LEAD SCORE CASE STUDY

WAQAS YAQOOB

PROBLEM STATEMENT

- 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 30%.

Now, although X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as 'Hot Leads'. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

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%.

GOAL OF BUSINESS

- There are quite a few goals for this case study:

1. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.
2. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

SOLUTION TO PROBLEM

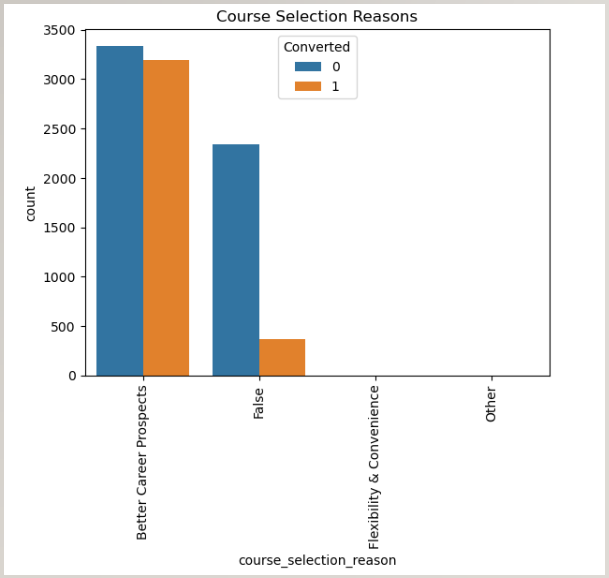
I have followed the following steps to solve this data problem

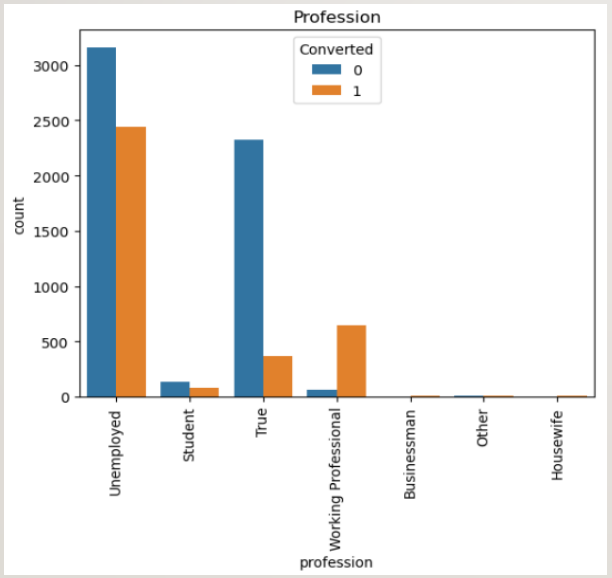
1. First of all cleaning the data
2. Checking the duplicating values and missing values.
3. Imputing the values in dataframe if necessary and also handling the outliers
4. Performing the Univariate analysis for categorical and numeric variables
5. Performing the Multivariate analysis.
6. Using logistic regression for model building and prediction.
7. Presentation of model

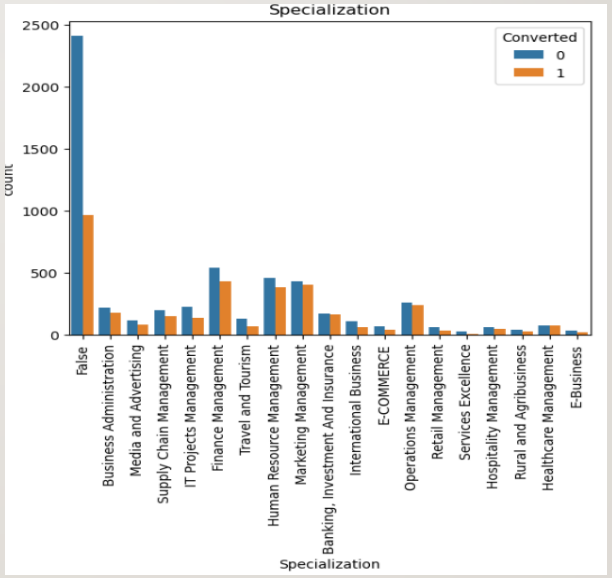
DATA ANALYSIS

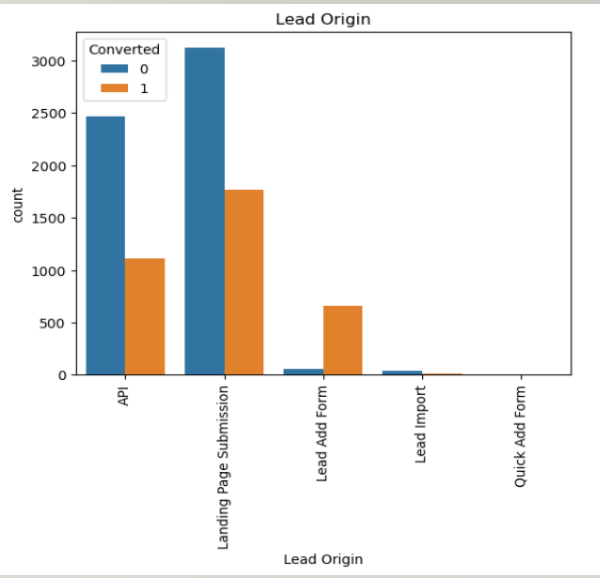
- The shape denotes 37 columns, and the total number of records is 9240.
- There is a huge value of null variables in some columns as seen above. But removing the rows with the null value will cost us a lot of data and they are important columns. So, instead, we are going to replace the NULL values with '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 almostall records have the same value.Because of heavy data imbalance, we can drop the following columns as well

UNIVARIATE ANALYSIS

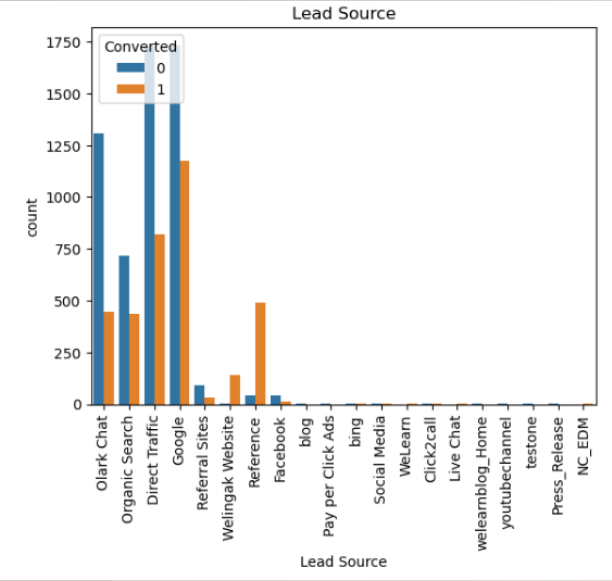


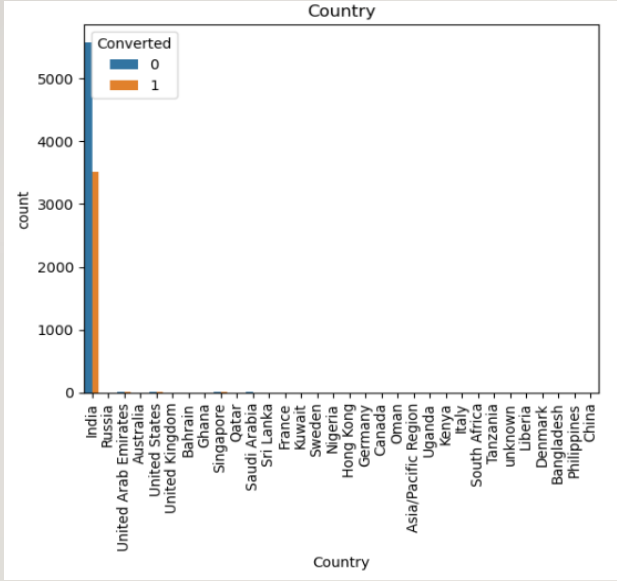


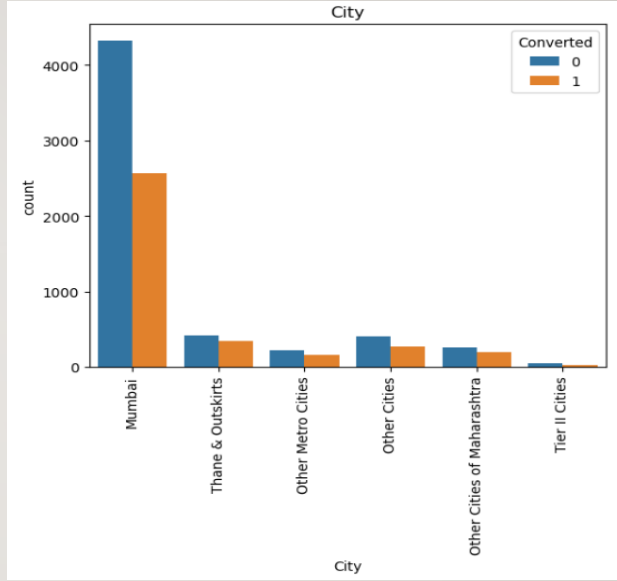




UNIVARIATE ANALYSIS



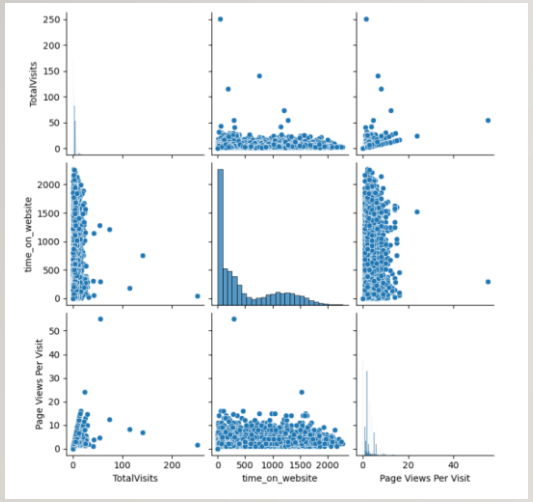


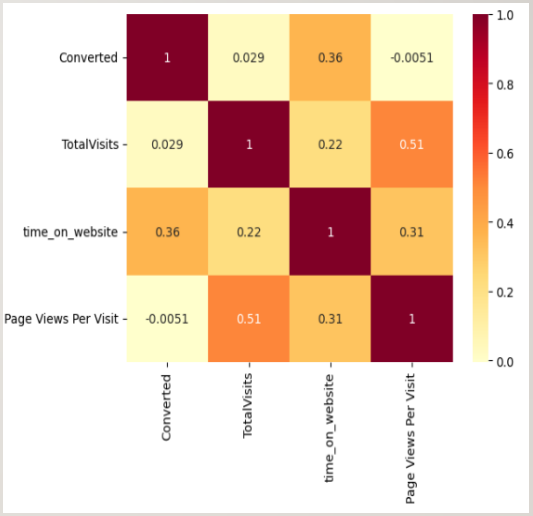


Observation:

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.

MULTIVARIATE ANALYSIS



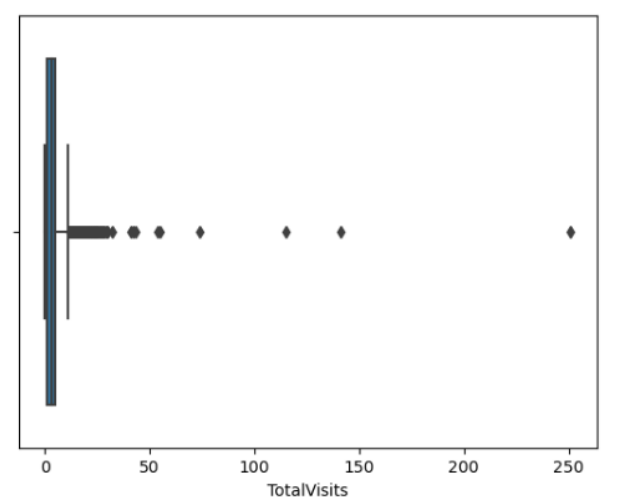


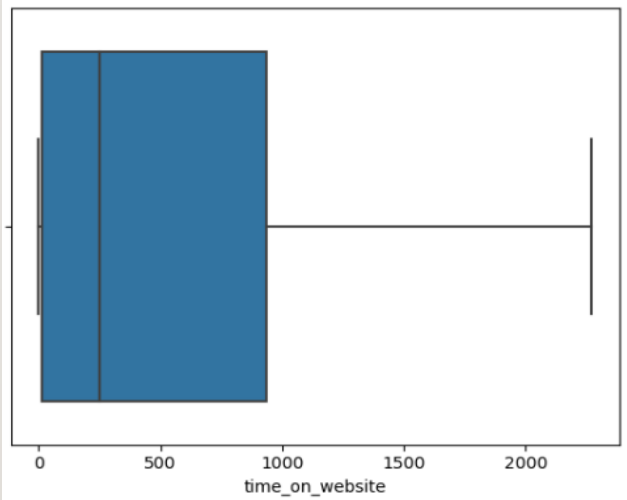
Observations:

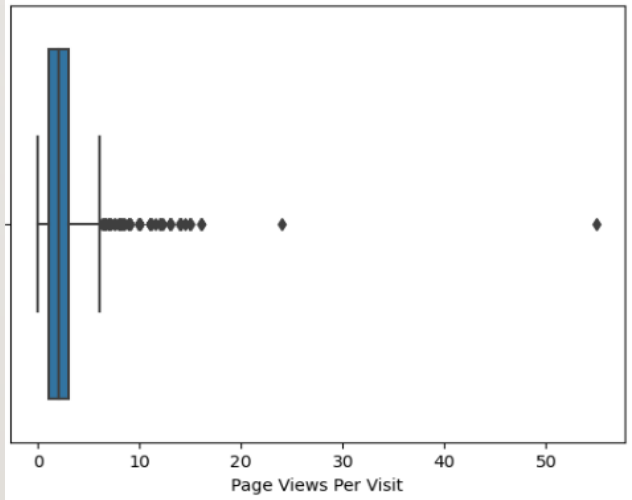
1) Time spent on website has strong and positive relation with target variable 'Converted'

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

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

- 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%.

ROC CURVE

The ROC curve demonstrates several things:

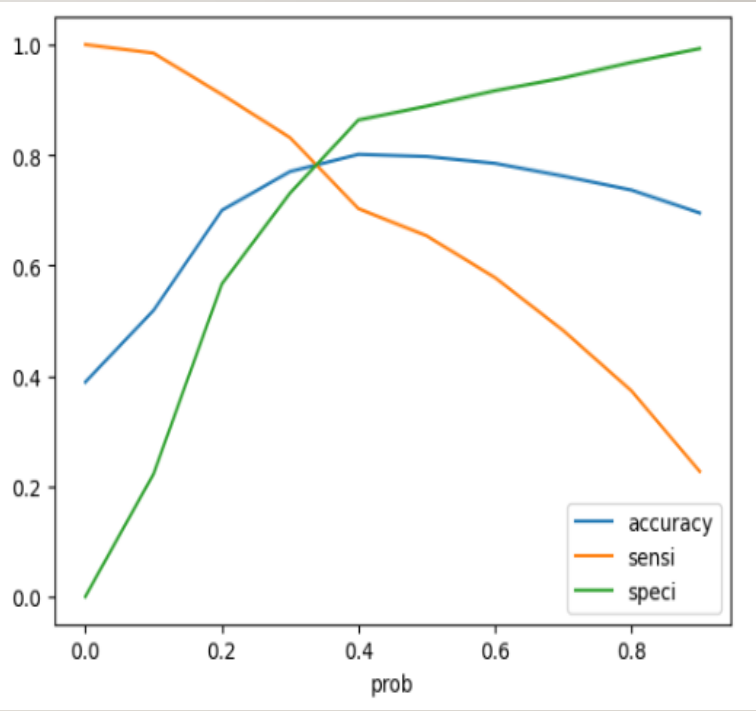
It shows a trade-off between sensitivity and specificity (any increase in sensitivity will be accompanied by a decrease in specificity).

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 %

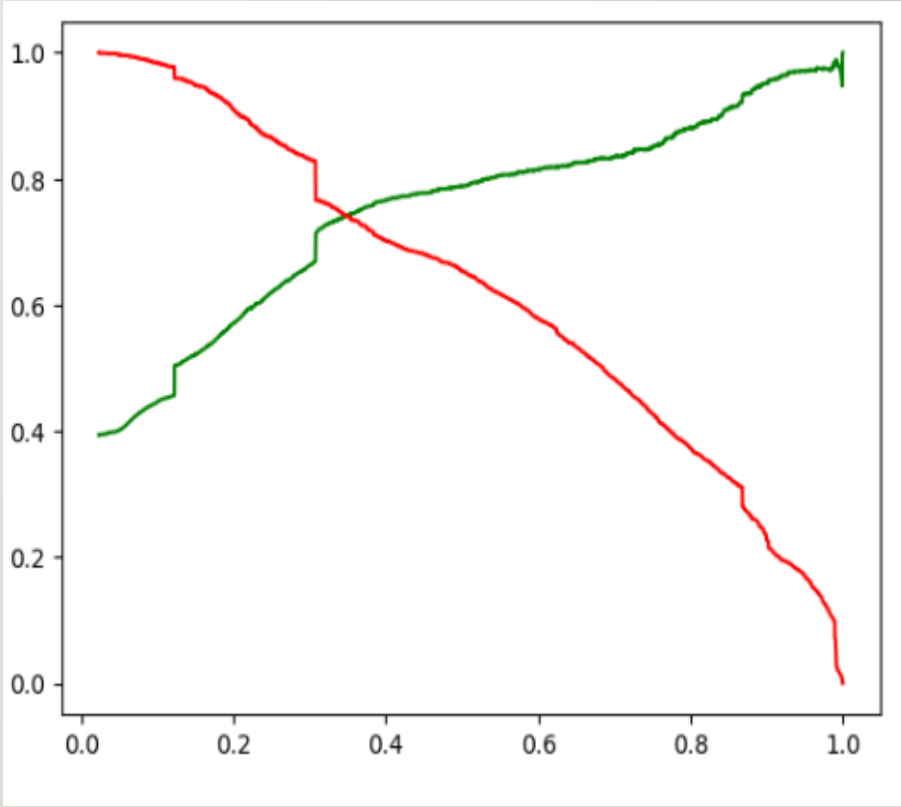


PREDICTION ON TEST SET

- 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



SUMMARY

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