Lead Scoring Case Study

By

Shubham Kumar Gupta

Cohort ID:4401

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- When these people fill out a form with their email address or phone number, they are classified as leads.
- The typical lead to successful sale conversion rate at X education is around 30% which is very poor.
- *Objective*:
  - The company wants to identify the most potential leads, also known as 'Hot Leads' to increase the conversion rate by focusing more on these leads and improving the sales.
  - The company wants a model to be built which can assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Approach of the Analysis

- Data Understanding
- Data Cleaning
  - Removing Missing values
  - Imputing missing values
  - Dropping Imbalanced Variables/Columns
- Data Visualization
  - Univariate Numeric Analysis
  - Bivariate Numeric Analysis
  - Multivariate Numeric Analysis
- Data Preparation
  - Dummy Variables using One Hot Encoding
- Model making and Evaluation

Data understanding

The data dictionary includes attributes such as:

- Lead Source
- Total Time Spent on the Website
- Total Visits
- Last Activity, etc.

which may or may not be useful in ultimately deciding whether a lead will be converted or not.

The target variable is the column 'Converted', which tells whether a past lead was converted or not, where 1 means it was converted and 0 means it wasn't converted.

Data Cleaning

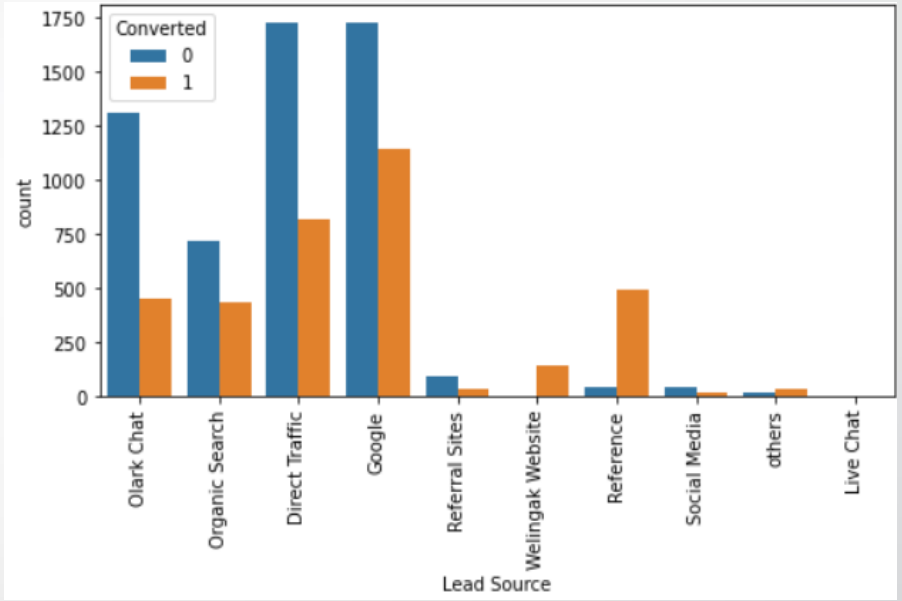
- There were total 9240 rows and 37 columns.
- After removing and imputing the missing values, there were 9103 rows and 28 columns.
- After dropping imbalanced columns, 9103 rows and 14 columns left.
- During univariate numeric analysis, the outliers were being removed after which 8953 rows and 14 columns left.

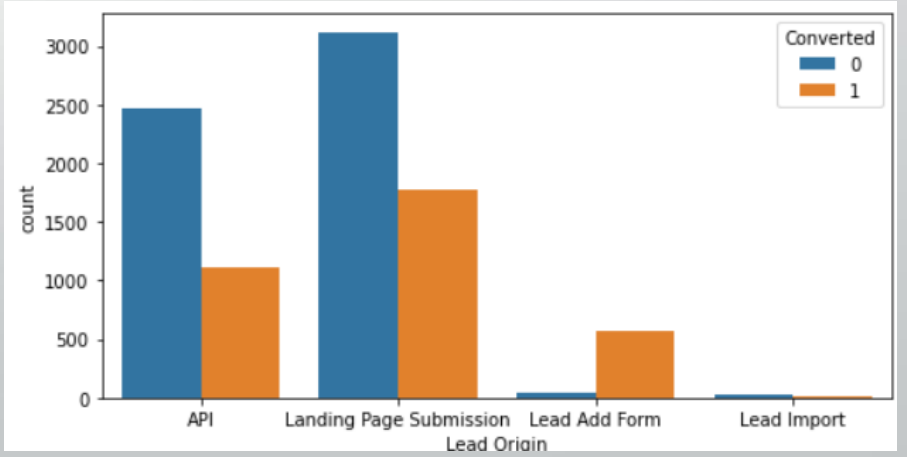
Visualization

Categorical Analysis

Inferences:

- The leads came through ‘Reference’ has high conversion rate.
- The leads with 'lead add form’ as their origin has a high conversion rate.
- Both have low counts, therefore more leads should be generated through these categories.

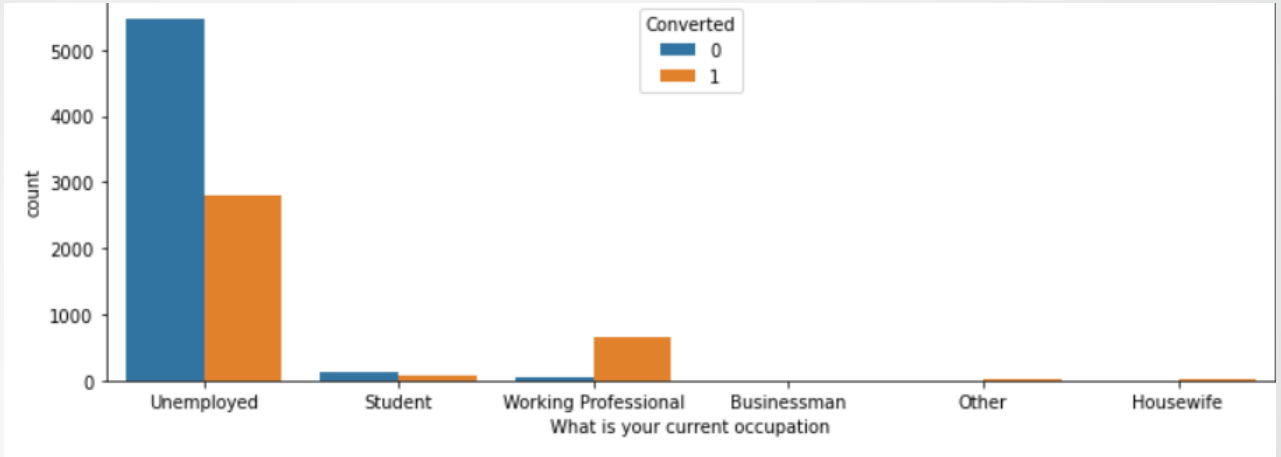


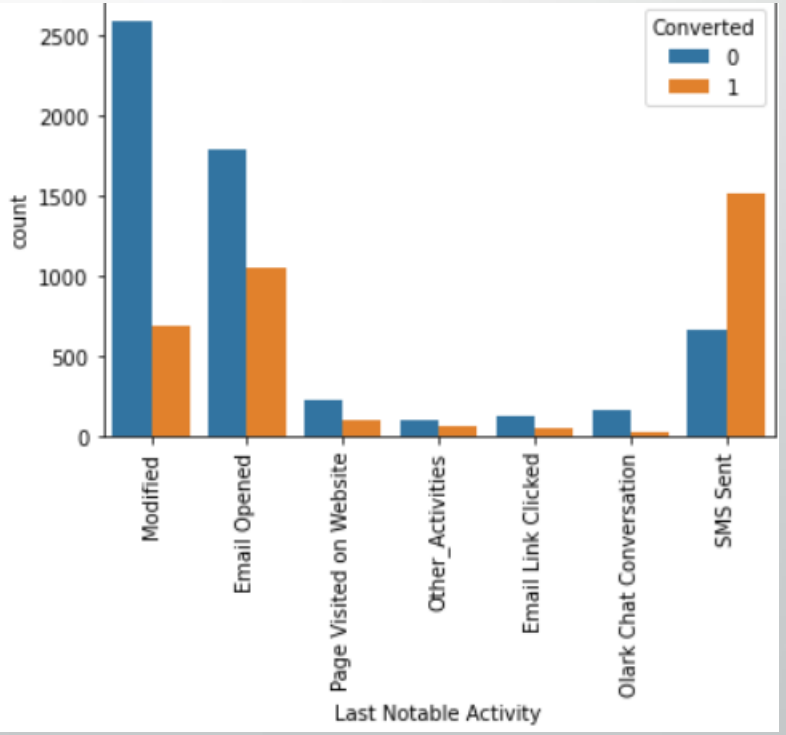


Categorical analysis

Inferences:

- The leads who are "working professionals" and have their last notable activity as "SMS sent" have high conversion rate.
- Therefore, more focus should be given on the leads from the aforementioned categories.

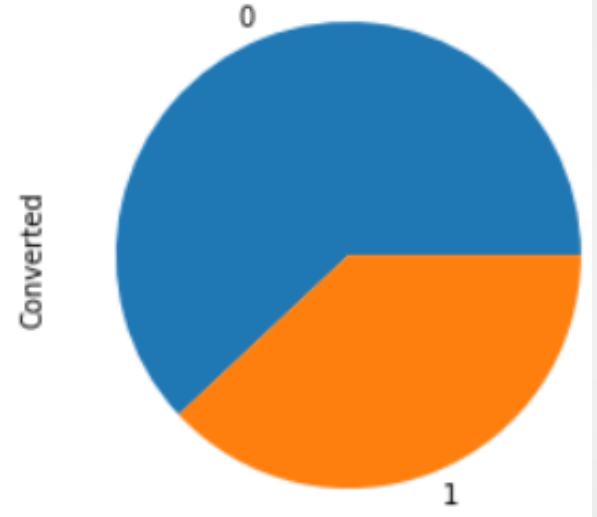


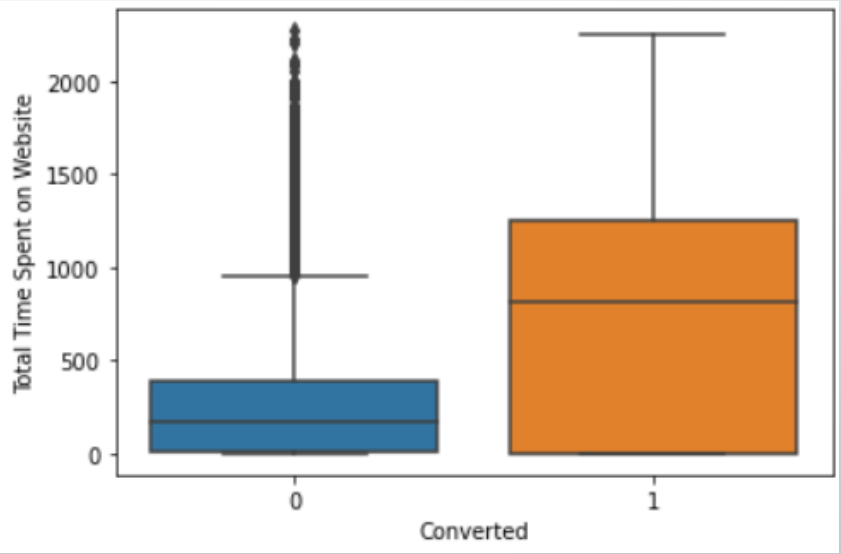


Numeric Analysis

Inferences:

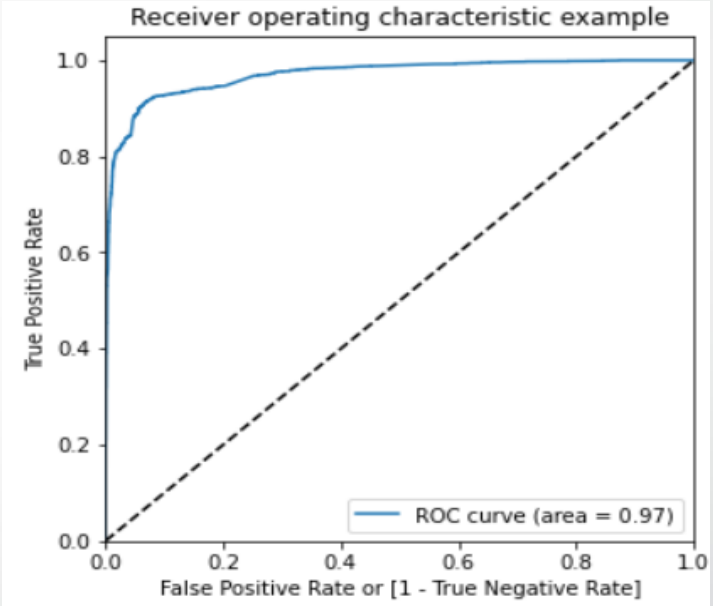
- The pie plot depicts that only 38.02% of the whole data are being converted.
- The box plot depicts that leads spending more time on Website have high conversion rate.
- Therefore, the website should be made more interactive for the users in order to increase the count of leads.



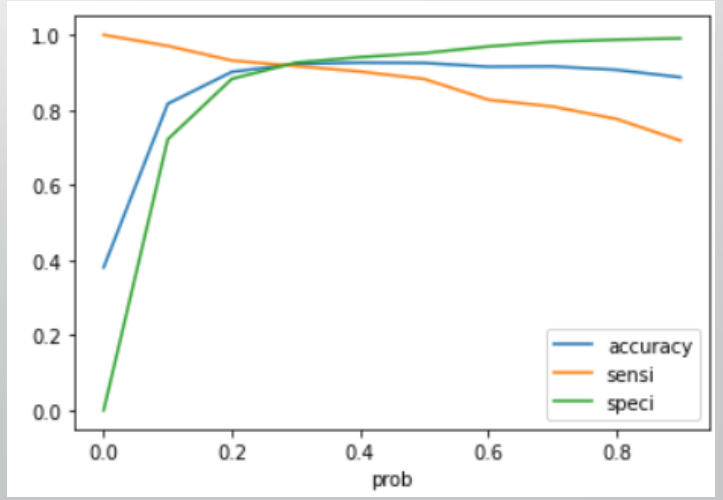


Model building

- Split of whole data into train and test data.
- Feature selection done by RFE by selecting 15 features.
- After running the model with the features selected by RFE, one feature i.e. Last Notable Activity_SMS Sent has highest VIF(~6.2) so after its removal and re-running the model rest of the features had good VIF(<3).
- Both times the arbitrary cut off taken as 0.5 upon which the predictions were made.
- ROC(Receiving operating characteristic) curve was being plotted. ROC area = 0.97
- The optimal cutoff was being found as 0.3 through the plot after which final prediction made.



ROC plot



Optimal cut off plot

Model Evaluation

After running the model on the Test data following results have been found:

Observations from Train data

Accuracy = 92.7%

Sensitivity = 91.78%

Specificity = 93.25%

Observations from Test data

Accuracy = 92.26%

Sensitivity = 91.69%

Specificity = 92.6%

This shows that the model is working well and predicting lead score at about 92% accuracy.

Recommendations based on Results

To increase the conversion rate and improve the sales, the sales team have to focus on the leads:

- spending most time on the Website.
- coming through Direct Traffic, Referral Sites and Welingak Website as their source.
- whom Origin is from Lead Add Form.
- Who have their Last Activity as SMS Sent.
- Those who have been tagged as Closed by Horizzon, Interested in other courses, Lost to EINS, Other_Tags, Ringing, Will revert after reading the email.
- Who have their Last Notable Activity as Modified and Olark Chat Conversation.