800334 = 87.6%
- Precision : 0.7920978363123237 = 79.2%
- Recall : 0. 8005554361306325 = 80%



Model Evaluation

Test Data Set Prediction:

Confusion Matrix:

```
array([[1359,  330],
       [179,  863]])
```

After running the model on the Test Data , we obtain:

- Accuracy : 0.8136213841083852 = 81.4%
- Sensitivity : 0.8282149712092131 = 82.8%
- Specificity : 0.8046181172291297 = 80.5%

Recommendation based on Final Model

- Lead Origin_Lead Add Form (4.013872): It appears that forecasts are much improved by this feature. It suggests that leads obtained via the lead add form have a higher conversion rate. As a result, the business could wish to put more of an emphasis on and resources into this lead generation channel.
- What is your current occupation_Working Professional (2.607540) This implies that professional leads have a higher conversion rate. To better appeal to this group, the business could modify its course offerings or marketing tactics.
- Last Notable Activity_Unreachable (2.047690): Predictions are significantly improved by leads who were identified as unreachable in their most recent notable activity. The business could find it beneficial to devise plans to reconnect with these prospects or investigate the reason behind their initial unreachability.
- Last Notable Activity_SMS Sent (1.906964): It appears that SMS sending increases conversions. To effectively engage leads, the business might think about adding SMS campaigns to their marketing mix.
- Last Activity_Others (1.855867): This category appears to have a good effect on predictions, but it's difficult to make particular recommendations without knowing exactly what "Others" entails. To find out what activities fit into this category and how they affect conversions, the organization needs look into it more.
- Total Time Spent on Website (1.078158): Predictions benefit from the amount of time spent on the website. It suggests that leads have a higher chance of converting if they spend longer time on the website. The business might concentrate on raising website engagement and giving users useful information.

Recommendation based on Final Model

- Lead Source_Olark Chat (0.964421) is the primary source. Predictions benefit from leads obtained through Olark Chat. To take advantage of this lead source, the business could spend more money on chat-based customer service or sales campaigns.
- Last Activity_Email Opened (0.493766): Email openings have a beneficial effect on forecasts, but not as much as other indicators. However, conversion rates may still rise as a result of email marketing campaign optimization.
- Lead Origin_Landing Page Submission (-0.942038): This indicates that the leads that originate from landing page submissions do not contribute positively to forecasts. To increase conversion rates from this source, the business might need to evaluate the layout of their landing pages or their optimization techniques.
- Specialization_Others (-0.956577): As with "Last Activity_Others," it's difficult to provide particular recommendations in the absence of information about what exactly qualifies as "Others." To fully comprehend this category's effect on conversions, more research is required.
- Last Activity_Olark Chat Conversation (-1.034059): Remarkably, forecasts appear to suffer when the final activity is an Olark Chat discussion. The business might have to reconsider how they manage leads from this channel or evaluate how successful their chat discussions are.
- Do Not Email (-1.215419): Predictions suffer from leads who choose not to receive emails. Although it's crucial to honor preferences, the business should nonetheless look into other avenues of contact or devise non-intrusive strategies to encourage email subscriptions.


Thank
you