LEAD SCORE CASE STUDY

Building a Logistic Regression Model

-By Shashank,

Nisith Sahoo,

Dharani

Problem Statement

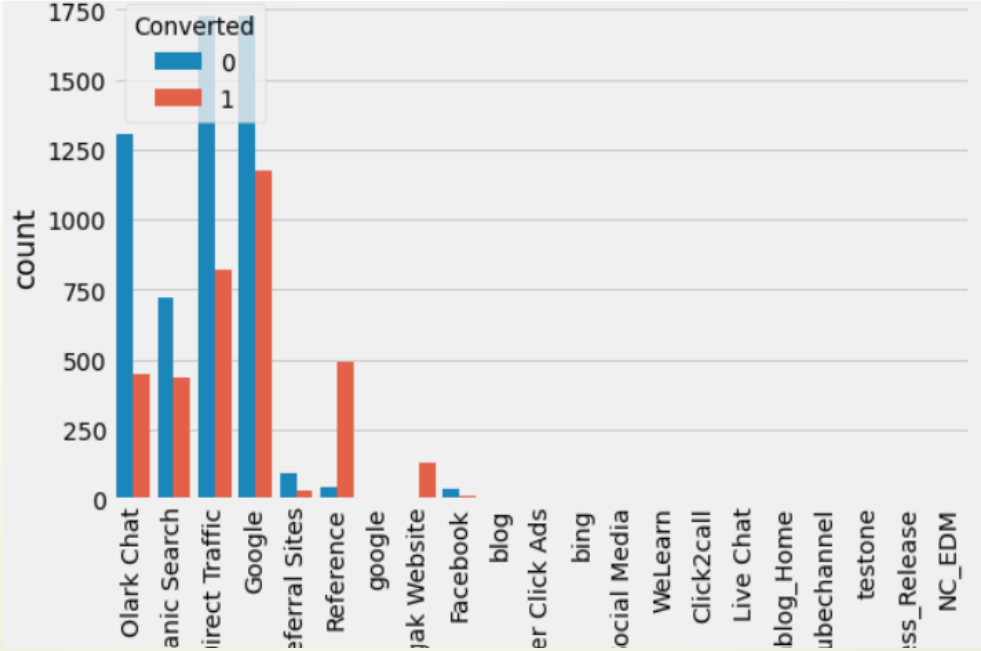
- X Education company is an online course selling company.
- Company markets on platforms like Google, FaceBook, YouTube, etc.
- Lead
  - When a person enters their email Id or phone number via website.
  - When a person is referred by past referrals.
- In order to convert the lead to valuable customer, sales team contacts all the lead, try to explain about the courses they offer.
- The typical lead conversion rate is only 30%.
- Design a model to find the Hot-Leads, such that the conversion rate of Hot Leads is around 80%.

EDA
Data Cleaning

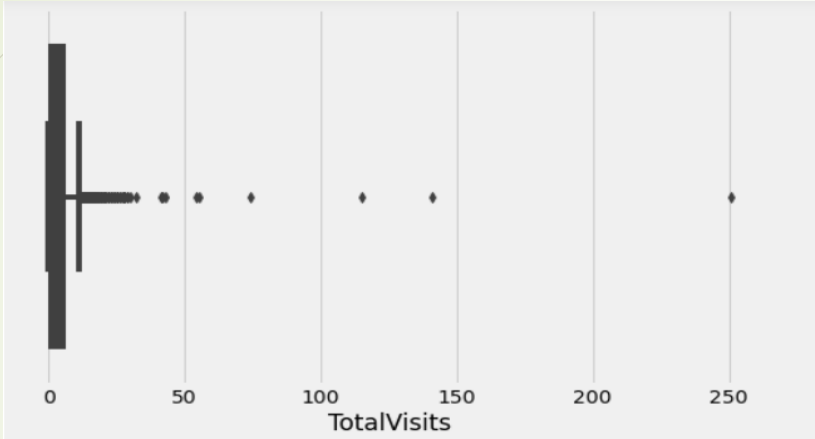
- Few columns has value as “Select” which can be treated as null/missing values as per the input, thus, converted “Select” to null values.
- Removed columns with more than 40% of null/missing values.
- Irrelevant columns such as country and city has been removed.
- If the column has less then 5% null values, replaced with median or mode.
- Columns with 5-40% of null values, handled manually based on the column nature.
- Removing all the redundant and imbalance columns, which doesn’t have any pattern

EDA
Univariate Analysis

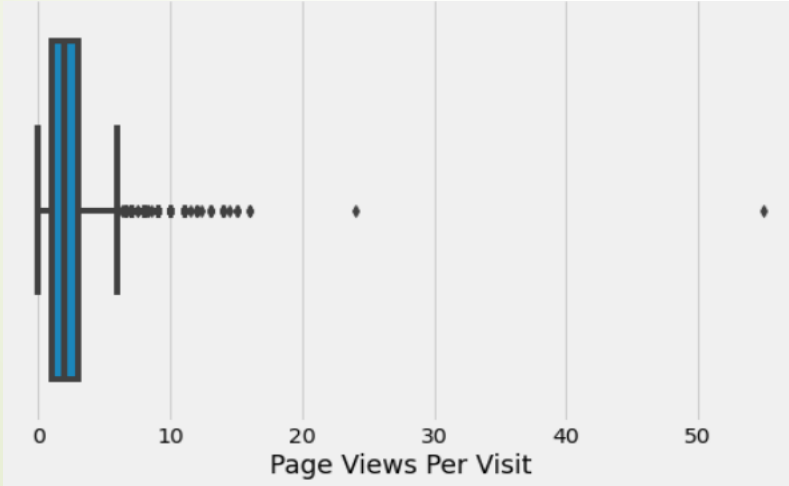
- **Categorical Variables**
  - Last Notable activity and Last Activity are the more or less same. So Last activity is removed.
  - **Lead Source -** Many categories does not carries significant amount of data So replace them all with others



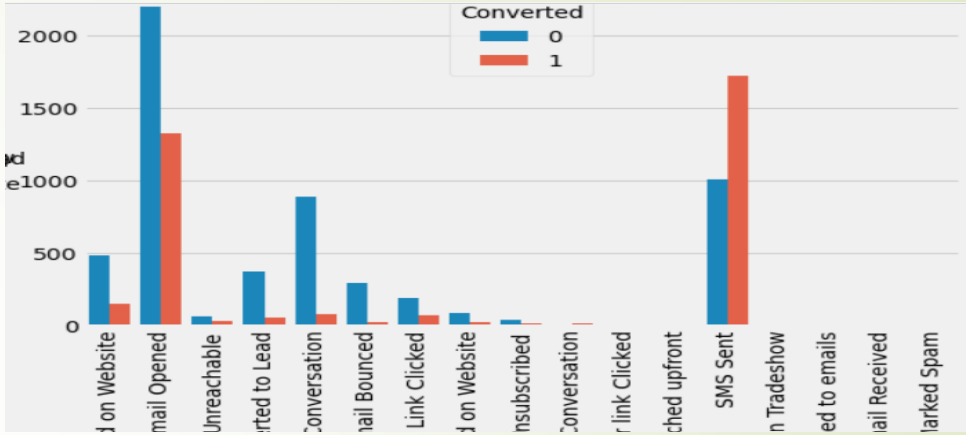
- These outliers for total visits are replaced with 95th percentile value



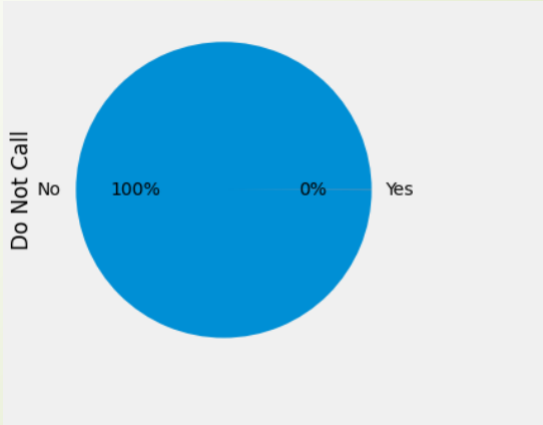
- Page views are replaced with 95th percentile value.



- We Can Club some categories of last activity to "Other Activities" as they having very few data

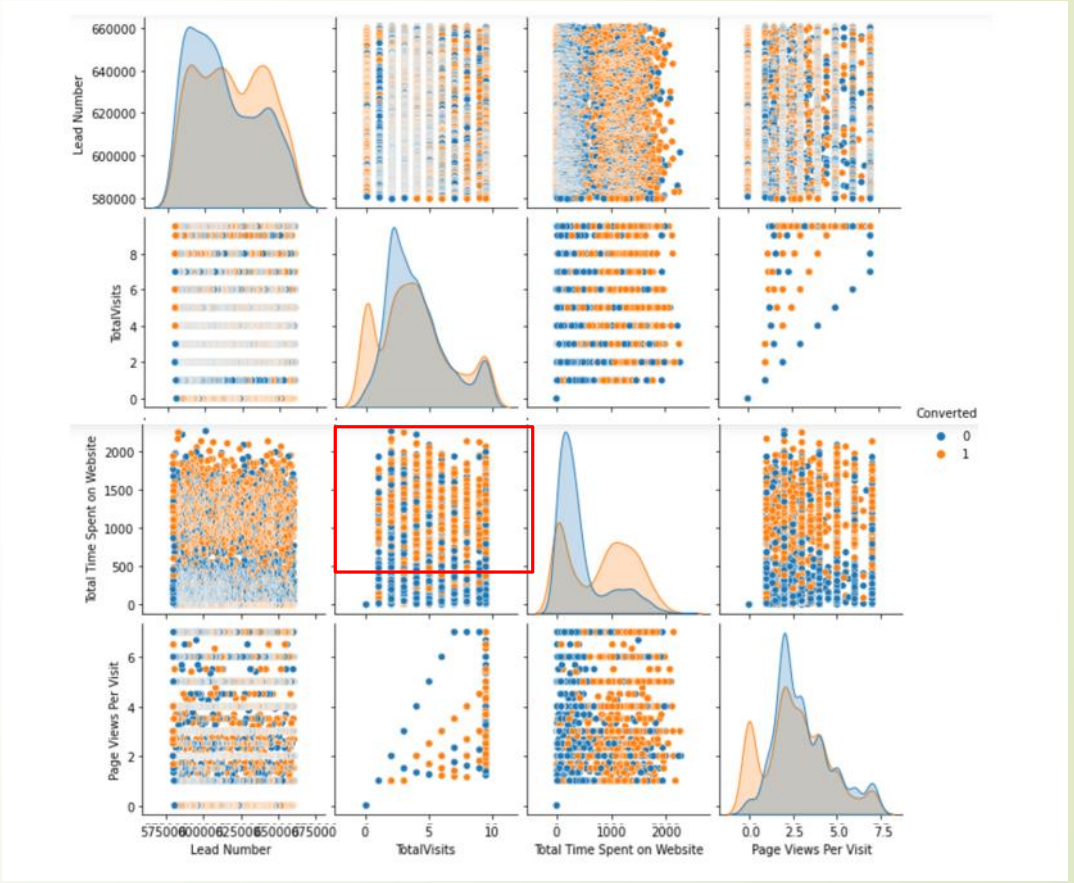


- As the "Do Not Call" column have only one category of data that is "No" so this not so significant for analysis. So the columns with same imbalance are removed.



EDA
Bivariate Analysis

- Most of the customers before taking the course have visited the website and also spent total time >500



Pre-Model Preparation

- Converted Columns with “Yes” and “No ” to 1 and 0 respectively.
- Created Dummy variables for categorical variables with or more than 2 categories
- Standardizing Data using standard scaler. i.e mean is 0 and standard deviation is 1
- Removed Prospect ID and kept lead number for mapping purpose, as these are not necessary for model building.
- Splitting Data to train and test data set to 7:3 ratio.

Model Building
Model 1

- Model was built with all the variables, to get an essence. And concluded that there are lots of insignificant features.
- We have selected top 20 features by using recursive feature elimination method to deal with the above case

Lead_Origin_Lead Import	19.6116	1.77e+04	0.001	0.999	-3.47e+04	3.48e+04
Lead_Origin_Quick Add Form	19.1723	1.77e+04	0.001	0.999	-3.47e+04	3.48e+04
Lead_Source_Facebook	-19.2753	1.77e+04	-0.001	0.999	-3.48e+04	3.47e+04
Specialization_Business Administration	-0.2655	0.226	-1.176	0.240	-0.708	0.177
Specialization_E-Business	-0.1612	0.462	-0.349	0.727	-1.068	0.745
Specialization_E-COMMERCE	0.1168	0.325	0.359	0.720	-0.521	0.755
Specialization_Finance Management	-0.1591	0.196	-0.810	0.418	-0.544	0.226
Specialization_Healthcare Management	-0.0919	0.301	-0.305	0.760	-0.683	0.499
Specialization_Hospitality Management	-0.8564	0.338	-2.531	0.011	-1.520	-0.193
Specialization_Human Resource Management	-0.2127	0.199	-1.071	0.284	-0.602	0.176
Specialization_IT Projects Management	-0.1016	0.234	-0.433	0.665	-0.561	0.358
Specialization_International Business	-0.7033	0.299	-2.350	0.019	-1.290	-0.117
Specialization_Marketing Management	-0.0973	0.199	-0.488	0.626	-0.488	0.294
Specialization_Media and Advertising	-0.3003	0.278	-1.082	0.279	-0.844	0.244
Specialization_Operations Management	-0.1316	0.218	-0.604	0.546	-0.559	0.295
Specialization_Others	-1.6083	0.207	-7.777	0.000	-2.014	-1.203
Specialization_Retail Management	-0.7287	0.381	-1.911	0.056	-1.476	0.019
Specialization_Rural and Agribusiness	-0.1181	0.403	-0.293	0.769	-0.908	0.672
Specialization_Services Excellence	-0.2508	0.516	-0.486	0.627	-1.263	0.761

Model building

Steps for dropping a feature:

1. Drop the feature with high p value and high VIF or high p-value and low VIF
2. Drop the feature with low p value and high VIF once the above criteria is met
3. Keep the feature with low p value and low VIF

- Built a model using the features selected by RFE.
- Checked insignificance of each feature using p value
- Checked multi-collinearity using VIF.

Model Building
Part 3

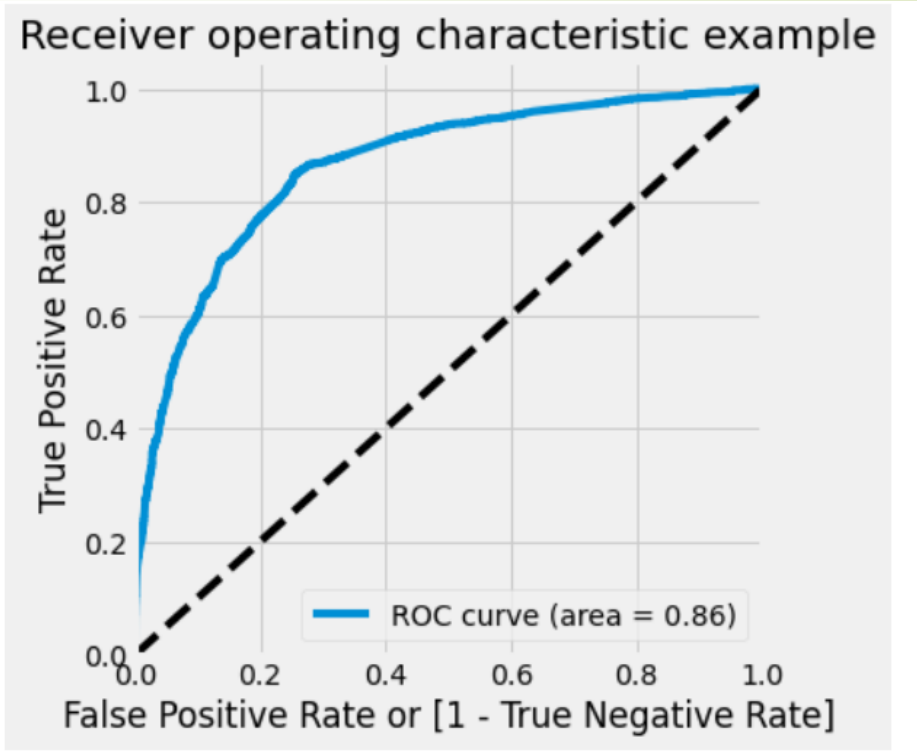
MODEL NO	HIGH P-VALUE	HIGH VIF	ELIMINATING FEATURE
2	Specialization_Services Excellence	Lead Origin_Lead Add Form	Lead Origin_Lead Add Form
3	Specialization_Services Excellence	All features have low VIF	Specialization_Services Excellence
4	Lead Origin_Lead Import	All features have low VIF	Lead Origin_Lead Import
5	Last Notable Activity_Other_Activity"	All features have low VIF	Last Notable Activity_Other_Activity"
6	Specialization_Retail Management	All features have low VIF	Specialization_Retail Management
7	All features have low P-value	All features have low VIF	---

Model Building
Evaluation Metrics

- Final Model:
  - Model is good not much of insignificance
  - Accuracy score = 78.9%
  - Sensitivity = 65%
  - Specificity = 88%

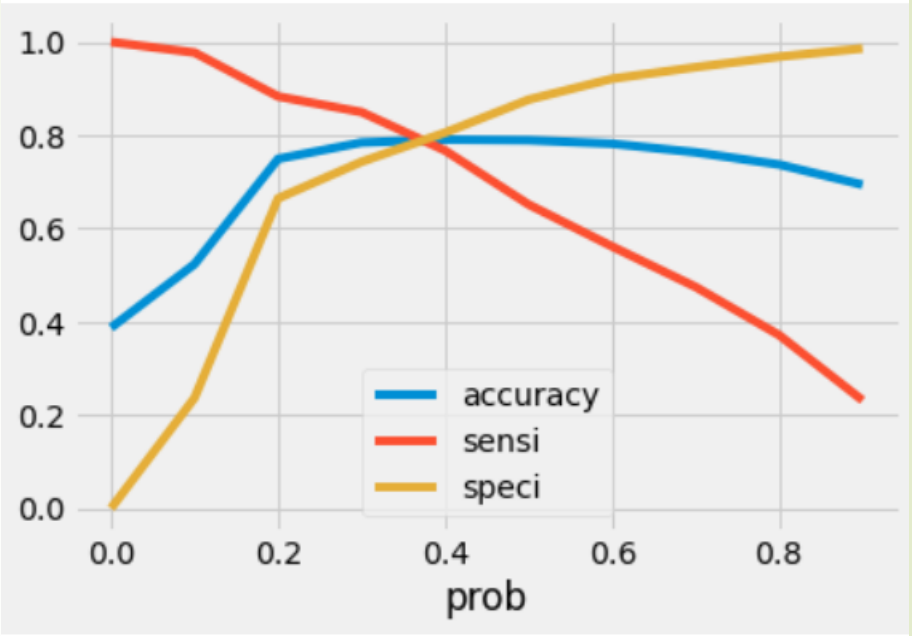
ROC

- Plotting ROC curve for converted and converted_prob



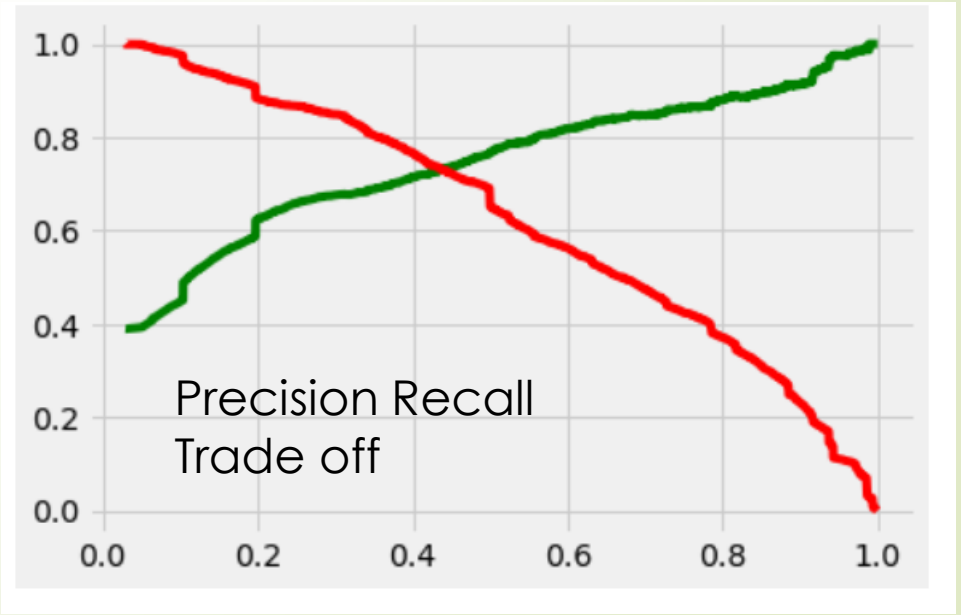
Evaluation

- Finding the Optimal cutoff point
- Looking at the graph, 0.38 is the optimum point to take it as a cutoff probability.



After taking cutoff as 0.38

- Accuracy score = 78.7%
- Sensitivity=78%
- Specificity = 79%



Making predictions on Test Data set

- Accuracy score = 47.03%
- Sensitivity =98.19%
- Why this model is good?
  - Sensitivity is the accuracy of predicting the positive classes, in this case, it is the lead conversion probability.
  - Higher the sensitivity, higher the lead conversion probability.

Observation from final Model

- Total Time Spent on website – has positive relation, if this increased then the conversion probability increases.
- A free copy of Mastering The Interview – has low negative relation
- Lead Origin – 'Landing page submission' has negative relation
- Lead Source – "Olark Chat" has low positive relation, "Reference","Welingak Website" has high positive relation
- Specialization – "Hospitality Management"," International Business"," Others" has negative relation
- Last Notable Activity – "Had a Phone Conversation"," SMS Sent"," Unreachable" has positive relation. "Modified"," Page Visited on Website" has negative relation.

Recommendations

- By looking into the amount of time spent by the customer on the website and once that customer is identified send the SMS regarding the courses , also try to reach them through Welingak Website, Olark Chat.
- Connect with the customer through a call and try to explain the importance and the ROI of the course
- Try to ask your current student for the reference and for that try to lure the current student with some sort of referral bonus for getting a new customer.