Lead Scoring Case Study

By
Laltesh Choudhary
Kunal Khandelwal
Royal Singh

INTERNAL

Table of Contents

Problem Statement & Objective of the Study

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 is 30%.

As you can see, there are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion.

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 a higher lead score have a higher conversion chance and the customers with a 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%.

Recommendations for Lead Conversion

- Focus on features with positive coefficients for marketing strategies.
- Develop strategies to attract leads from top-performing lead sources.
- Sales team should engage working professionals with SMS,messaging and Emails.
- Leads who have spent time on sending massages and opening the emails are also potential hot leads and hence, effective Communication channels should be used.
- Retention rate of existing leads needs to be identified from the time of enrolment.
- Social Media can be used as a potential source for understanding and interacting. It can be effectively used for marketing purpose as well by encouraging people to take action.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.
- Focus on leads with high potential : Based on our analysis coefficient factor was high for following :
    1) Total Visits : 8.63
    2) Total Time Spent on Website:4.5145
    3) Lead Origin_Lead Add Form:3.87
    4)What is your current occupation_Working Professional:2.46
    5)Lead Origin_Lead Add Form :3.8735
    6)Lead Source_Welingak Website:1.6791
    7)Last Activity_Had a Phone Conversation: 2.1141

Steps Followed

- Importing Necessary Libraries
- Data Loading & Preparations
- Missing Values Check
- Univariate and Bivariate Analysis
- Dummy Variable Creation
- Test-Train Split
- Scaling Features
- Model Building
- Model Predictions
- Finding the optimal cutoff

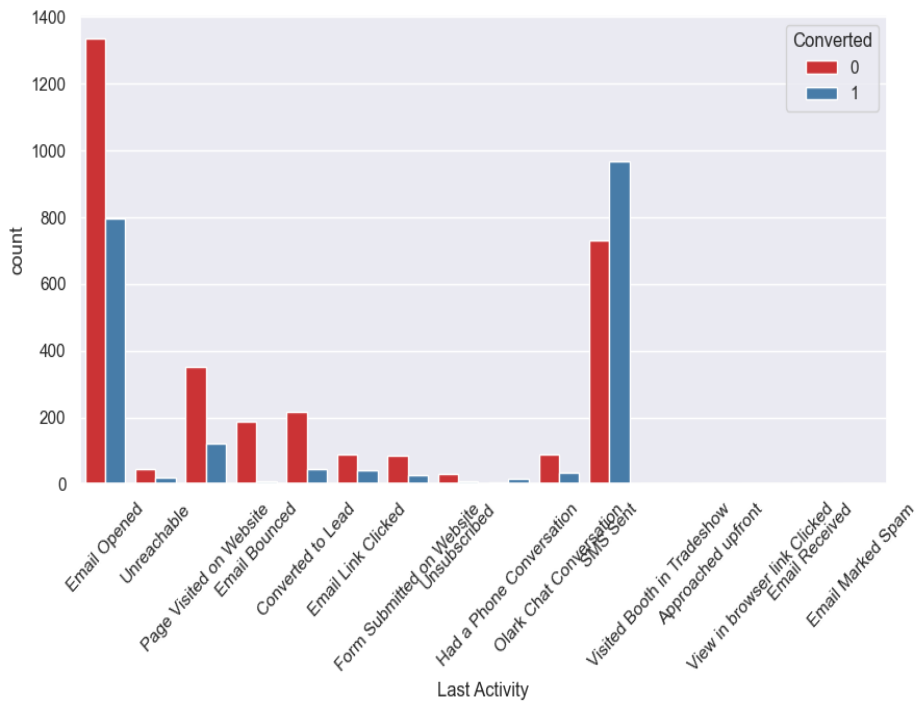
Data Cleaning

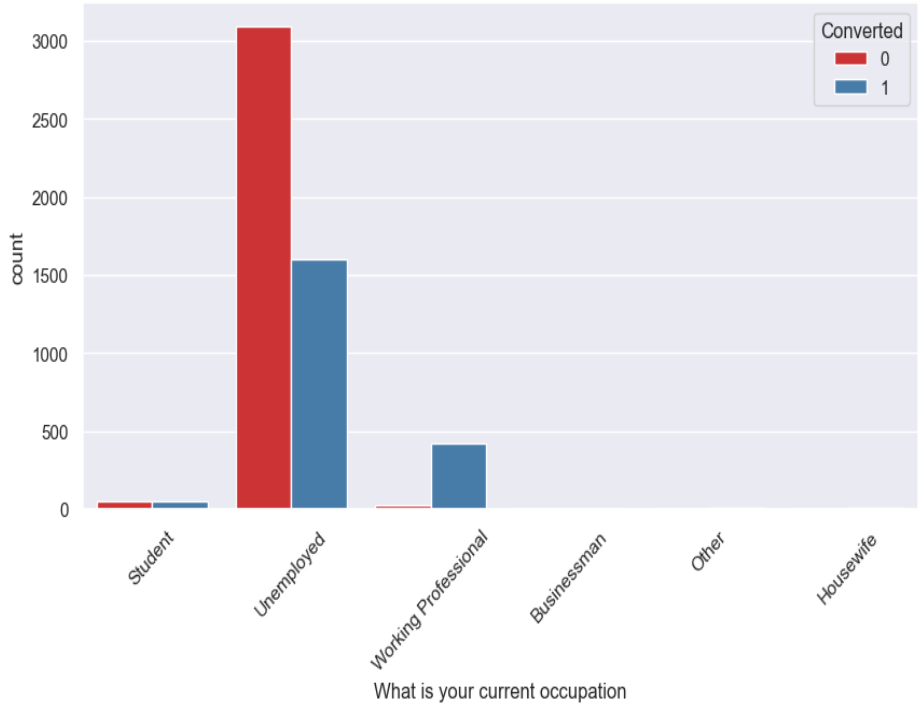
- “Select” level represents null value for some categorical variables, as customers did not choose any option from the list.
- There are few columns with over 40% null values were dropped
- We handled the missing values for categorical columne were based on Value counts
- We have dropped certain columns from our analysis which do not add insights useful for analysis i.e. City, tags, Country, Specialization, Newspaper etc.
- Columns with no use for modelling (Prospect ID, Lead number) can we have dropped.
- Other cleaning activities were performed to ensure data quality and accuracy
- Skewed category columns were checked and dropped for example. “Do not Call”, “Search”, “Magazine”, “Newspaper Article”, “Through Recommendations”, “Get updates on DM content” etc.

EDA

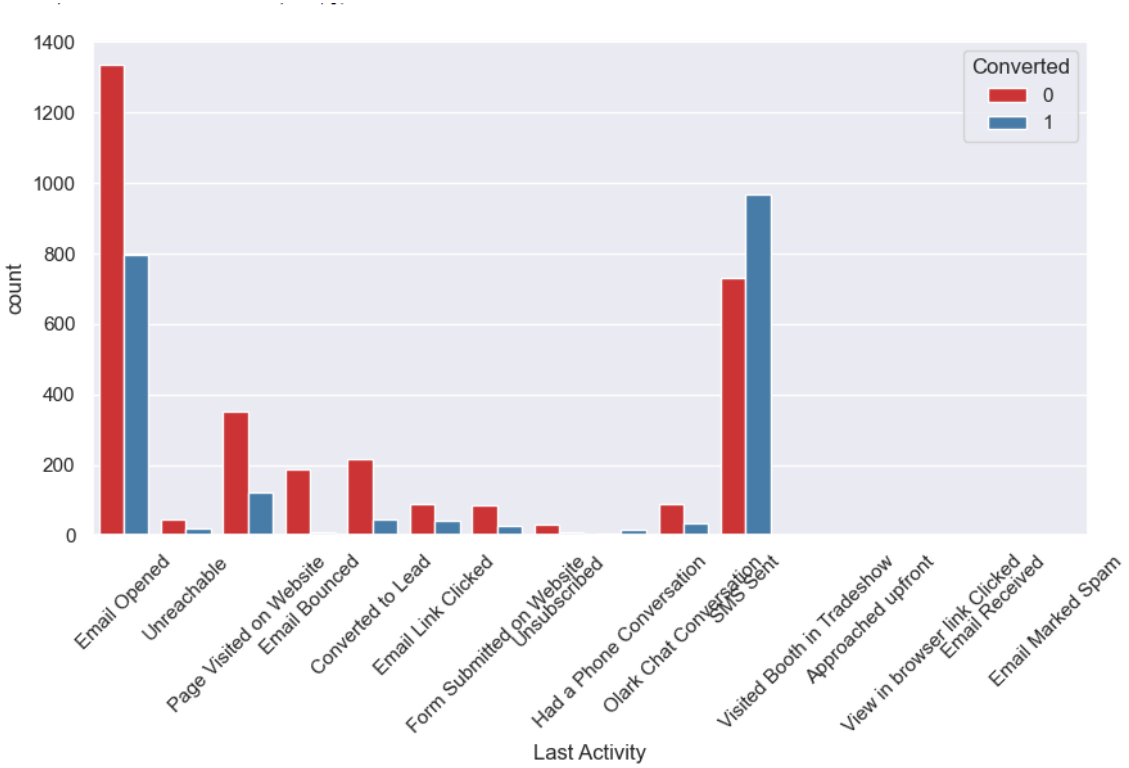
- Data is imbalanced while analyzing target variable.

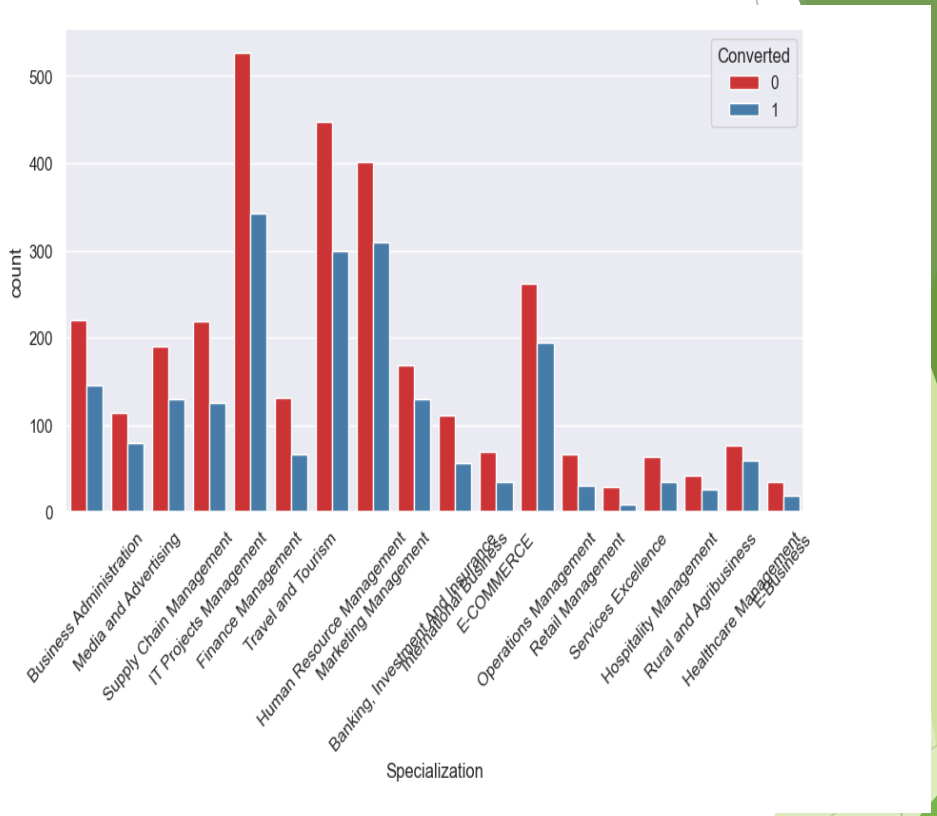
UNIVARIATE ANALYSIS FOT CATEGORICAL VARIABLE





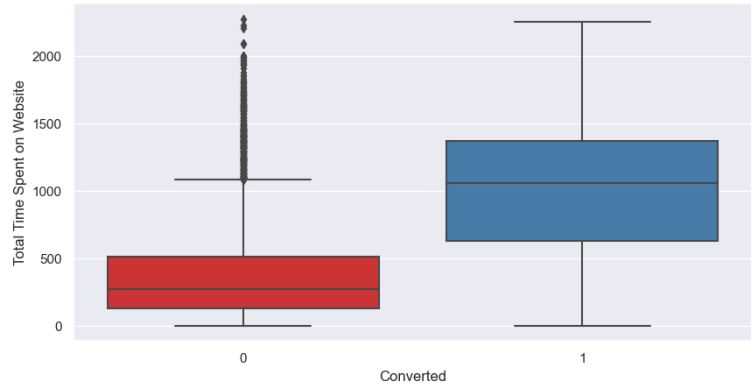
BIVARIATE ANALYSIS



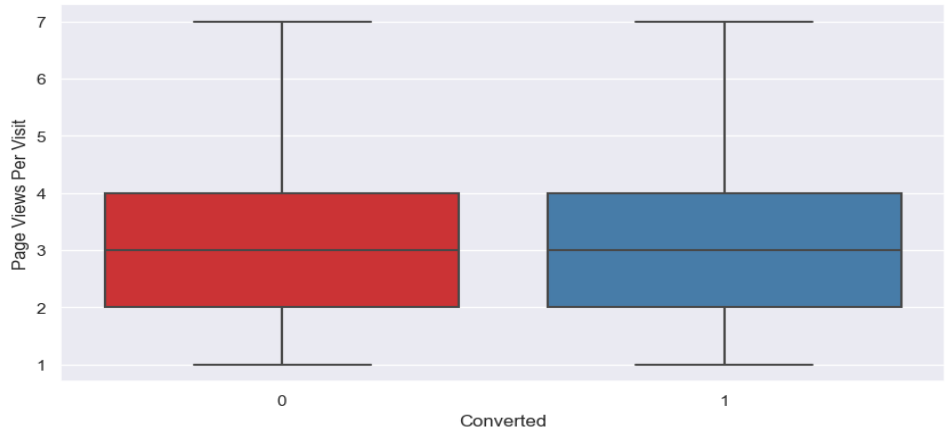


BIVARIATE ANALYSIS FOR NUMERICAL VARIABLE

```
In [566]: sns.boxplot(y = 'Total Time Spent on Website', x = 'Converted', data = df,palette='Set1')
Out[566]: <Axes: xlabel='Converted', ylabel='Total Time Spent on Website'>
```




```
In [571]: sns.boxplot(y = 'Page Views Per Visit', x = 'Converted', data =df,palette='Set1')
Out[571]: <Axes: xlabel='Converted', ylabel='Page Views Per Visit'>
```



Data Preparation before Model building

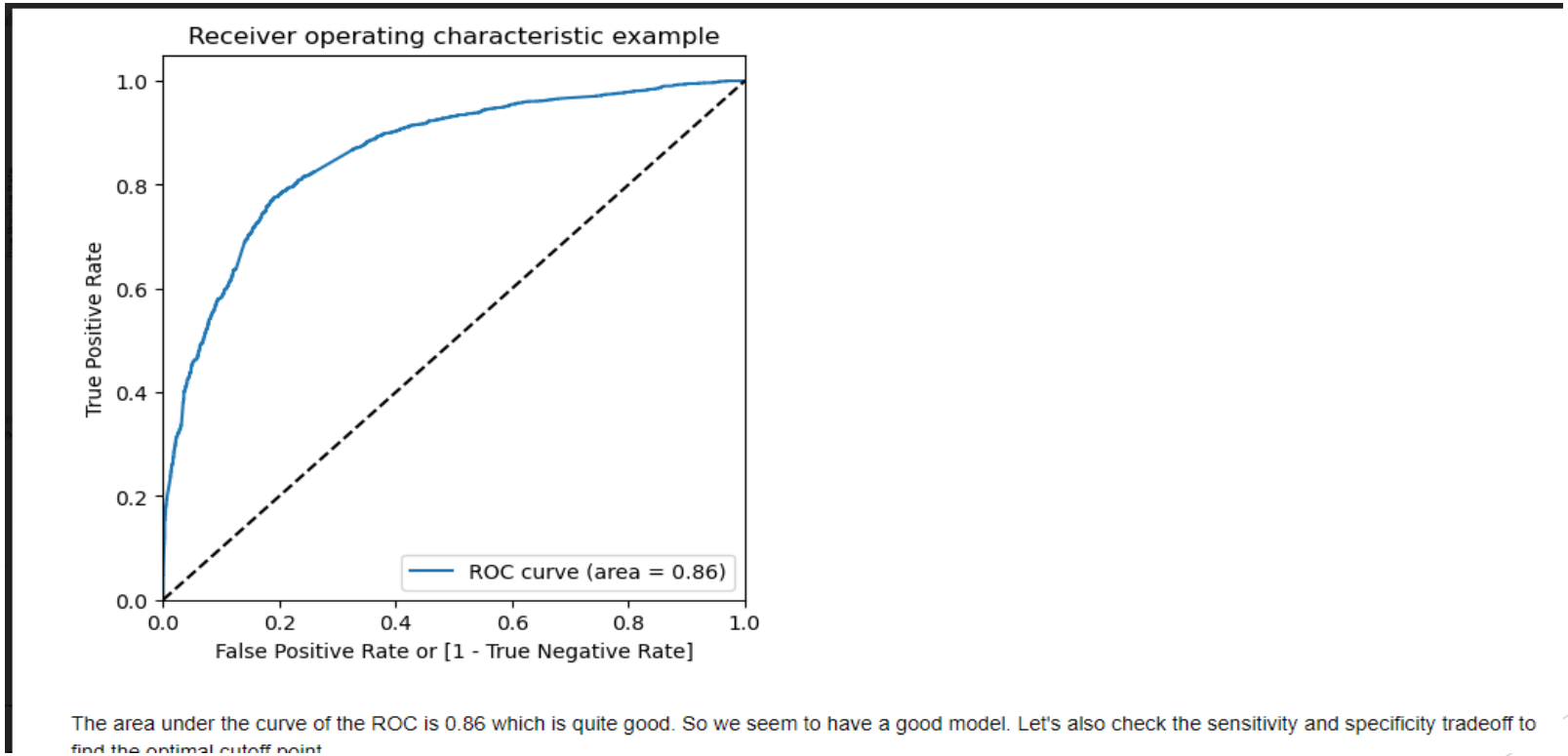
- Some categorical columns i.e. “Do not Email” were mapped to 1/0.
- Created dummy features (one-hot encoded) for categorical variables- Led Origin, Lead source, Last activity, current occupation
- Splitting the dataset into 70% train and 30% test
- There a few numeric variables present in the dataset which have different scales
- Correlations needs to be checked
- Variables with positive coefficient were used for predictive analytics
- Variables with VIF value greater than 0.5 are neglected
- Total 4 models were built before arriving at any conclusion
- Logm4 was selected as final model with 11 variables

Top 3 Models

- Total Visits
- Total Tim Spent on Website
- Lead origin Lead Add Form

ROC Curve

1. Area under ROC Curve is 0.86 out of 1 which indicates a good predictive model.
2. The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all the threshold values.



The area under the curve of the ROC is 0.86 which is quite good. So we seem to have a good model. Let's also check the sensitivity and specificity tradeoff to find the optimal cutoff point

Thank You