LEAD SCORING CASE STUDY

Prepared By:
Vaibhav Swarnkar
Dipali Visapure

Problem Statement & Input Data Sets

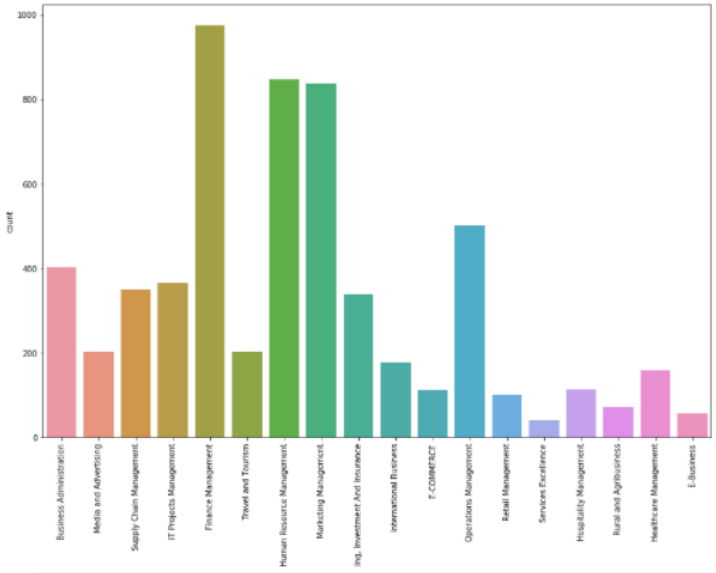
- An education company named X Education markets and sells its courses on several websites and search engines. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their contact details, they are classified to be a lead. Lead can also be got through past referrals
- Once these leads are acquired, employees from the sales team contacts them through various mediums, so the leads get converted.
- The typical lead conversion rate at X education is around 30%.
- X Education wants us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- Objective of the assignment –
  - We have to prepare a logistic regression model wherein we need to assign a lead score 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.
  - Predict whether a lead will convert or not, here the target lead conversion rate need to be around 80%.
- Input Data Sets –
  - Leads .csv

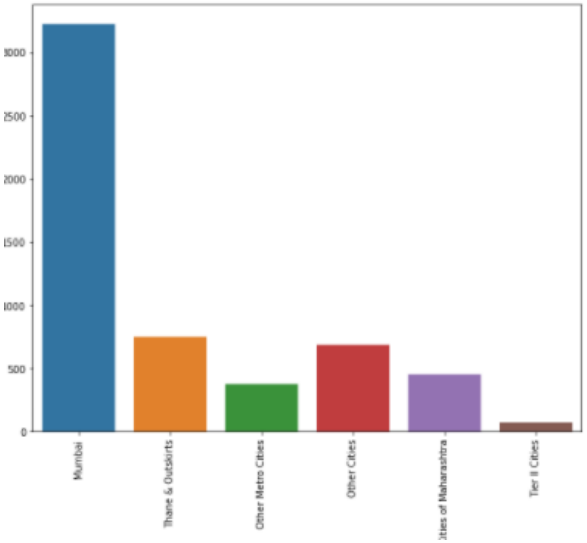
Understand and Clean the Data

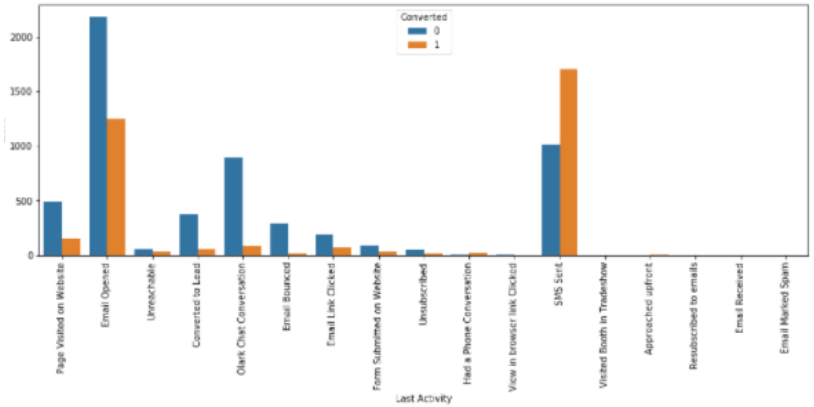
- Import the leads.csv
- Inspecting the dataset.
  - ❖ Use functions like , describe, shape, info, columns, dtypes etc.
- Data quality and missing value checks
  - ❖ Finding the percentage of missing values for all columns
  - ❖ Removing columns with **>= 40% null** values
  - ❖ Check and remove **highly skewed** columns
  - ❖ Identify columns with multiple categories with less data and group them together into separate category.
  - ❖ Identify and Impute the missing Values using techniques like median for categorical variables..
  - ❖ Check for outliers for continuous variables and use the techniques to cap the outliers, such as variable Page view per visit
  - ❖ Create dummy variables
  - ❖ Check the correlations using heatmap, drop the highly correlated variables like Last Notable Activity_SMS Sent, Lead Source_Others, Page Views Per Visit, Last Activity_Olark chat Conversation

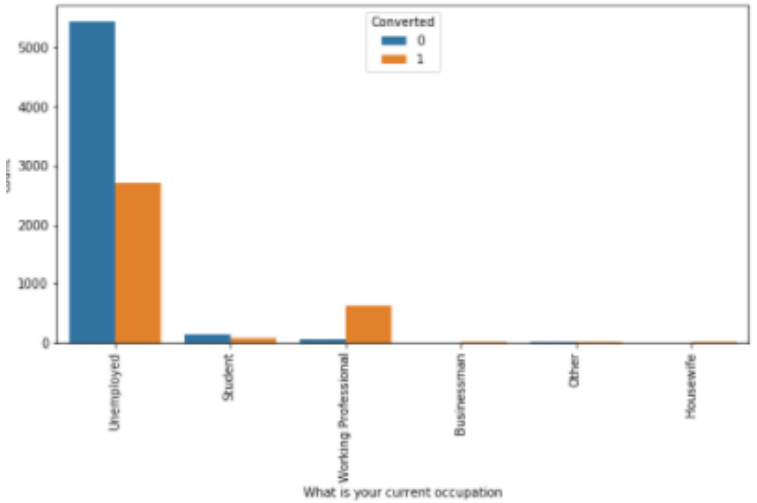
Data Visualizations

- Univariate and Bivariate Analysis



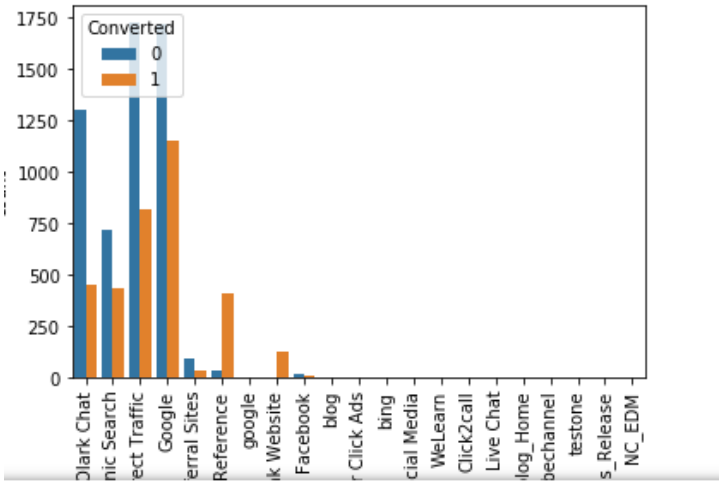


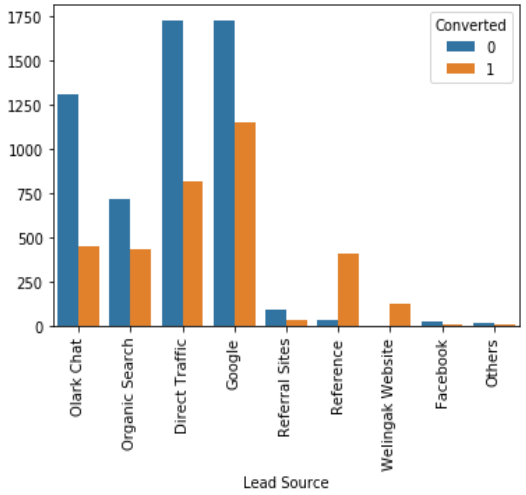


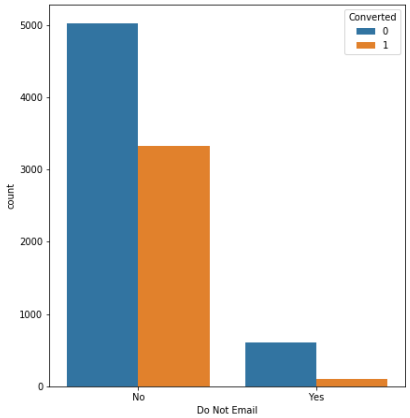


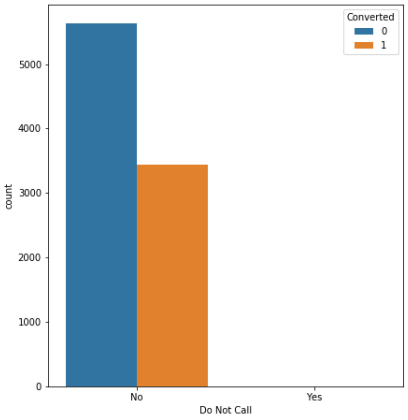
Data Visualizations

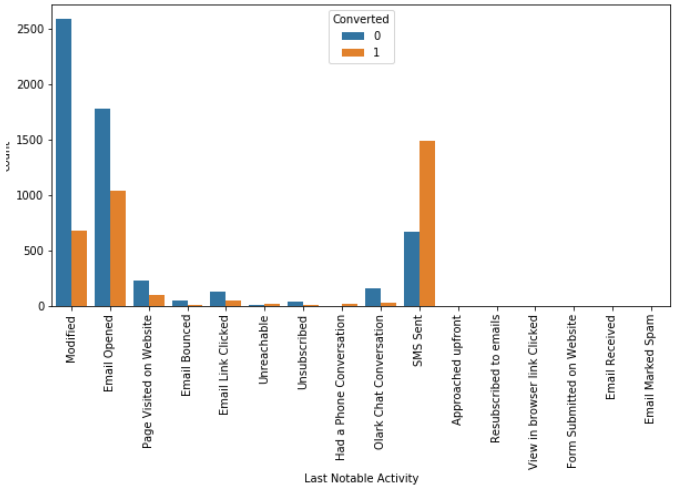
- Univariate and Bivariate Analysis





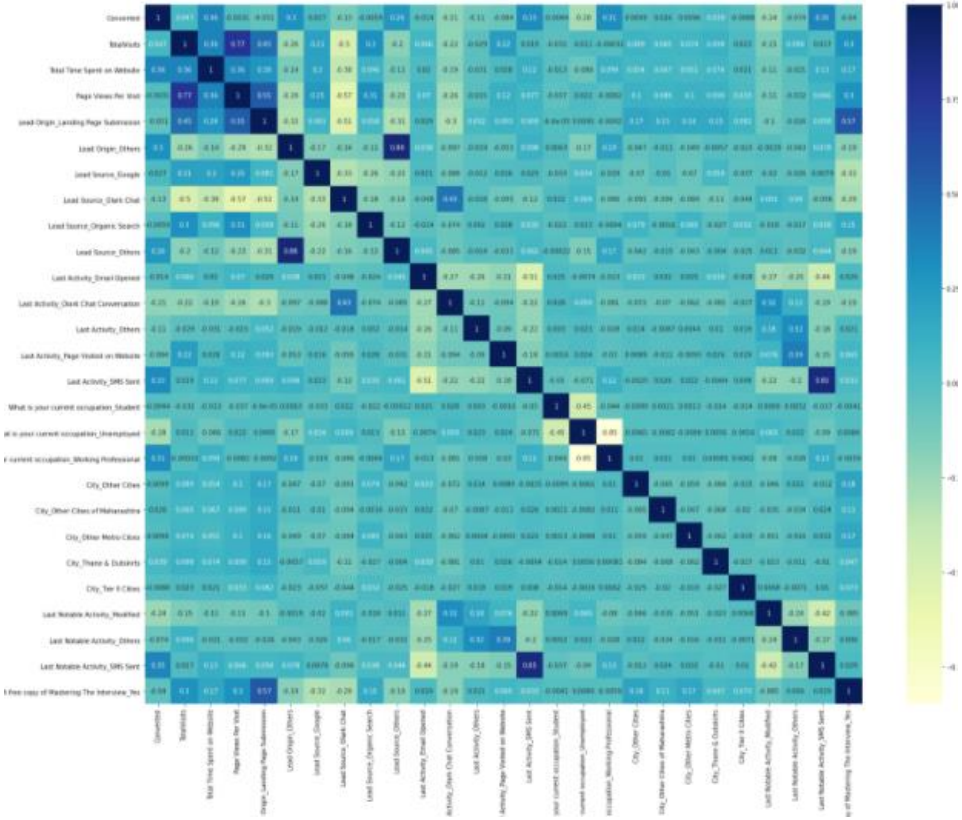






Data Visualizations

- Check the correlations between variables with heatmap



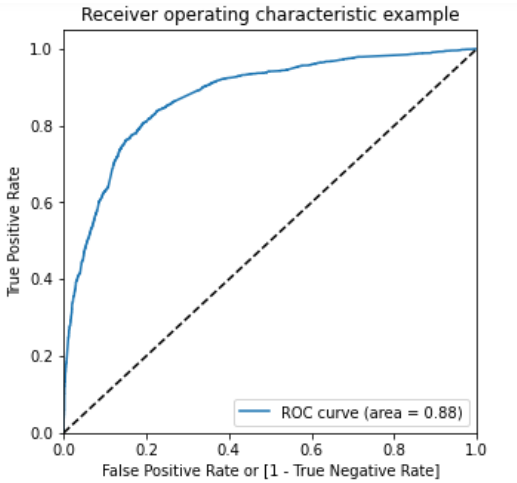
- Drop the columns which are hioghly correlated like, Last Notable Activity_SMS Sent', 'Lead Source_Others', 'Page Views Per Visit', 'Last Activity_Olark Chat Conversation'

Model Building

- Splitting the data set into train & test sets
- Scaling of data – Scale the data using StandardScaler method
- Drop the Target Variable
- Build the Logistic Regression model
- Using Recursive Feature Elimination method to select top 18 columns
- Building model with the columns that are chosen by RFE.
- Make multiple iterations of model building by manual elimination of features by checking p-values and VIF after each iterations.
- Features like City_Other, Last Notable Activity_Others, Last Activity_Page Visited on Website, What is your current occupation_Student, What is your current occupation_Unemployed are removed after manual eliminations

Checking Predictions on Train Data

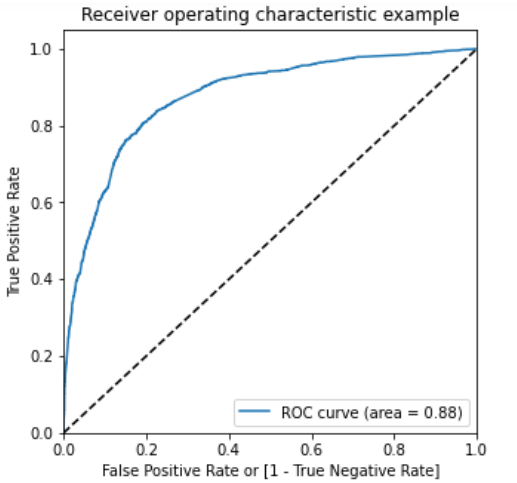
- Check various predictions on the prepared model
- Create a new column 'predicted' with cut-off 0.5
- Prepare confusion matrix
- Check various parameters like accuracy(81), sensitivity(70), specificity(87) etc
- Plot ROC - Receiver operating characteristic curve and create table of probability cut-offs for diff probabilities.



	prob	accuracy	sensi	speci
0.0	0.0	0.388837	1.000000	0.000000
0.1	0.1	0.593074	0.967396	0.354920
0.2	0.2	0.734539	0.921670	0.615482
0.3	0.3	0.796228	0.842545	0.766759
0.4	0.4	0.812152	0.771769	0.837845
0.5	0.5	0.810761	0.706561	0.877055
0.6	0.6	0.790662	0.593638	0.916013
0.7	0.7	0.771336	0.506561	0.939793
0.8	0.8	0.745826	0.395229	0.968884
0.9	0.9	0.703772	0.257654	0.987604

Checking Predictions on Train Data

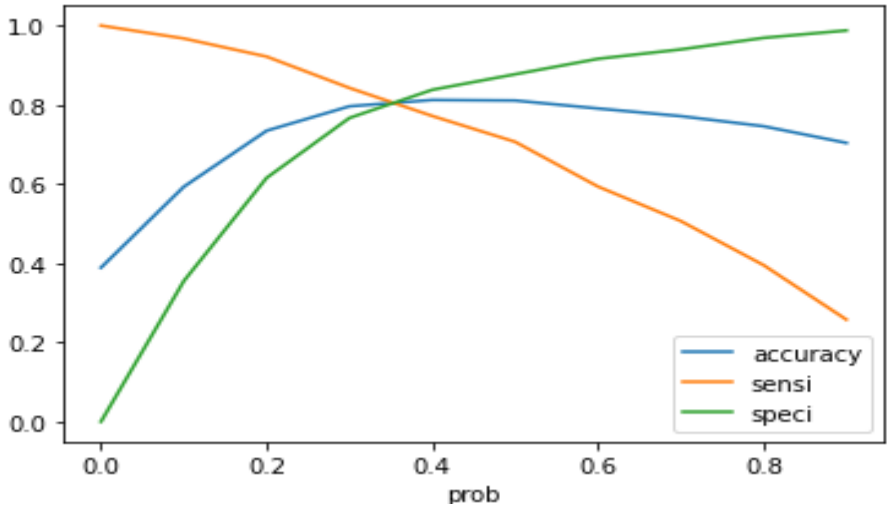
- Check various predictions on the prepared model
- Create a new column 'predicted' with cut-off 0.5
- Prepare confusion matrix
- Check various parameters like accuracy(81), sensitivity(70), specificity(87) etc
- Plot ROC - Receiver operating characteristic curve and create table of probability cut-offs for diff probabilities.



	prob	accuracy	sensi	speci
0.0	0.0	0.388837	1.000000	0.000000
0.1	0.1	0.593074	0.967396	0.354920
0.2	0.2	0.734539	0.921670	0.615482
0.3	0.3	0.796228	0.842545	0.766759
0.4	0.4	0.812152	0.771769	0.837845
0.5	0.5	0.810761	0.706561	0.877055
0.6	0.6	0.790662	0.593638	0.916013
0.7	0.7	0.771336	0.506561	0.939793
0.8	0.8	0.745826	0.395229	0.968884
0.9	0.9	0.703772	0.257654	0.987604

Checking Predictions on Train Data

- Plot accuracy sensitivity and specificity for various probabilities.



- From the curve above, 0.3 is the optimum point to take it as a cutoff prob
- ability as we want to achieve good Sensitivity
- Assign Lead Score
- Check the parameters again for prob cut-off 0.3, like sensitivity(84.25), specificity(76.67) etc

Checking Predictions on Test Data

- Splitting test data into X and y.
- Perform Scaling on test data as well for continuous variables
- Check predictions of various parameters like accuracy(79), sensitivity(83), specificity(76) etc.

Summary of the Model

- **Top 5 variables that contribute the most in this model are:**

1. Lead Origin_Others - Lead Origin of type Other
2. What is your current occupation-Working professionals
3. Last Activity_SMS sent - Leads to whom SMS is sent
4. Lead Source_Olark Chat
5. Total Time Spent on Website

- **Final Suggestions from the model:**

- Call leads whose origin are from Lead Add Form, Lead Import or Quick Add Form.
- Leads to whom SMS was sent and are Working Professional must be chosen.
- Consider leads whose source is from Olark Chat and have highest time spent on website.
- Employees may focus on sending email as it is seen leads who are opening their emails and reading the offer are also contributing highly in conversion.
- Making UI/UX of the website will also help in the converting the leads as this improves the trust of users on the course and the company.
- Optimizing SEO also helps heavily in conversion rate. It is seen that leads that are coming from Google are also getting converted into customer at a higher rate as it is considered that these kinds of users they know what they are looking for and more often or not these users also know the company they want buy from.