Leads Scoring Case Study

By - Utkarsh Updhayay

Mayank Singh Soni

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, 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.

Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

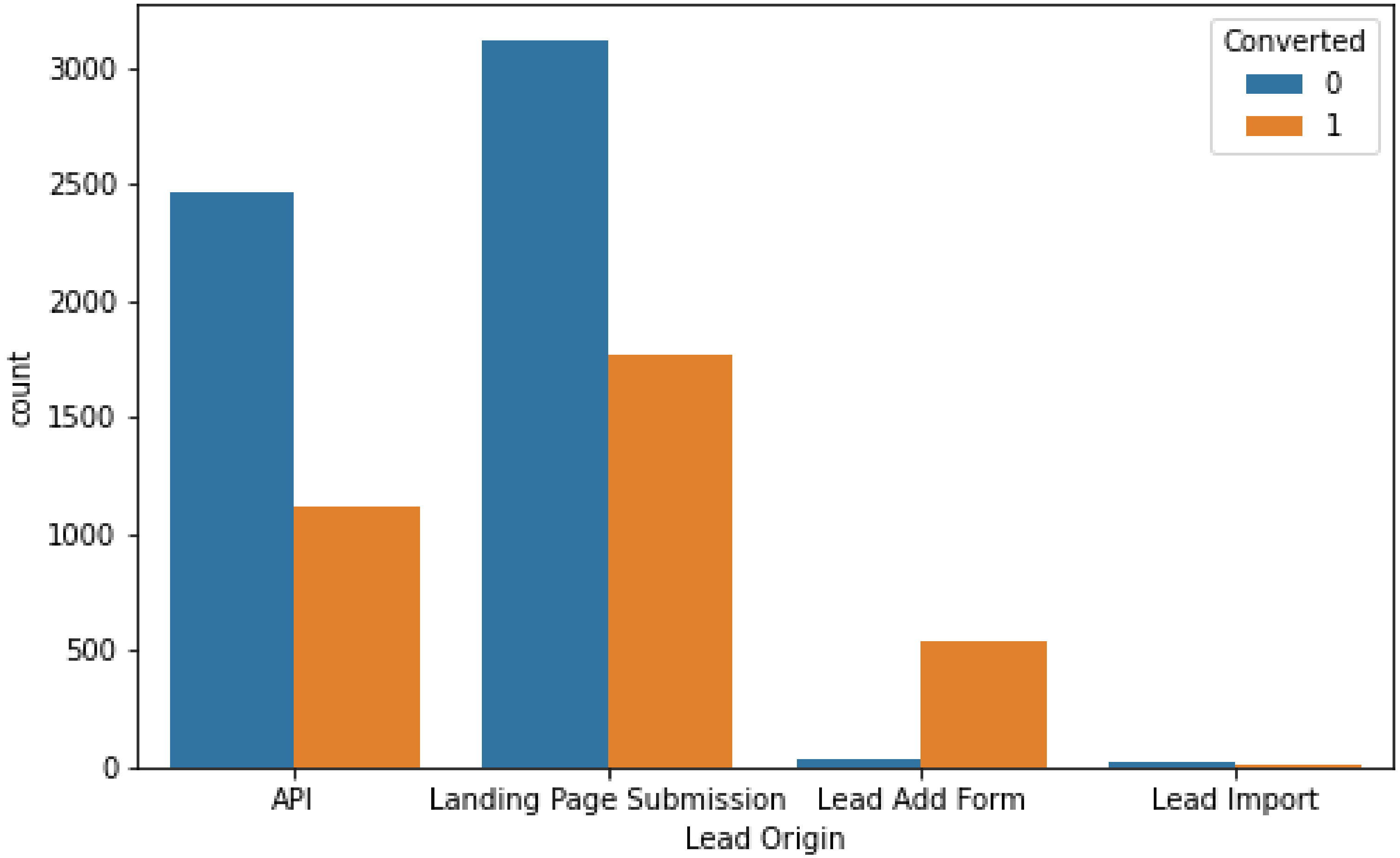
- Data cleaning and data manipulation.
  - Check and handle duplicate data.
  - Check and handle NA values and missing values.
  - Drop columns, if it contains large amount of missing values and not useful for the analysis.
  - Imputation of the values, if necessary.
  - Check and handle outliers in data.
- EDA
  - Univariate data analysis: value count, distribution of variable etc.
  - Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations

Data Cleaning and Data Manipulation

- The Shape of the data is 9240 rows and 37 columns.
- Inspecting the data and found “Prospect ID” and “Lead Number” should be dropped due to having low variance data.
- Inspecting Null values and dropped those columns which have null values greater than 40% as they are not going to be a part of final model.
- After further inspection, found missing values which is imputed with the highest occurring value in individual column (if required)
- Dropping those columns who are highly screwed in nature as they are not to be considered for final prediction.

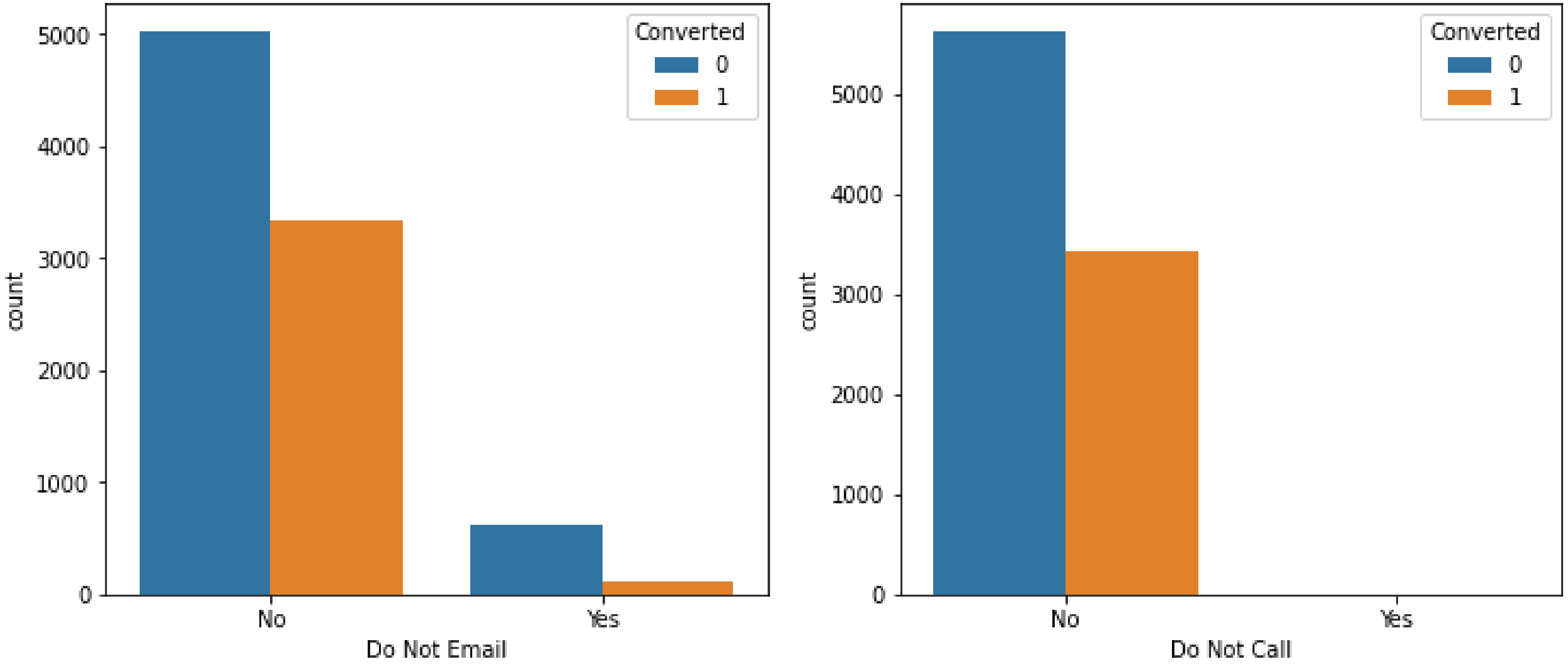
EDA

Leads w.r.t. Converted



- Plotting Leads with respective to our target variable that is Converted.
- Highest lead count and conversion rate is from Landing Page Submission followed by API
- Lead Add Form has low number of lead but very high conversion rate
- Lead Import have lowest Leads and Conversion Ratio

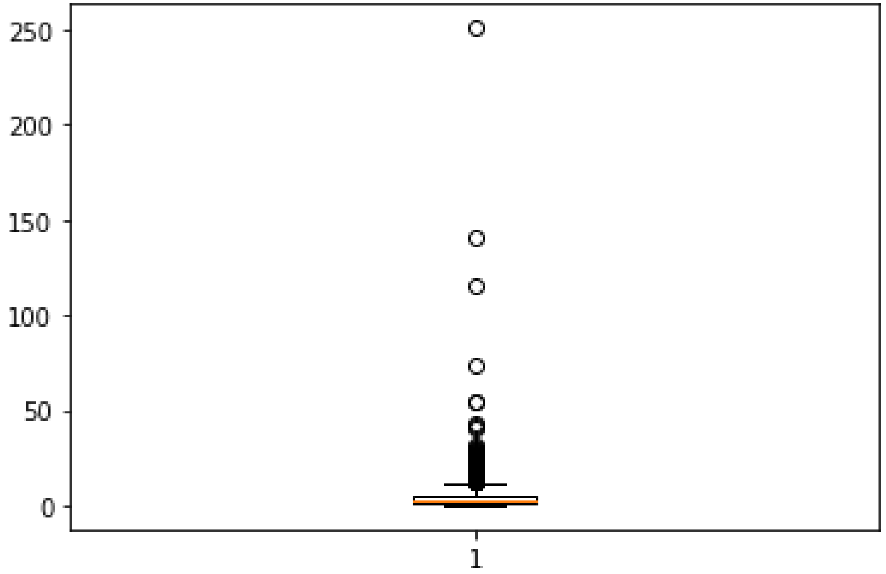
Do Not Email & Do Not Call w.r.t Target Variable



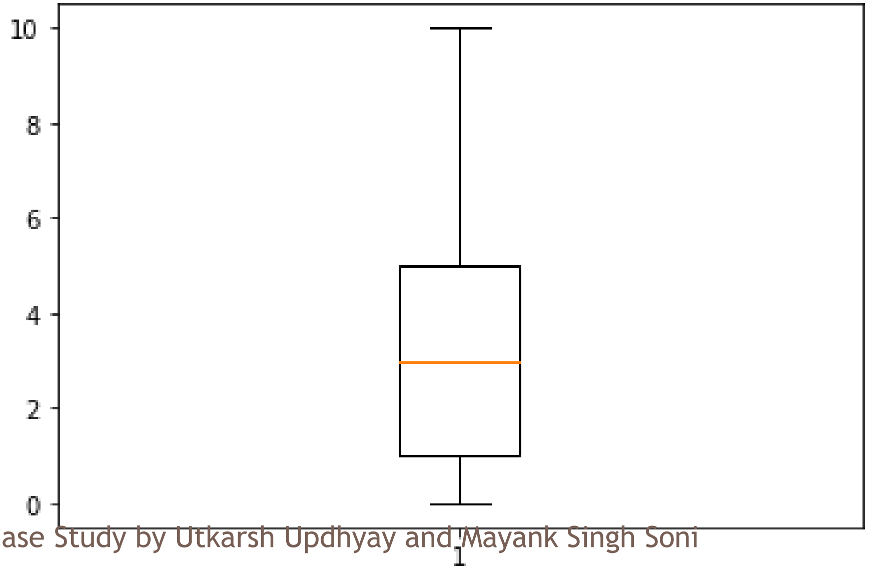
EDA

- Plotting 'Do Not Email' & 'Do Not Call' w.r.t. target variable Converted.
- People who respond to calls and email have high chances to being converted as Hot Lead
- Do not call have most values as No in it so it is imbalanced and can be dropped from analysis

Before Outlier Treatment



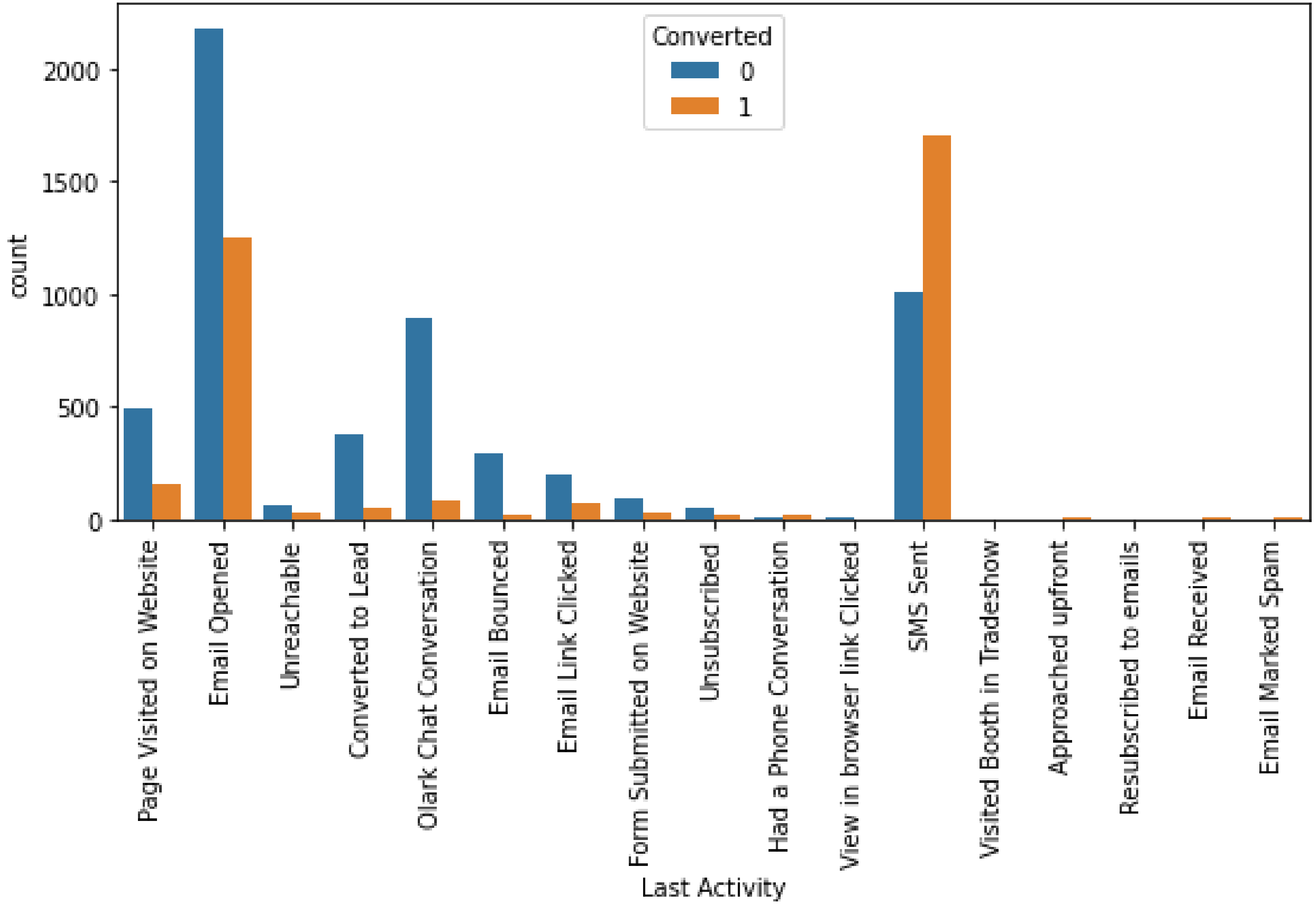
After Outlier Treatment



EDA

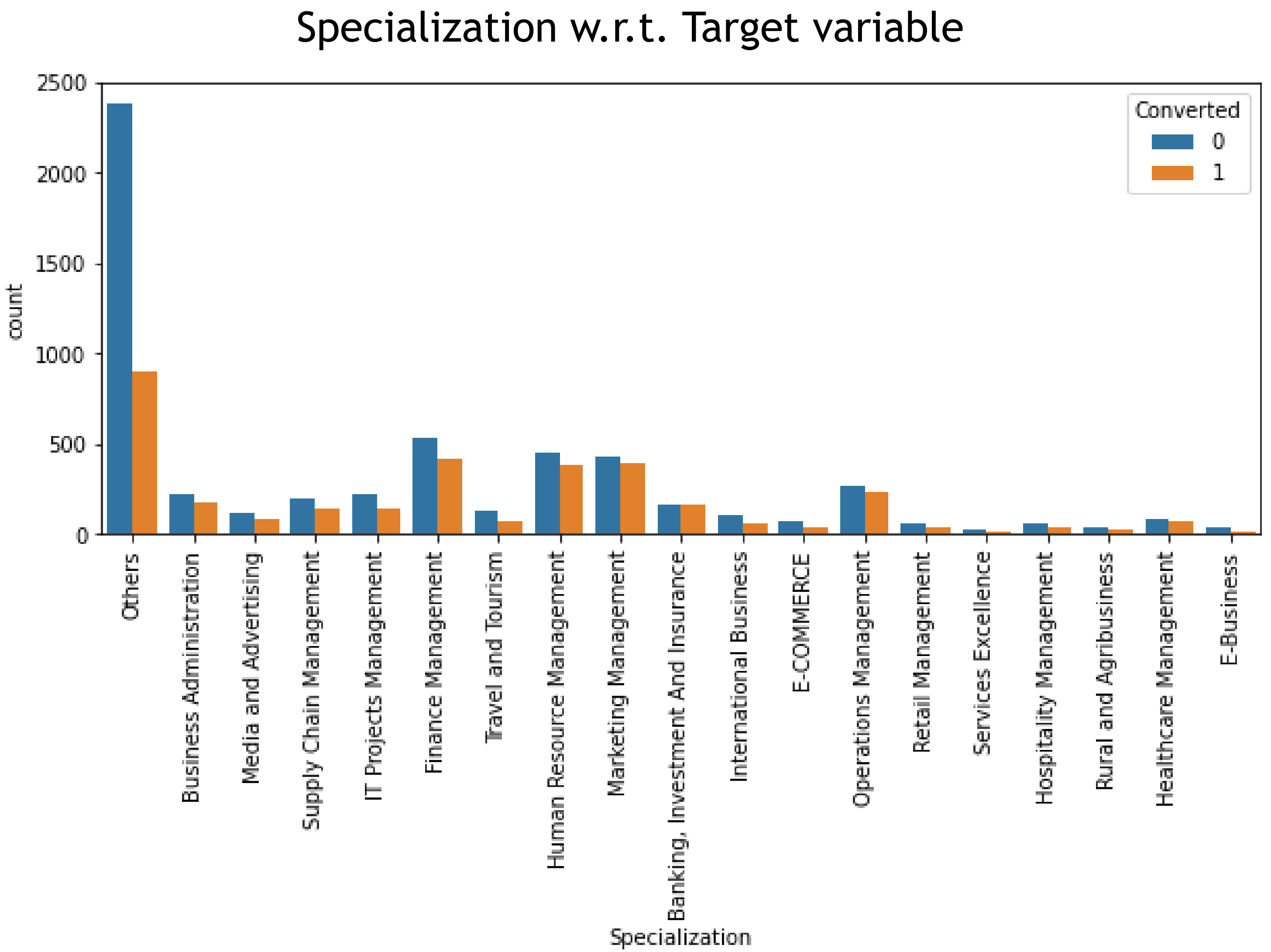
- Plotting box plot for 'Total Visits‘ variable.
- We can see there are a lot of outliers present, we can cap the outliers at 95th percentile.
- Now that outliers are treated, comparing it against Target Variable.

Last Activity w.r.t. Target variable



EDA

- Plotting 'Last Activity' w.r.t. target variable 'Converted‘.
- Most of the people has 'Email Opened' as their Last Activity.
- People having 'SMS Sent' as their Last Activity have the highest conversion.



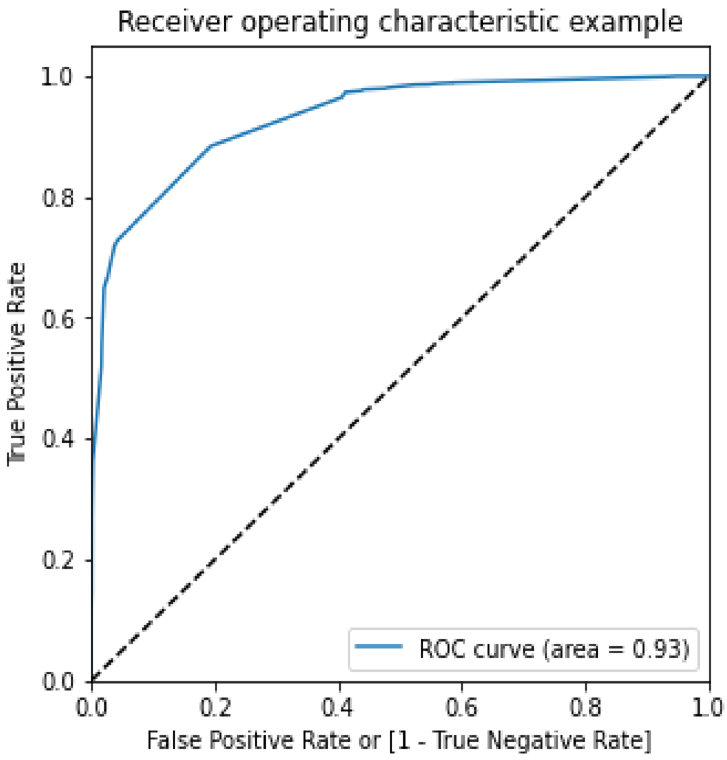
EDA

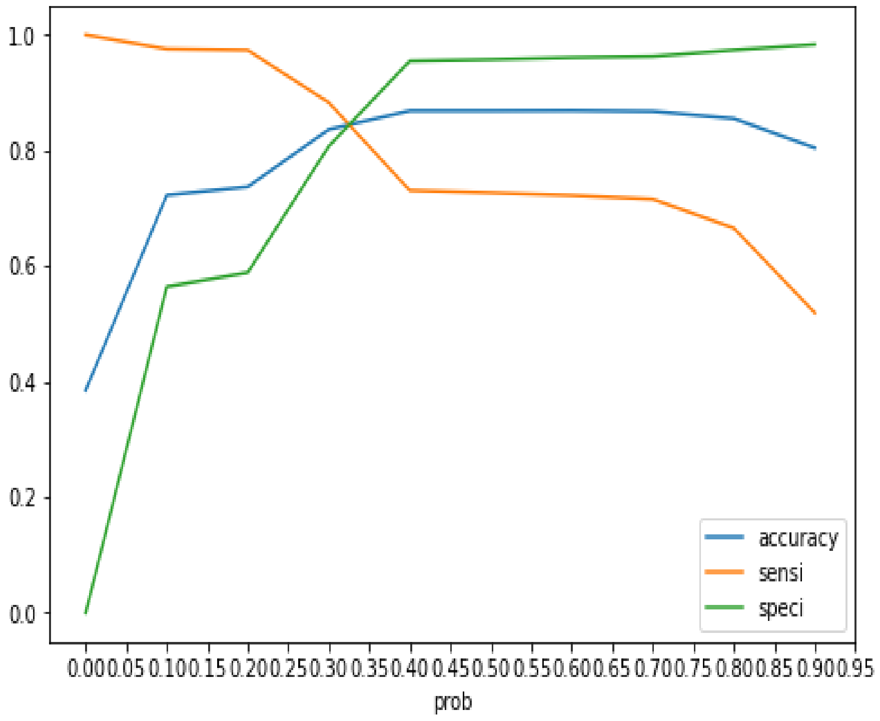
- It can be a case where the Specialization is not mentioned because it was not a option in drop down as we are mostly seeing 'Management' but there can be others like Analyst, Architect, CTO, CFO etc. or does not have any Specialization or are they are students.
- We Can replace such missing value with 'Others‘.
- Various Management and Business Specialization have high conversion rate, so more focus should be there.

Model Building

- Splitting the Data into Training and Testing Sets.
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection.
- Running RFE with 20 variables as output.
- Building Model by removing the variable whose p- value is greater than 0.05 and VIF value is greater than 5.
- Predictions on test data set.
- Overall accuracy 86.83%.

ROC Curve





- Finding Optimal Cut off Point
- Optimal cut off probability is that probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.33

Conclusion

So Looking at the Final Model, we can conclude that below are the factors that contribute the most in getting a lead converted

1. Lot of focus should be made to leads with current status of lead being ‘BUSY’, ‘CLOSED BY HORIZZON’, ‘RINGING’, ‘LOST TO’ & ‘WILL REVERT AFTER READING EMAIL’.
2. Another focus point should be the Lead Source ‘WEBLINGAK WEBSITE’ and Lead Origin ‘LEAD ADD FORM’
3. People having Current Occupation as ‘WORKKING PROFESSIONAL’ should be targeted as they are high potential lead source
4. Another are of improvement for better Lead conversion is people with Last Activity ‘HAD A PHONE CONVERSATION’ and Last Notable Activity as ‘SMS SENT’.
5. If there are limited sales representatives, then score cut-off should be higher to ensure a higher conversion probability people are contacted. In case there are more resources available in the sales team (i.e., interns, etc.), then the score cut-off can be lowered for better reach and more number of leads.

Thank - You