Lead Score Case Study

Abhay Singh
Gaurav Chugh
Vinit Pahwa

Problem Statement

An education company called X Education sells online courses to industry professionals. On any given day, many professionals who are interested in courses land on their website and search for courses.

The company sells its courses on several websites and search engines such as Google. Once these people get to the website, they can browse the courses or fill out a course form or watch some videos. When these people fill out a form with their email address or phone number, they are classified as potential customers. In addition, the company also gets leads through past referrals. Once these leads are obtained, the sales team members begin calling, writing emails, etc. Through this process, some leads will convert, while most will not. A typical lead conversion rate in X Education is around 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very low. For example, if they get 100 leads per day, only about 30 of them will convert. To make this process more efficient, the company wants to identify the most potential leads, also known as "hot leads." If they successfully identify this set of leads, the lead conversion rate should increase because the sales team will now focus more on communicating with potential leads rather than calling everyone. A typical lead conversion process can be represented using the following path:

Lead Conversion Process - Shown as a Funnel Lead Conversion Process - Shown as a Funnel As you can see, there are many leads generated in the initial stage (above), but only a few emerge as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educate the potential customers about the product, communicate constantly, etc.) to get higher lead conversion.

X Education has hired you to help them select the most promising leads, the leads most likely to convert into paying customers. The company requires you to create a model where you assign a lead score to each lead such that customers with a higher lead score have a higher chance of converting and customers with a lower lead score have a lower chance of converting. In particular, the CEO stated that the target lead conversion rate would be around 80%.

Business Objective

- X education wants to know the most promising potential customers.
- To do this, they want to create a model that identifies hot contacts.
- Deploying the model for future use.

Solution Approach

- Data cleaning and manipulation
  1. Check and process duplicate data.
  2. Check and process NA values and missing values.
  3. Drop columns if they contain a large number of missing values and are not useful for analysis.
  4. If necessary, imputation of values.
  5. Check and handle outliers in the data.

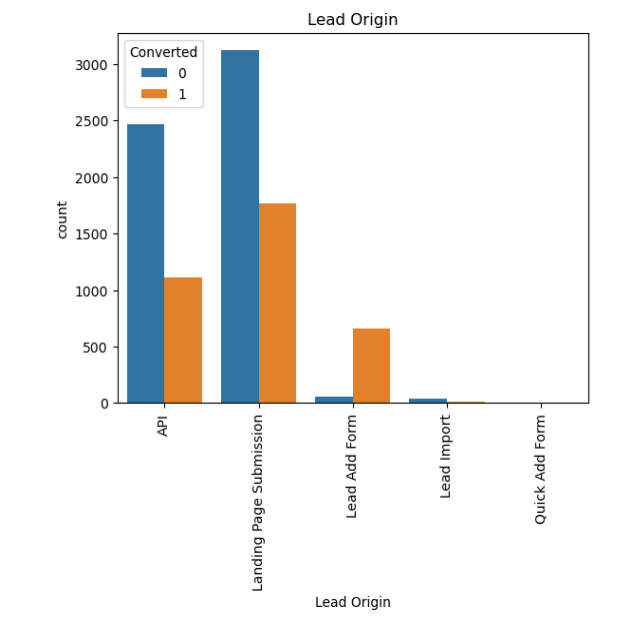
- EDA
  1. Univariate data analysis: number of values, distribution of the variable, etc.
  2. Bivariate data analysis: correlation coefficients and patterns between variables, etc.
- Classification technique: logistic regression used for model building and prediction.
- Model Validation
- Presentation of the model.
- Conclusions and recommendations.

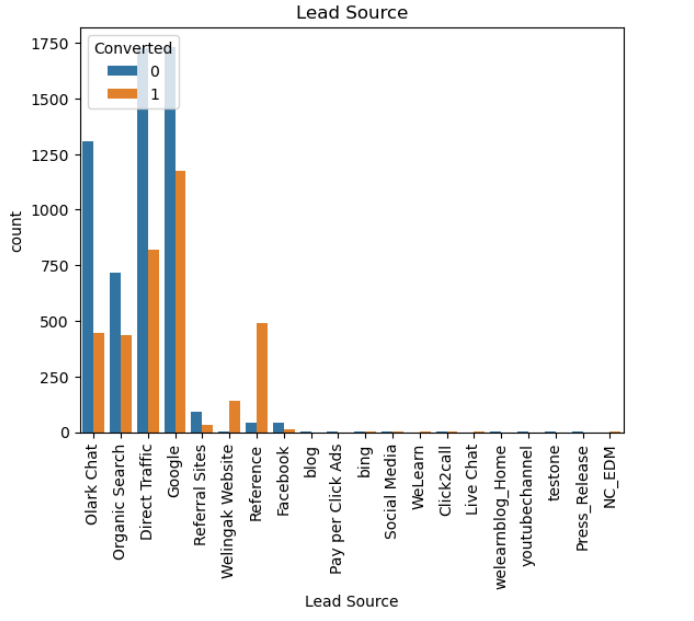
Data Manipulation

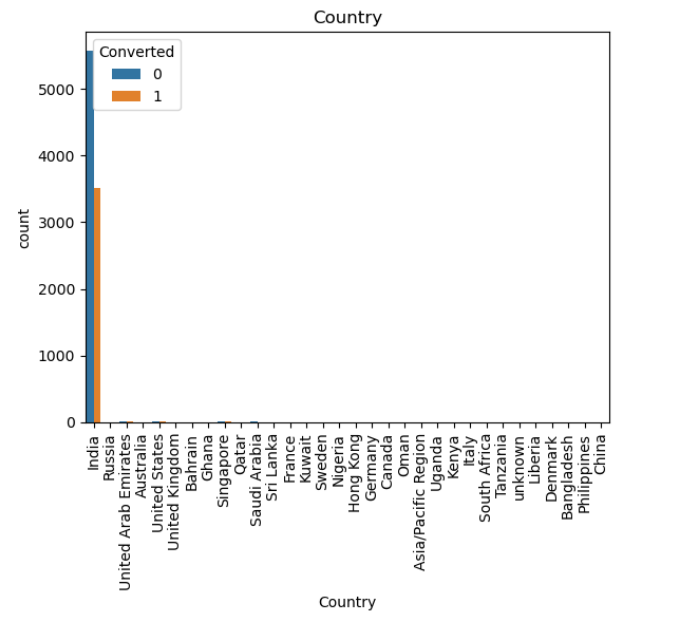
- The shape denotes 37 columns, and the total number of records is 9240.
- There is a huge value of null variables in some columns as seen above. But removing the rows with the null value will cost us a lot of data and they are important columns. So, instead, we are going to replace the NULL values with 'NA’.
- For columns 'Specialization', 'course_selection_reason’, and 'occupation', since all genuine data are well distributed across the records, let's impute the missing value proportionately
- For Columns 'City' and 'Country' the majority of data is of 'Mumbai' and 'India’ respectively, so let’s use mode() method to impute the data.
- Prospect ID and Lead Number are both unique identifiers. which don't server any purpose in data analysis, hence we can drop these columns.
- There few columns like 'Specialization', 'source' that contains values as 'Select', looks like its drop down to collect data and not mandatory fields, so we can convert these 'Select' words to NULL.
- There are few columns with only one unique values (Magazine , courses_updates , supply_chain_content_updates,dm_content_updates and cheque_payment).
- These are binary data columns, having only 'Yes' and 'No' as values , these columns show DATA IMBALANCE, as almost all records have the same value.Because of heavy data imbalance, we can drop the following columns as well

EDA

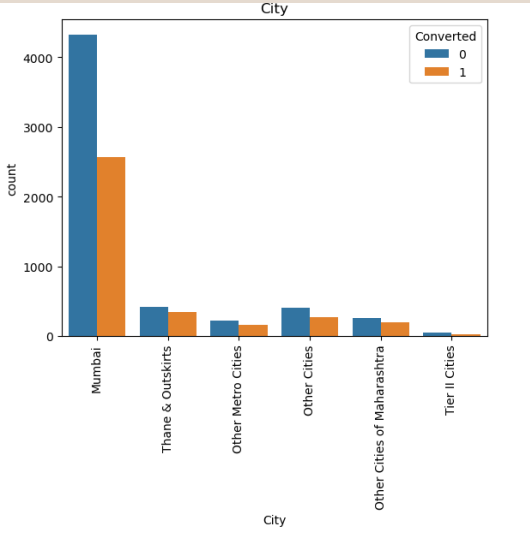
Uni-variate Analysis

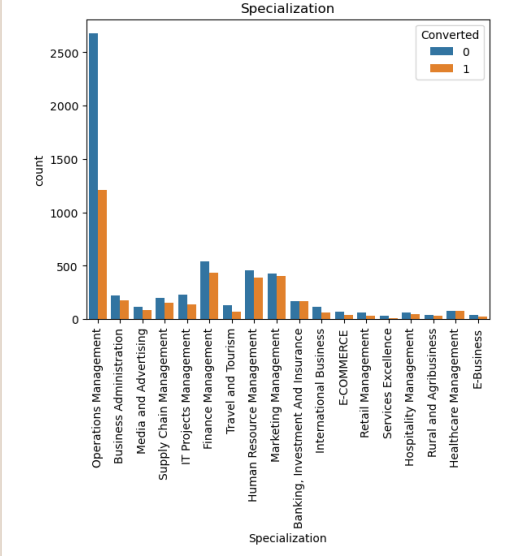


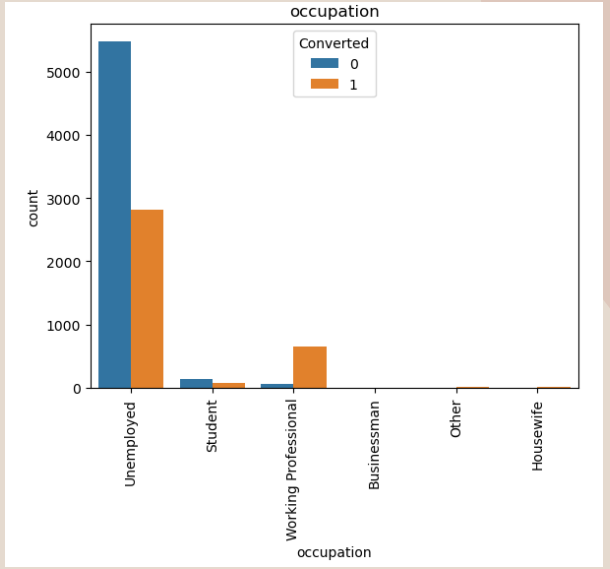


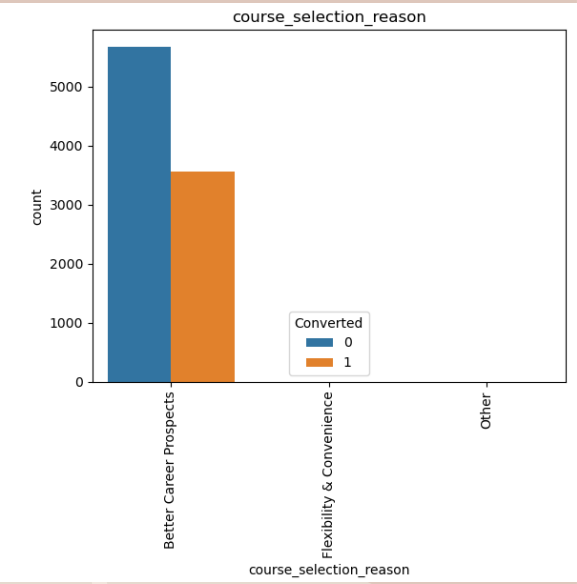


EDA





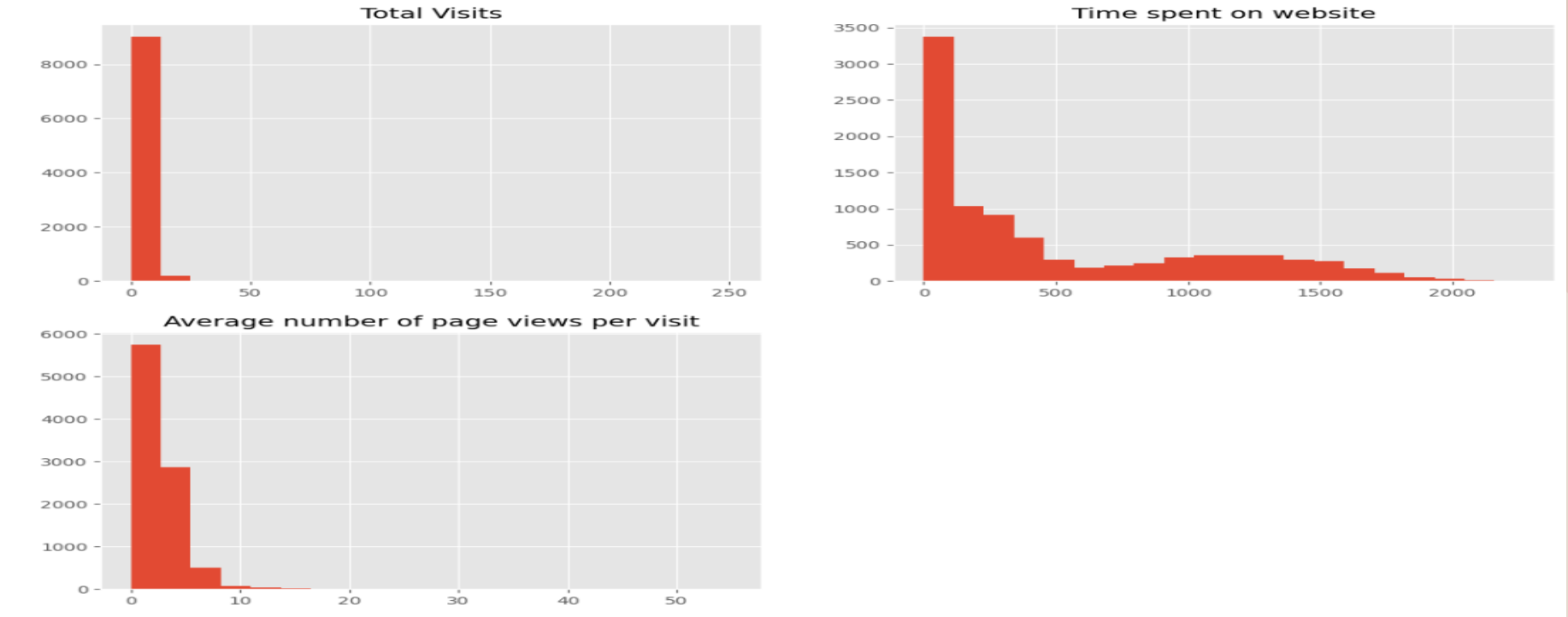




Observation:

- 1) Landing page submission is a good initial origin, it has more chance to convert than others
- 2) Online search engines are a good source of leads, they have 90% of the traffic
- 3) The data mostly belong to India, and mostly from Mumbai
- 4) The "Management" category is the most popular, with more than 70% of potential customers
- 5) The "Unemployed" section contains most of the questions
- ~~6) People go to the "Better choice of carrier" course.~~

Numrical data Analysis



Observation:
1) data have high peaks.
2) data looks skewed.

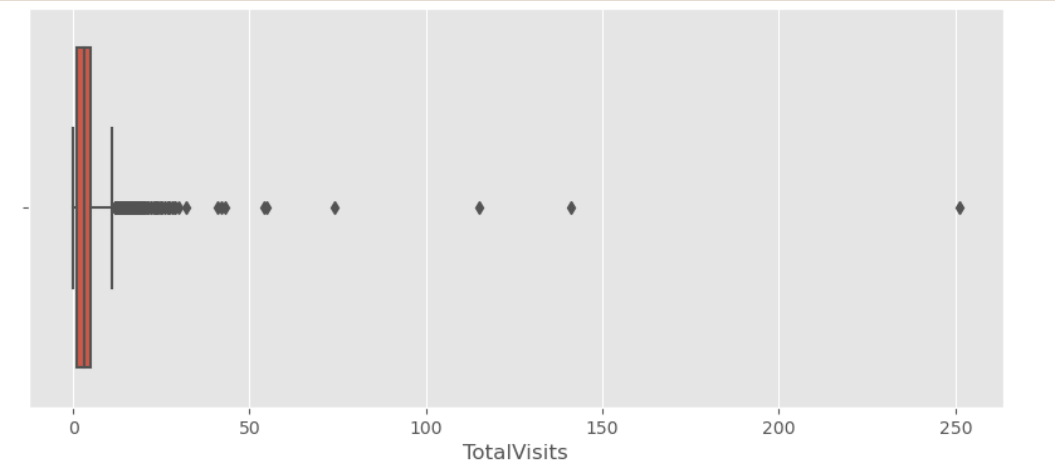
Numrical data Analysis

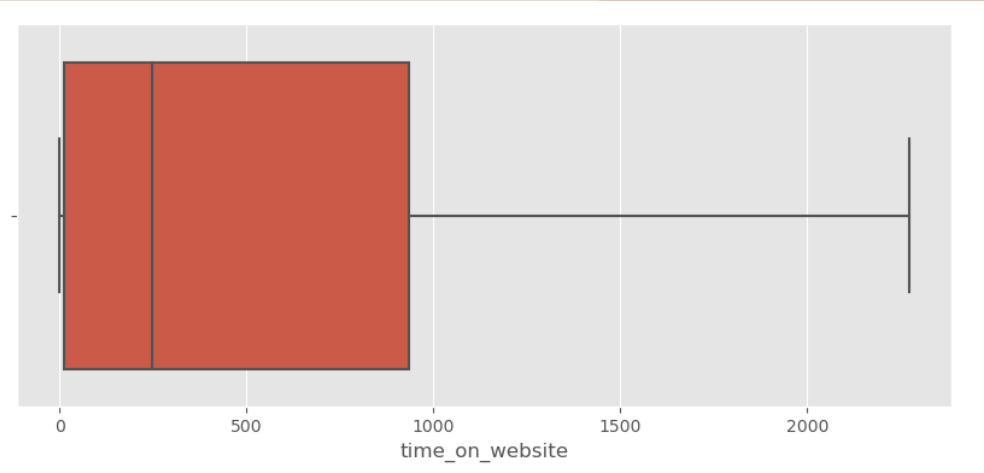
Observations:

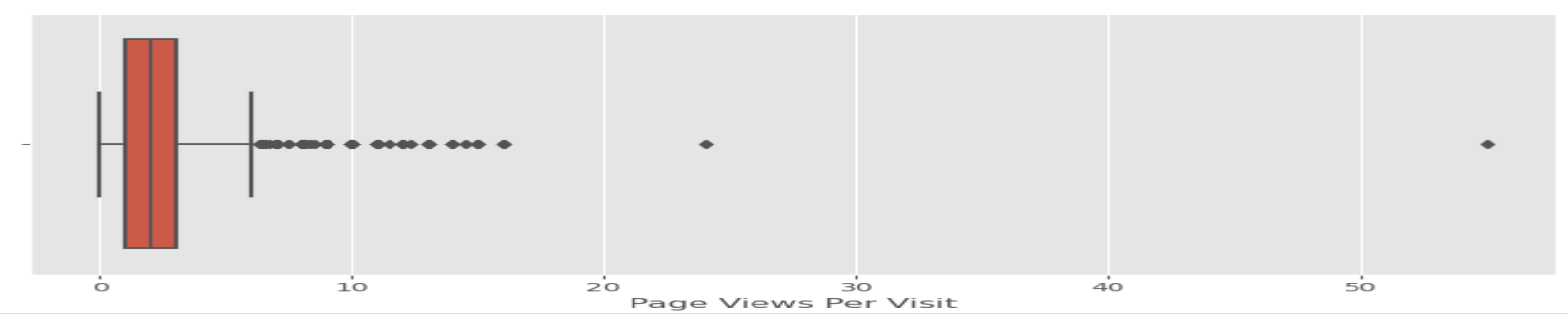
1) Time spent on website has strong and positive relation with target variable 'Converted'

2) Page views per visit has week and negative relation with target variable 'Converted'

Check Outliers







Observations:

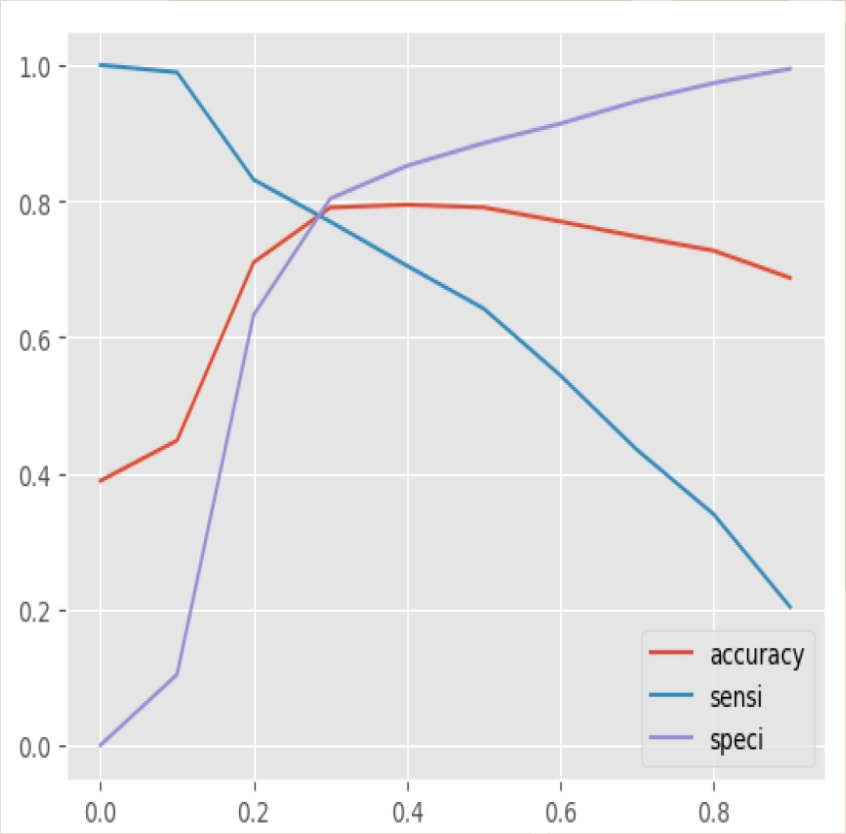
1) Look at 1st and 3rd box plots and the statistics, there are upper bound outliers in both total_visits and page_views_per_visit columns.
2) We can also see that the data can be capped at 99 percentile.
3) We are not going to treat Outliers as of yet.

Model Building

- Split the dataset into 70% and 30% for train and test respectively.
- Since number of feature is more let go with RFE method first and then we will use manual method to further fine tune the model
- Running RFE with 15 variables as output.
- Building Model by removing the variable whose p-value is greater than 0.05 and vifvalue is greater than 5
- Predictions on test data set
- Accuracy of this model is 79.25%.

Optimize Cut off (ROC Curve)

- The ROC curve demonstrates several things:
- It shows a trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).
- The closer the curve follows the left boundary and then the upper boundary of the ROC space, the more accurate the test.
- The closer the curve is to the 45 degree diagonal of the ROC space, the less accurate the test.
- The area under the ROC curve is 0.83.
- ROC Curve gave optimal cut-off value as 0.3 With the cut-off value as 0.3 the model parameters are Accuracy= 79.07 % sensitivity= 64.2 % specificity= 88.52 %

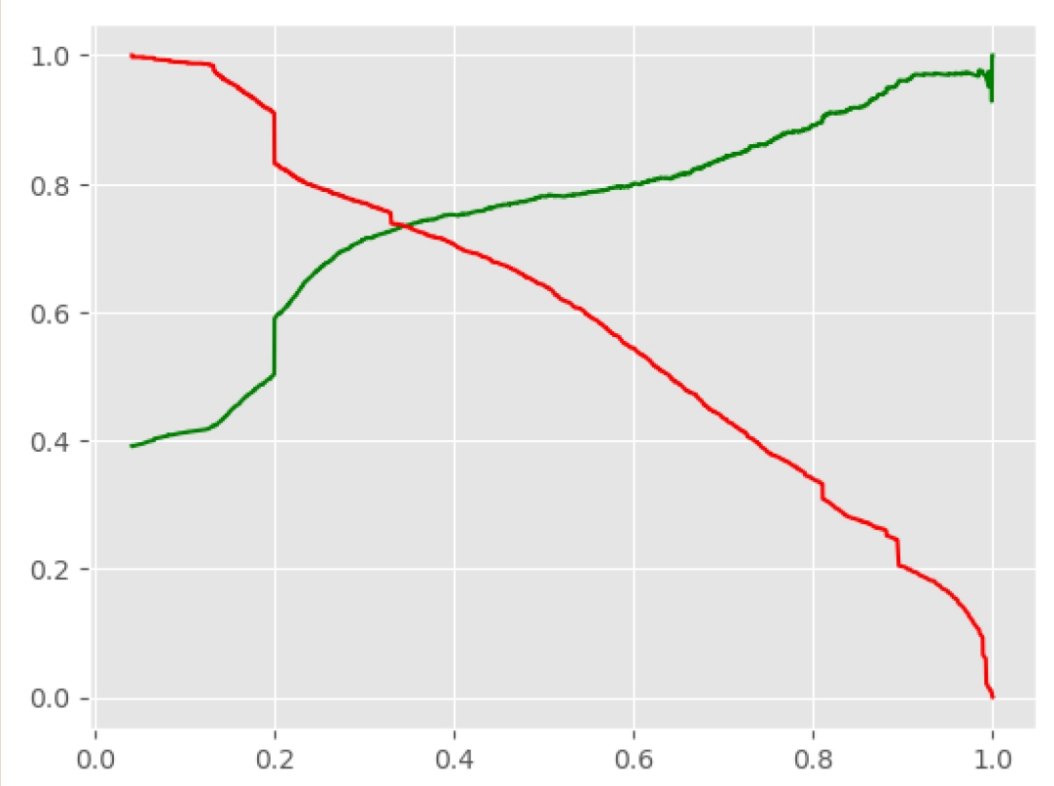


Prediction on Test set

- Model parameters on the test set is as followed Accuracy= 78.93 % sensitivity= 79.85 % specificity= 78.38 %

Precision-Recall

- The Precision-Recall curve shows that the cutoff is 0.35 With a cutoff of 0.35, the model parameters for the train data set are Precision= 79.07% Sensitivity= 64.2% Specificity= 88.52%
- With the same cutoff, the value of the model parameters for the test data set is Accuracy= 79.11% Sensitivity= 74.88% Specificity= 81.68% Overall:
- With the current limit of 0.35 we have Accuracy=79.22%, Sensitivity=74.88%, and Specificity around 81.68% for the test Data set.
- Observation: There was a slight improvement in the accuracy and specificity values, while a significant decrease in the sensitivity value with this new threshold of 0.35 on the test data set



Conclusion

The parameters/features were found to matter that help predict course sales

- Do Not Email
- TotalVisits
- time_on_website
- Number of Pageviews per Visit
- Origin_Landing Page Submission
- Lead Source_Olark Chat
- Lead Source_Reference
- Lead Source_Welingak Website

Specialization_Operations Management Occupation_Working Professional

Business Interpretation

a) The sales team should focus on leads that are more active on X Education’s website, now they can take a different parameters to collect these leads, these parameters could be more number of visits, total time spent on the website or they browse many pages every time they visit the website. As these leads have a better chance of buying the courses.

b) The sales team should target the leads that have come from the landing page of the website.

c) The sales team should focus on working professionals and not on students and the unemployed as the working professional are looking for better career opportunities and to full fill that they are opting for courses.

d) The sales team should focus on leads that have come from references as they have a high probability of buying the courses.

e) The sales team should keep Unemployed Students on their secondary list

f) They should not focus on leads that have come from search websites like Google and all, as they are just browsing as of now.

thank you

Abhay Singh
Gaurav Chugh
Vinit Pahwa