Lead Score Case Study

Submitted by :

Racherla Hemanth Raj

Lead Score Case Study for X Education

Problem Statement :

X Education sells online courses to industry professionals. 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%.

Business Goal:

X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

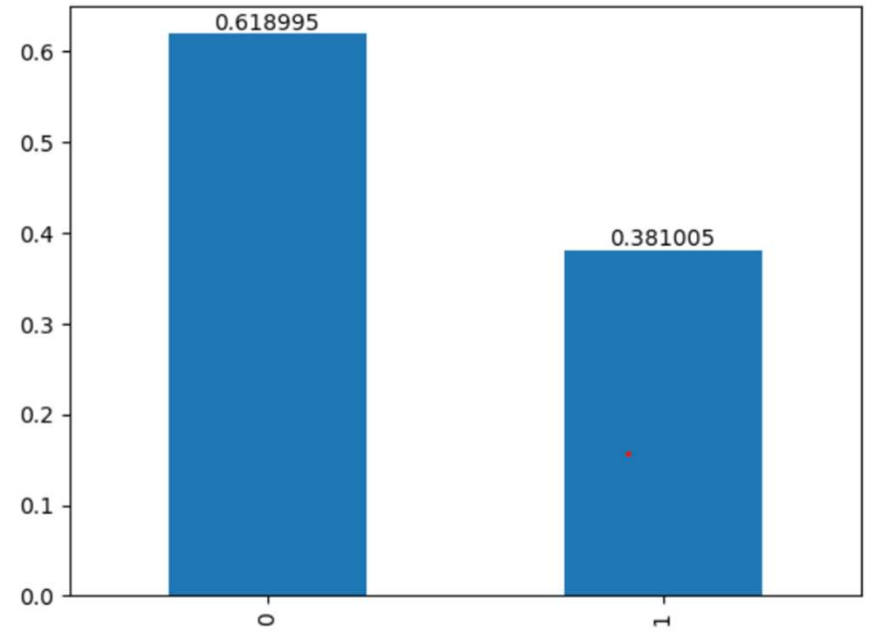
The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Problem Approach

- Sourcing the data for analysis
  1. Data has been provided for us in csv format
- Clean and prepare the data
  1. Some columns have values Select which are missing data so we convert them into null values
  2 .Checked for the null values and remove columns having missing values greater than 40
  3. Imputation is done for values and null values is removed from data
- Exploratory Data Analysis.
  1. Univariate and Bivariate analysis on dataset
- Feature Scaling
  1.Dummy value creation and Scaling is done
- Splitting the data into Test and Train dataset.
  1. Data is divided with 80 percent as Training data and remaining test data
- Build a logistic Regression model and calculate Lead Score.
- Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall.
- Applying the best model in Test data based on the Sensitivity and Specificity Metrics.

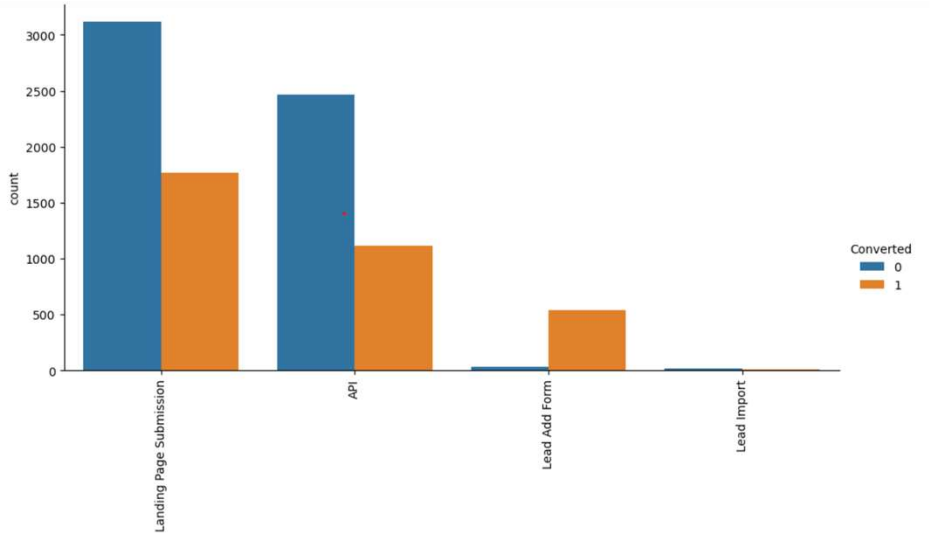
Conversion rate

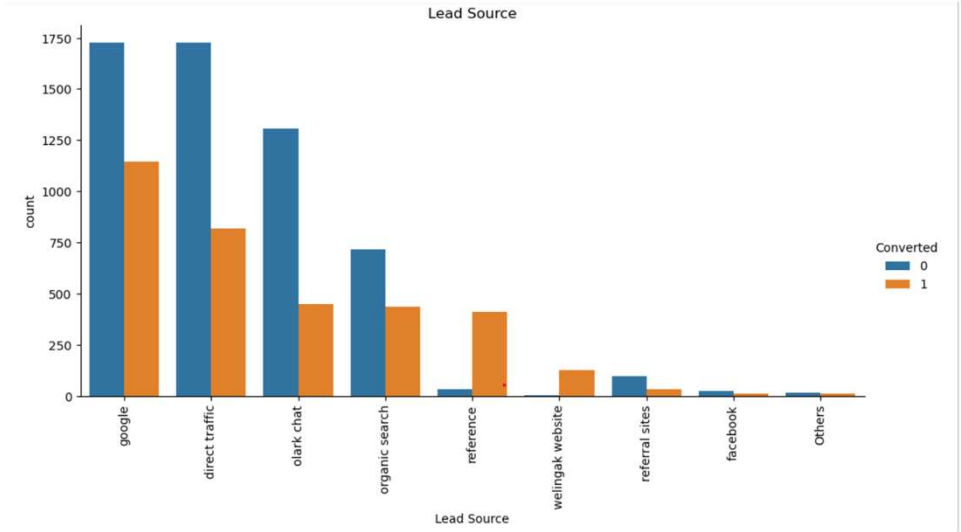


Our Conversion rate is nearly 39 percentage

EDA on Data set

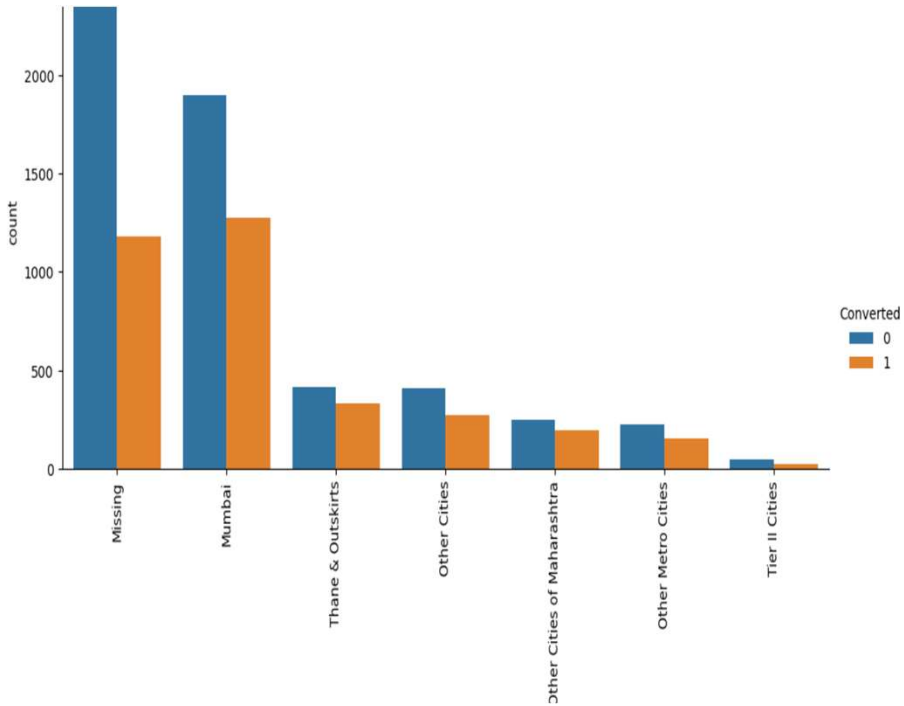
The Data set is first plotted on Categorical variables and some of the significant fields are lead source and origin



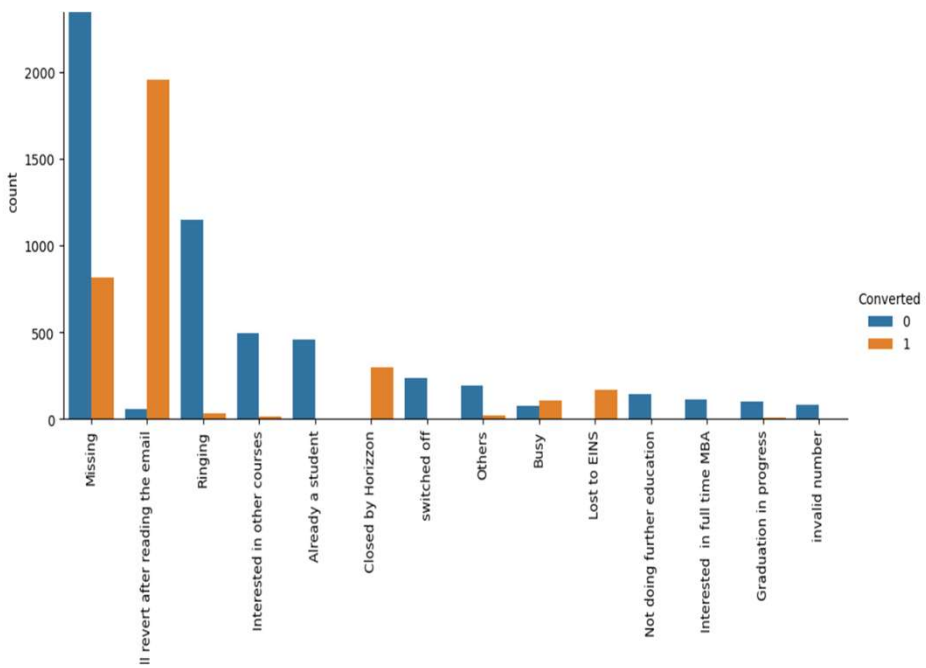


- Lead Add form and reference wellingak website have good conversions

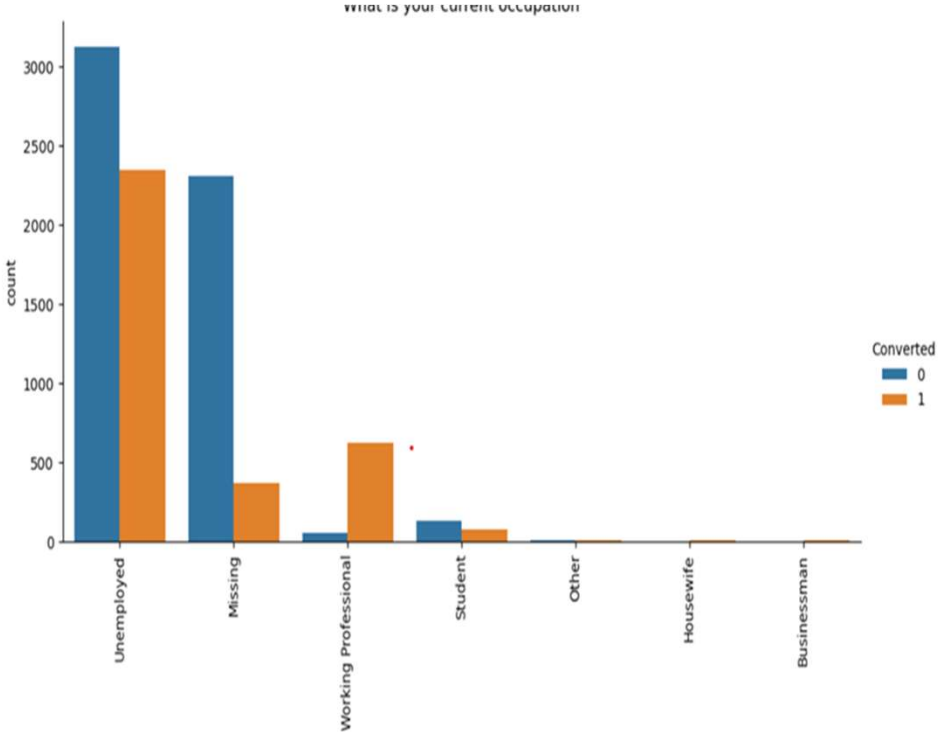
Mumbai is the major city for leads followed by other cities



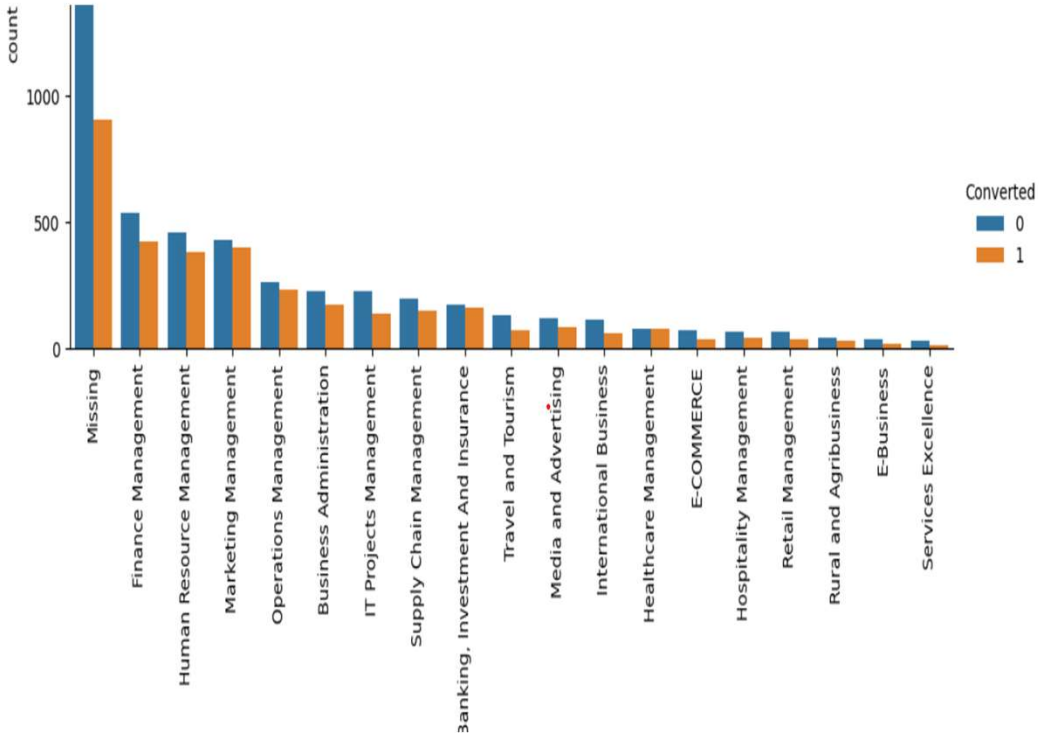
Major leads are reverting and closed by horizon tags



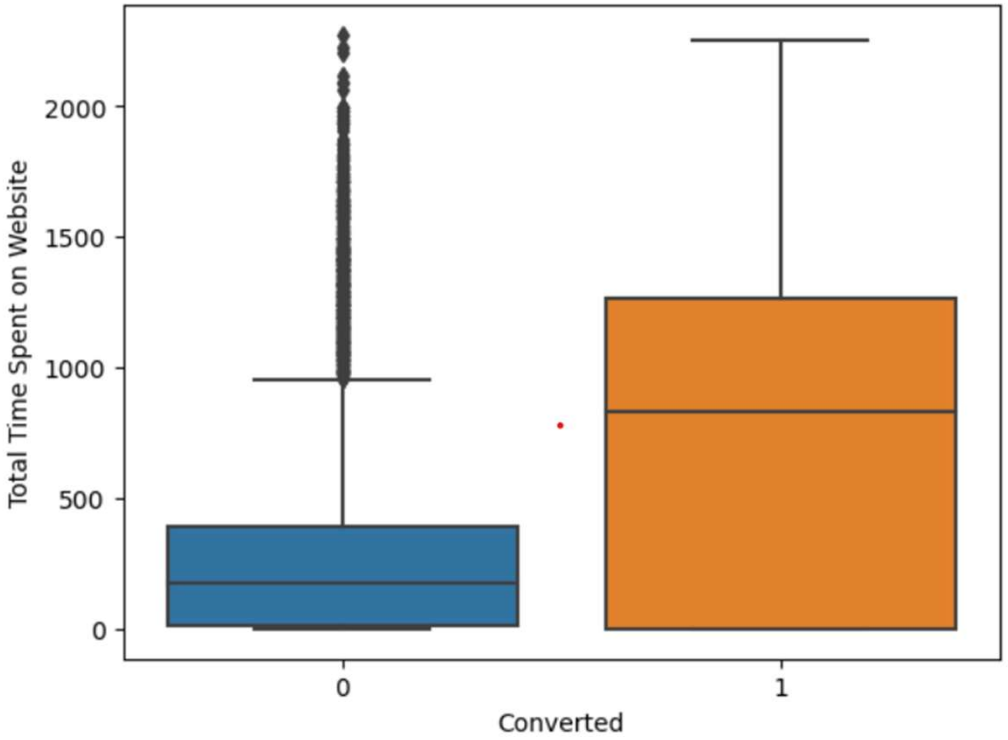
Unemployed and Working Professionals are major leads for us



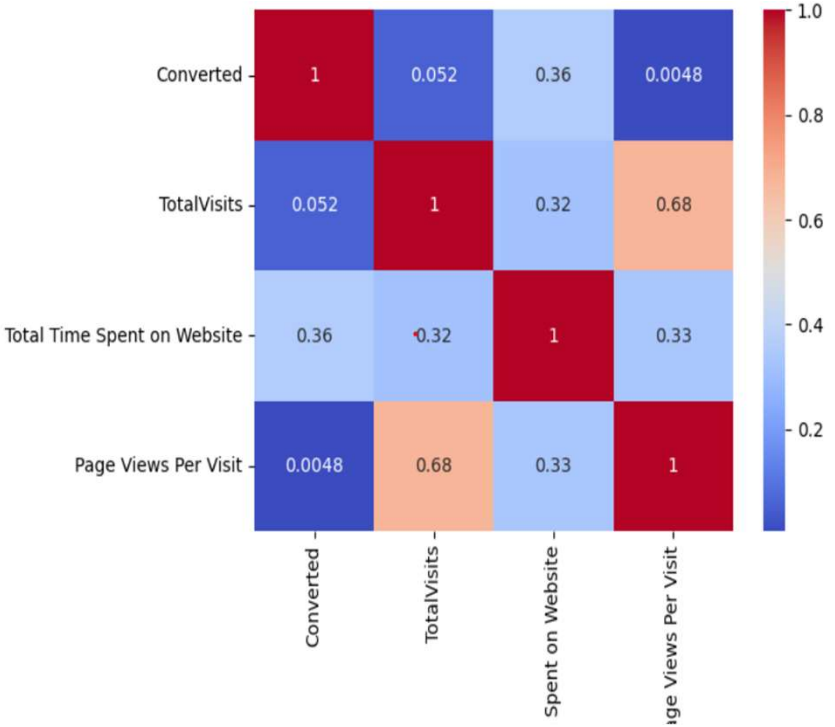
Leads are more from Management posts with Finance Management has highest



People spending average of 750-770 are more likely to convert leads



Most of our numeric values are not correlated to each other

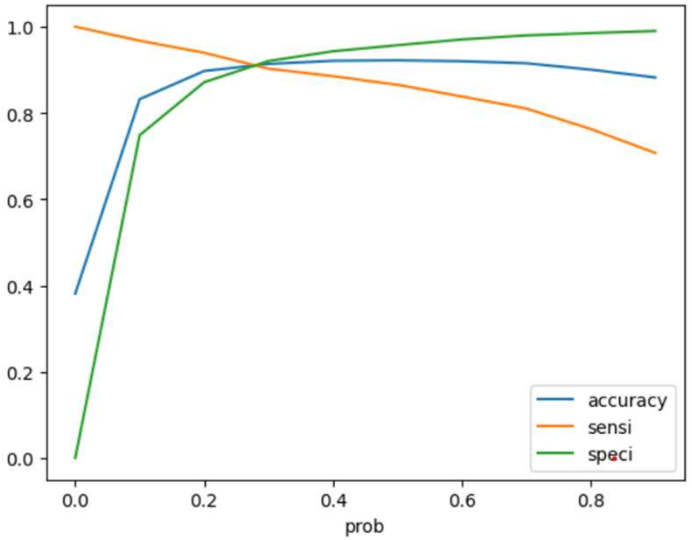


The Variables influencing our model

Tags_Closed by Horizzon
Tags_Lost to EINS
Lead Source_welingak website
Tags_Will revert after reading the email
Last Activity_SMS Sent
Lead Origin_Lead Add Form
Lead Source_olark chat
Last Activity_Others
Total Time Spent on Website
Do Not Email_No
Tags_Busy
What is your current occupation_Missing
Last Notable Activity_Olark Chat Conversation
Last Notable Activity_Modified
const
Tags_Ringing
Tags_switched off

Sensitivity and Specificity on Train Data Set

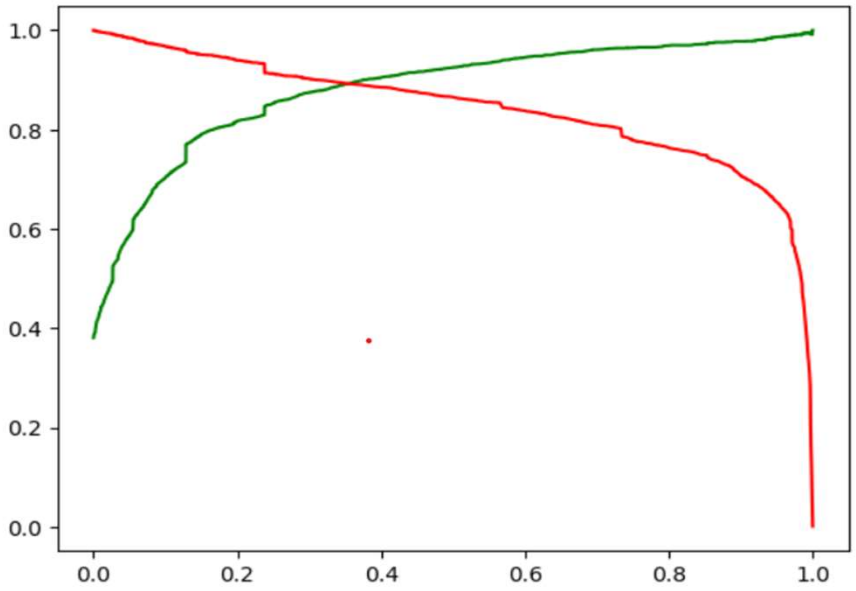
From graph optimal cut off of 0.34 based on Accuracy, Sensitivity and Specificity



- Accuracy - 92%
- Sensitivity - 9 0 %
- Specificity - 82 %
- False Positive Rate – 7.6%
- Positive Predictive Value – 87%

Model Evaluation- Precision and Recall on Train Dataset

From graph an optimal cut off of 0.34 based on Precision and Recall



- Precision for converted - 89 %
- Recall for converted- 89 %

Model Evaluation – Sensitivity and Specificity on Test Dataset

- Accuracy - 93 %
- Sensitivity - 9 2 %
- Specificity - 93 %

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 93%, 92% and 93% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 93%
- Hence overall this model seems to be good