LEAD SCORE CASE STUDY

PREDICT HOT LEADS FOR X EDUCATION COMPANY

PROBLEM STATEMENT

X Education company markets its courses on several websites and search engines like Google. Once 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%.

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.

Target Lead conversion rate to be around 80%.

DATA SOURCING

Leads.csv

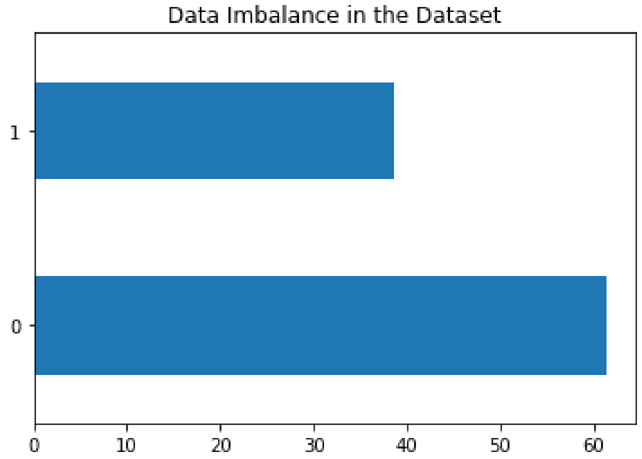
- Leads Dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. which may or may not be useful in ultimately deciding whether a lead will be converted or not. The target variable, in this case, is the column ‘Converted’ which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn’t converted.

Leads Data Dictionary.xlsx

- Learn more about the dataset from the data dictionary provided. This contains the definitions of all columns in the Leads Dataset

DATA CLEANING AND ANALYSIS

Data Imbalance

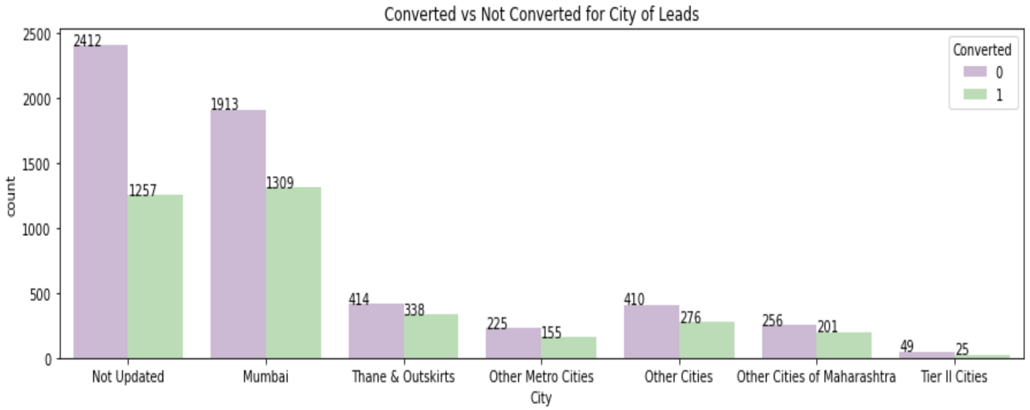


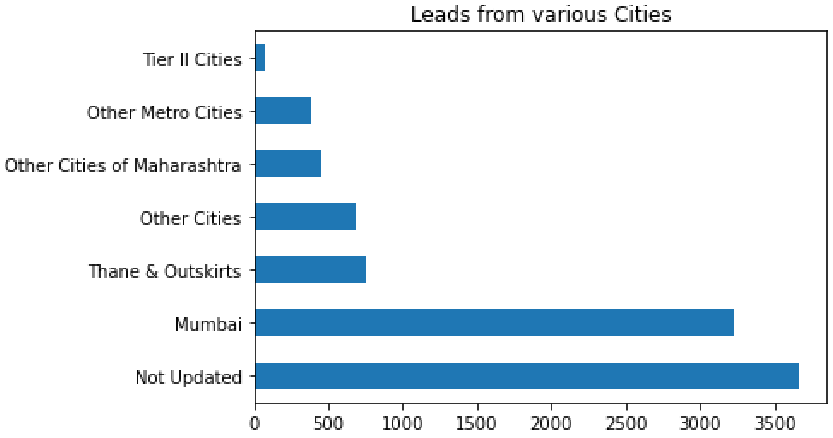
As we can see there is no data imbalance present in the given dataset.

The conversion vs non conversion ratio is ~3:2

Hence the given dataset is good for model building.

As we are going to predict if a lead is going to be converted to Hot Lead or not i.e. as our target variable is binary we are going to use Logistic Regression model for the same.



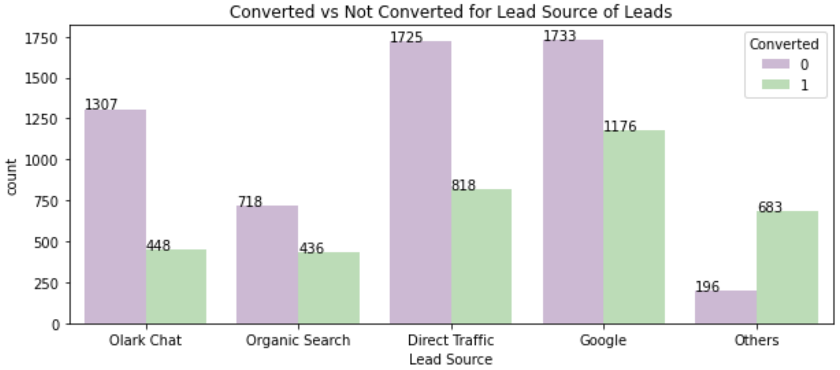


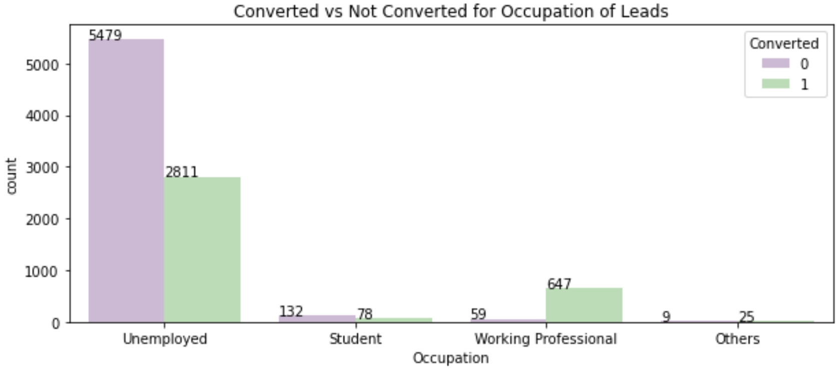
Customer Base for X Education

From the given data we understood that the major customer base for the company is India and Maharashtra.

Majority of the leads who filled the city information are from Mumbai.

Also we can see for the customers across all the cities the conversion rate is between 30% - 40%.


Converted vs Not Converted for Lead Source of Leads


Converted vs Not Converted for Occupation of Leads

Conversion Rate

Google has become the main marketing platform as majority of the leads source is Google.

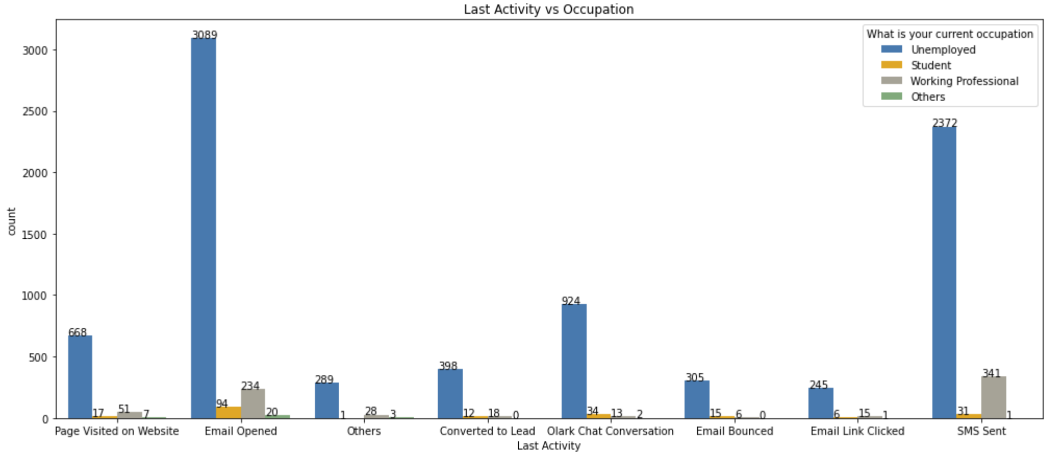
Also ~40% of Leads sourced from Google are converted to Hot Leads.

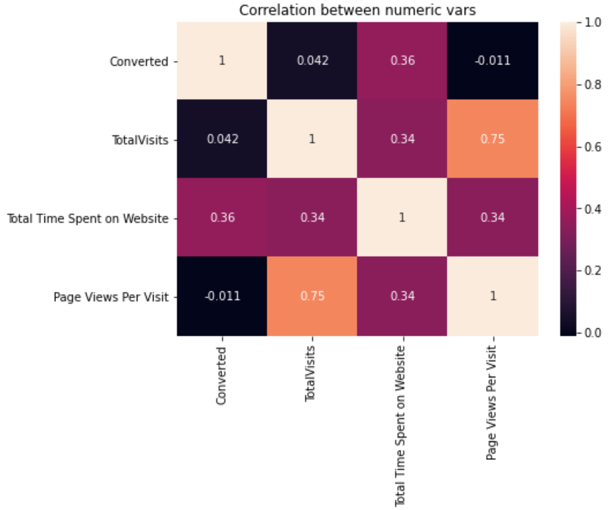
We have clubbed the categories which sourced less than 2% of leads into Others. Here the conversion rate is ~77%.

Also 89% of the leads are Unemployed.

From Unemployed and Student Category, the convertion is ~34% and ~37% respectively.

(Bi)Multivariate Analysis





When we compared the Last Activity of the Lead with their Occupation we found that majority of the under different categories of Last Activity are from Unemployed Category.

Also when we looked at the correlation between numerical variables we observed that TotalVisits and Page Views Per Visit columns have high correlation of 0.75

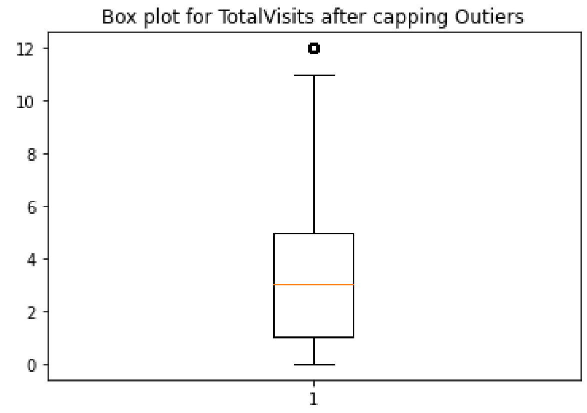
Treatment of Missing Values and Outliers

In the process of Data Cleaning and Analysis we have imputed many Null values in the dataset.

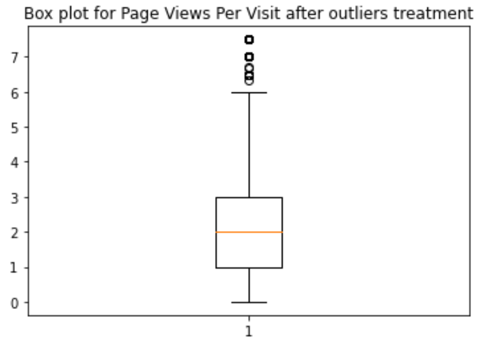
There are few columns with values as “Select”. These values are considered as null values as they are the default values in the dropdown when the lead does not select them from UI.

We also capped the upper outliers for numerical columns like TotalVisits, Page Views Per Visit.

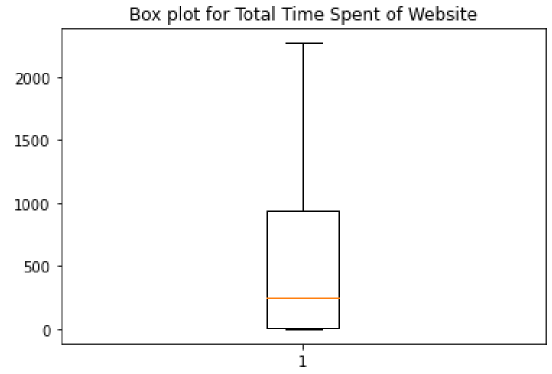
More details on the outliers and missing values treatment can be found in the Jupyter Notebook provided with this presentation.



Box plot for TotalVisits after capping Outiers



Box plot for Page Views Per Visit after outliers treatment



Box plot for Total Time Spent of Website

Model Building and Evaluation

Prepare Data for Modelling

We have created dummy variables for all categorical columns and then dropped the categorical columns.

We had split the data into train and test split with 70% and 30% of the leads data respectively.

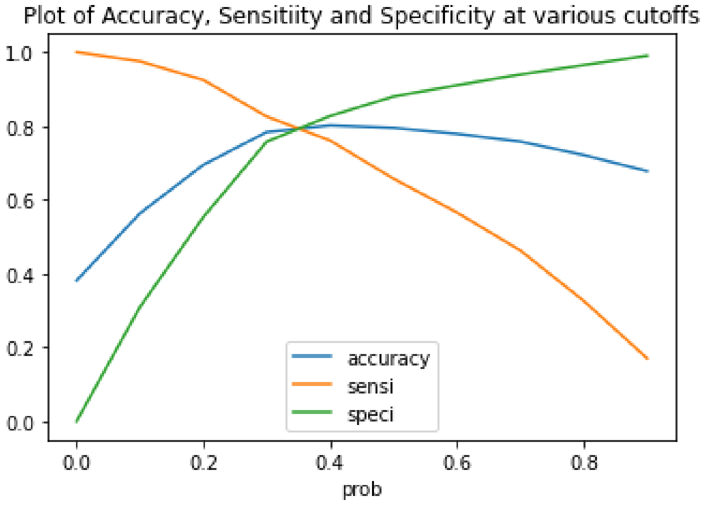
Then we used Standard Scaler for scaling the numeric variables like TotalVisits, Page Views per visit, Total Time Spent on Website.

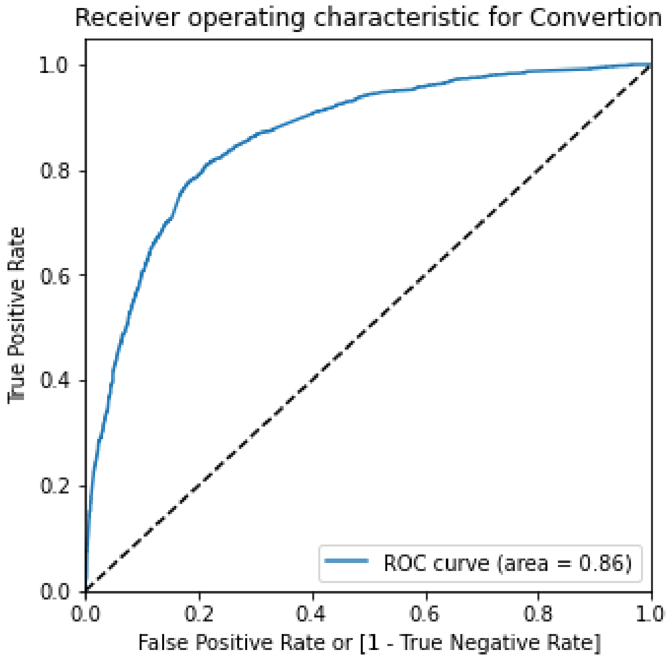
We built the initial model by selecting 25 features with RFE(Recursive Feature Elimination) technique.

Then we used the manual approach to eliminate features one by one which are having high p-value and high VIF values.

During this process we calculated the accuracy and sensitivity at a cut off probability of 0.5. This is to observe the change in those metrics as we proceed to arrive at final model.

We arrived at the final model with 15 features.





Logistic Regression Model

The final model we achieved has 15 features.

We were able to achieve 80% sensitivity with a cut off value of 0.34 for the Conversion Predictability.

As we can see in the graphs the accuracy, sensitivity and specificity curves meet at ~0.34.

Also the ROC curve shows a good tradeoff between sensitivity and specificity which means the predictive power of true positives is good in out model

Prediction

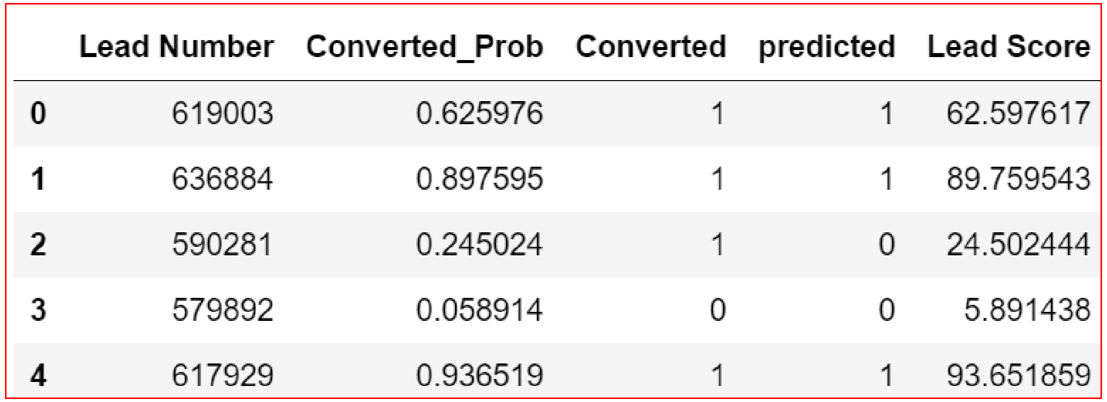
Predict Conversion in Test Data

We ran our final model on test data to predict conversion.

We achieved the accuracy and sensitivity of the model on test data is 79% and 80% respectively.

With 0.34 as a cut off we predicted the conversion of every lead.

We have assigned Lead Score corresponding to the Lead Number for each lead. A lead with high lead score implies 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.



	Lead Number	Converted_Prob	Converted	predicted	Lead Score
0	619003	0.625976	1	1	62.597617
1	636884	0.897595	1	1	89.759543
2	590281	0.245024	1	0	24.502444
3	579892	0.058914	0	0	5.891438
4	617929	0.936519	1	1	93.651859

Final Result

We decided to choose the cut of for the predicted probability as 0.34 for achieving the sensitivity of 80%.

The Sales team can predict the leads based on the given model and target the leads that have lead score greater than 34%.

Below are the features that are part of the final model.

- 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit', 'A free copy of Mastering The Interview', 'LeadOrigin_API', 'LeadOrigin_Lead Add Form', 'LeadSource_Olark Chat', 'LastActivity_Email Bounced', 'LastActivity_Email Link Clicked', 'LastActivity_Email Opened', 'LastActivity_Olark Chat Conversation', 'LastActivity_Others', 'LastActivity_Page Visited on Website', 'Specialization_Not Selected', 'City_Other Metro Cities'.

Thank You

- Team:
  - Srihari K S S.
  - Sandeep Rana
- Program:
  - Data Science March 2020.
- Email Ids:
  - srihariks s@gmail.com
  - sandeep.ece.111090@gmail.com