Lead Score Case Study

Shubhra Mishra

Rajeev K Balakrishnan

Problem Statement:

- Industry professionals can purchase online courses from X Education, a company that provides education. Many experts who are interested in the courses visit their website on any given day and search for courses. On their website, there is a process for filling out forms, following which the business receives the user as a lead.
- Once these leads are obtained, sales team members begin calling, sending emails, etc. Some leads are converted during this procedure, but most are not.
- At X Education, the normal lead conversion rate is roughly 30%. This implies that only around 30 of their leads will be converted if, for example, they receive 100 leads in a day. The business wants to discover the most promising leads, also known as Hot Leads, to make this process more effective.
- The lead conversion rate ought to increase if they are successful in finding this group of leads, since the sales team will now concentrate more on speaking with potential leads rather than calling everyone.

Business Objective :

- To assign a lead score between 0 and 100, create a logistic regression model.
- A greater number indicates whether a lead is hot or cold, or if it is likely to convert.
- The CEO aims for an 80% lead conversion rate.
- They want the model to be able to manage future limitations as well, such as activities that must be taken during peak times, how to use all available manpower, and what should be done once the goal has been reached.

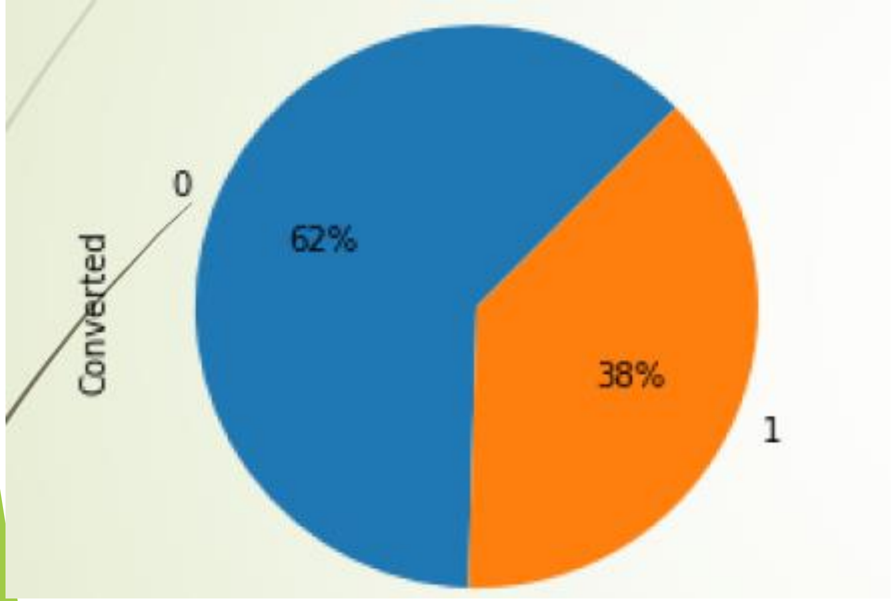
Solution Methodology

- Dataset Reading and Inspection
  - Reading the Data set
  - Inspecting the data
  - Checking data dimensions
  - Checking for All datatypes
  - Checking continuous values distribution
  - Checking for Null value content
- EDA
  - Analysis of univariate data: value count, variable distribution, etc.
  - Bivariate data analysis: patterns between the variables and correlation coefficients, etc.

Solution Methodology

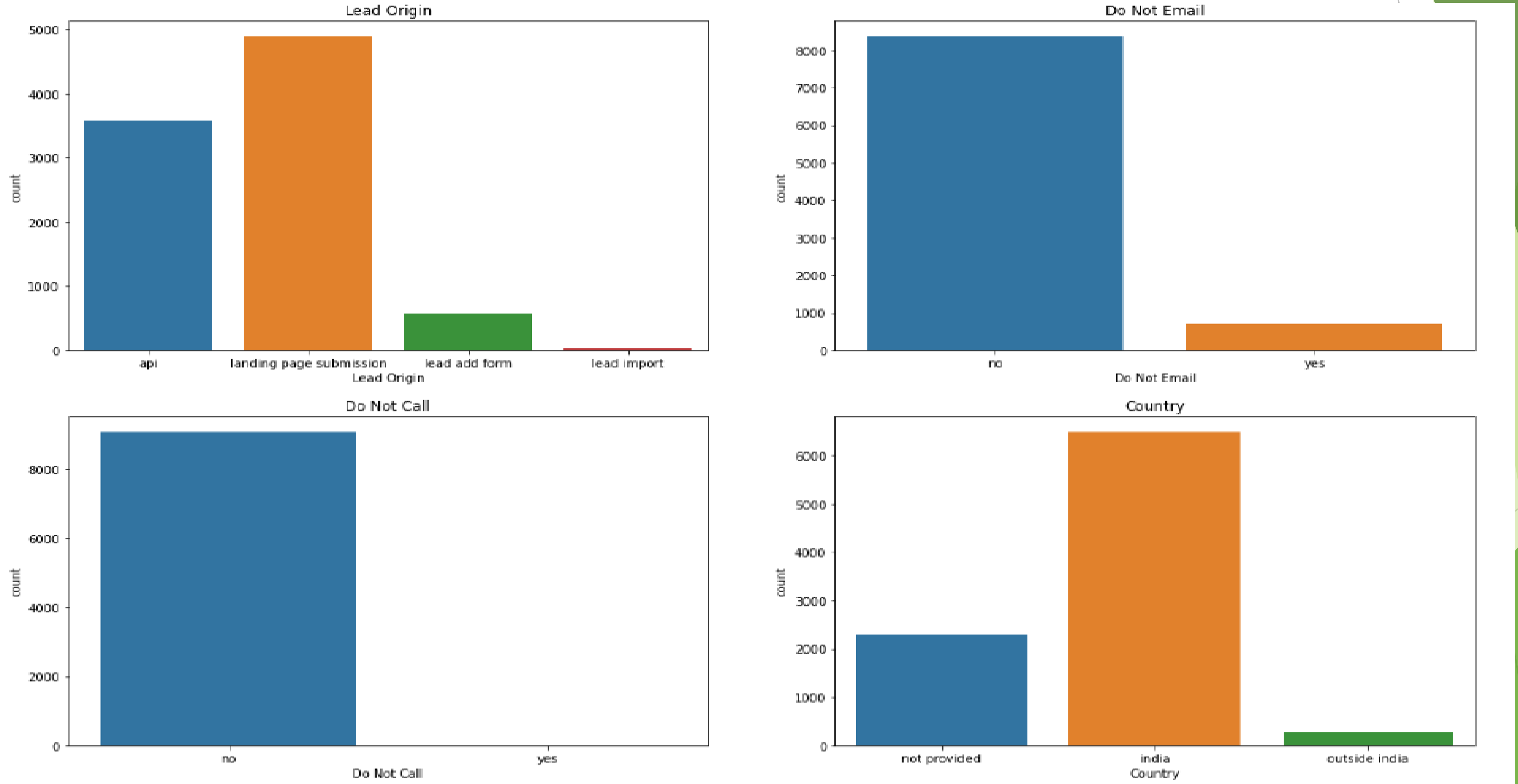
- Data Preparation and Splitting
  - Converting categorical data to binary
  - feature scaling, dummy variables, and data encoding.
- Model Building, Assigning lead scores and checking performance metrics
  - Logistic regression is a classification technique that is used to create models and make predictions.
  - Using RFE(Recursive feature elimination)
  - Model Building using features selected by RFE Validation of the model.
  - Evaluating all the
  - metrics.

Checking the Data Distribution :

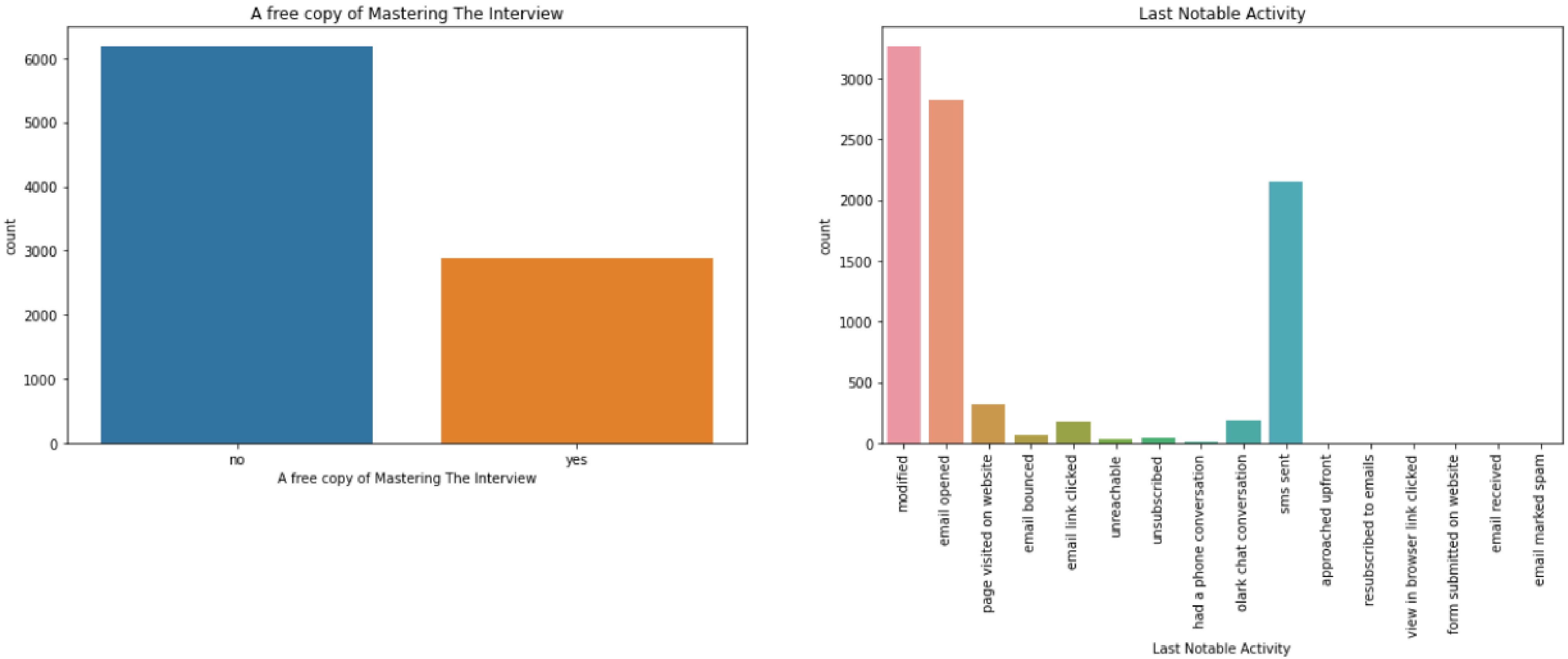


- The lead conversion rate, at 38% of the existing dataset, is low, as the figure makes abundantly clear. Lead X Education is currently dealing with this business, and that is the only issue we are here to address.

EDA of X_leads data



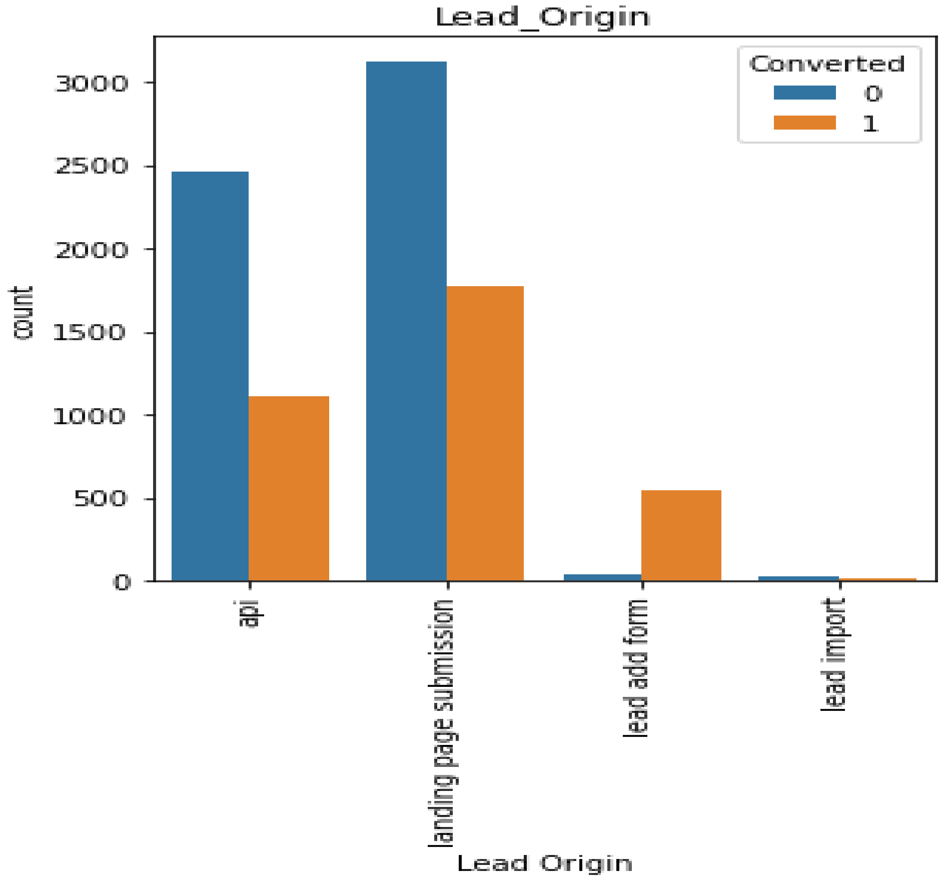
EDA of X_leads data

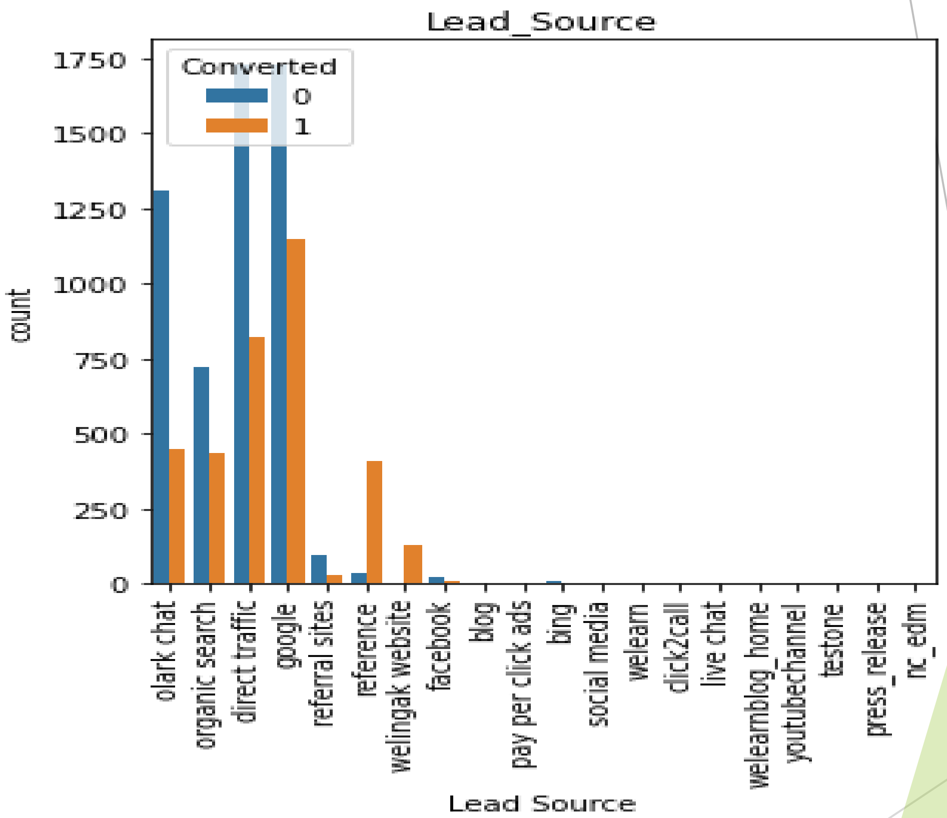


A free copy of Mastering The Interview

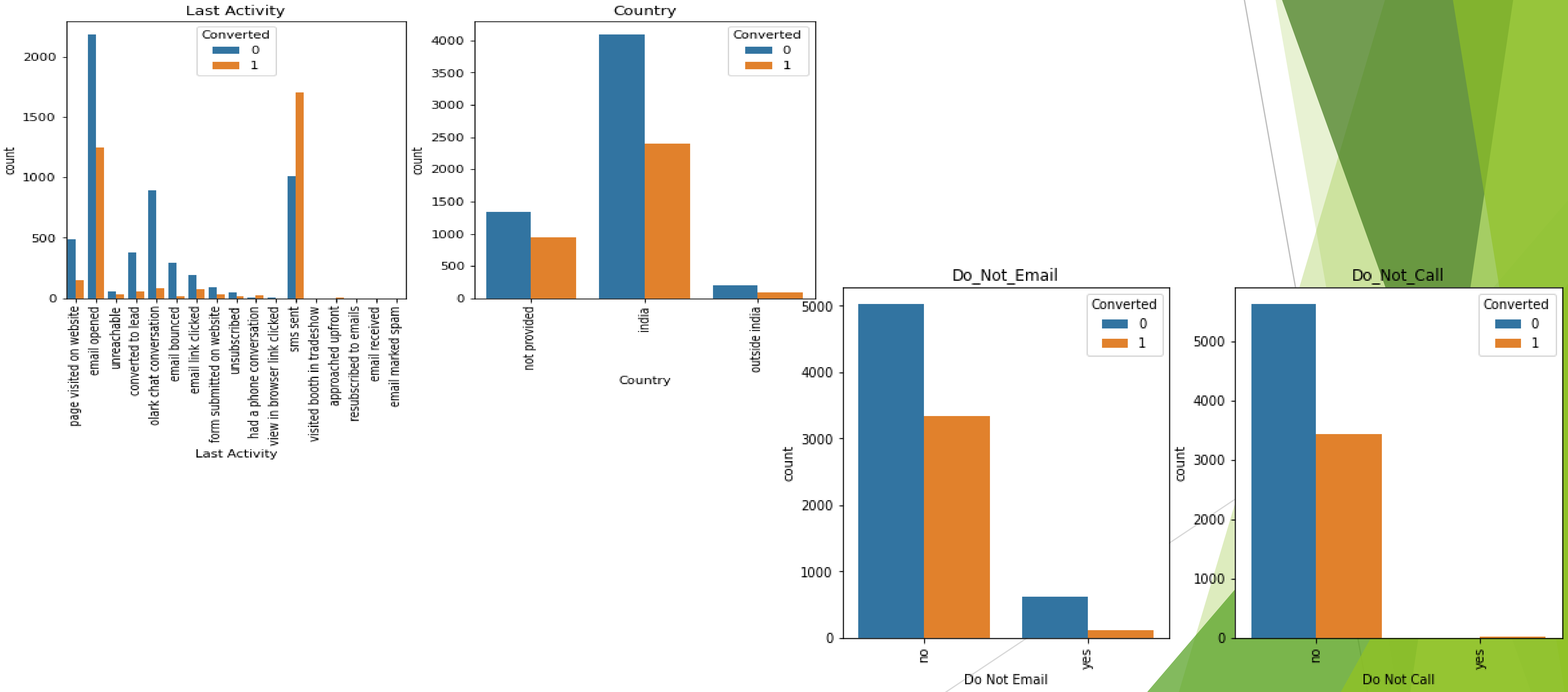
Last Notable Activity

EDA of X_leads data





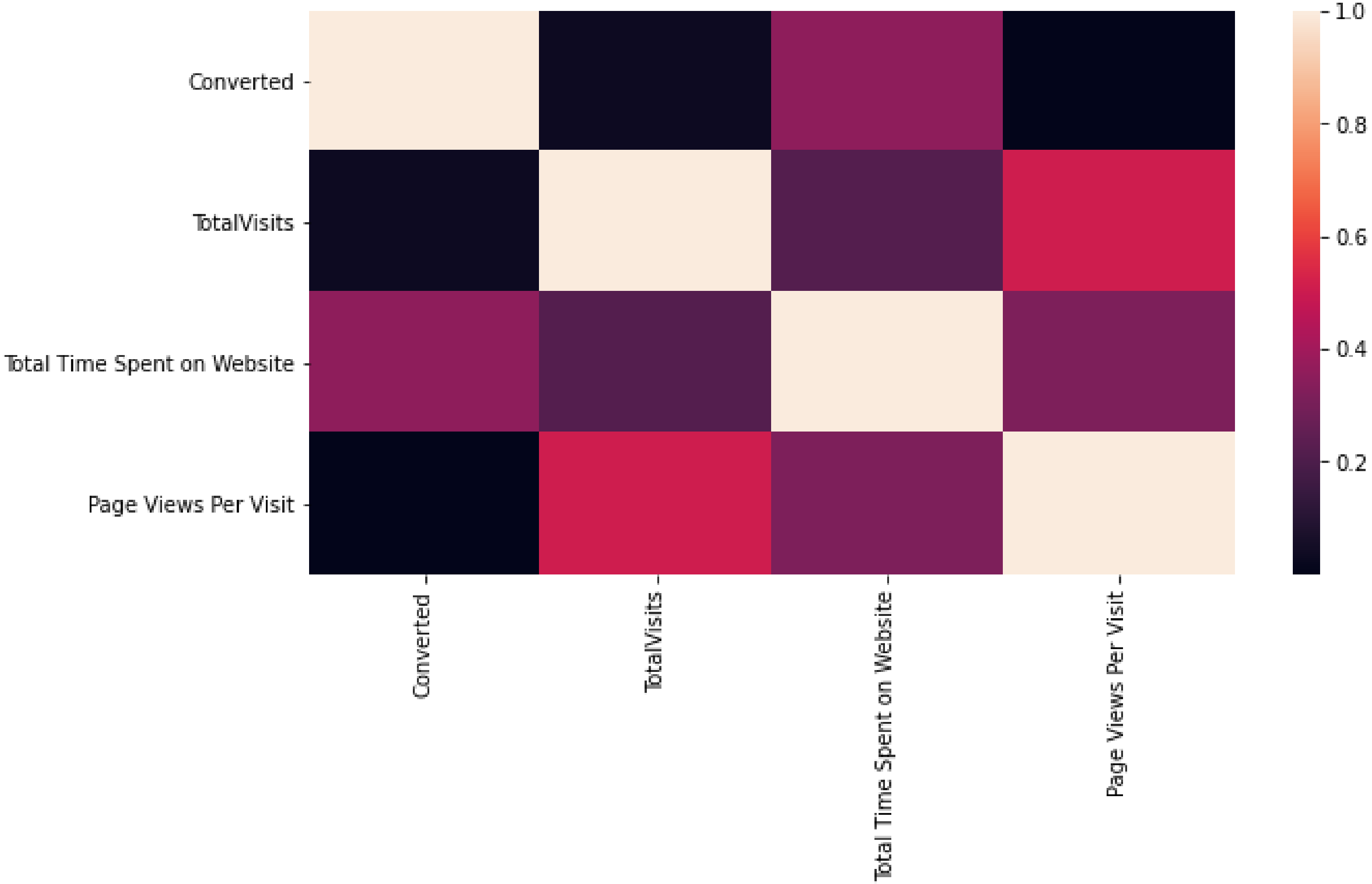
EDA of X_leads data



EDA of X_leads data:

- Some of the continuous type data columns contain outliers. Therefore, we eliminated any values above the 99% percentile in the columns for Total visits and Page per view per Visits. Later, we receive the distributions listed below. It demonstrates a more even distribution between the two columns.
- Although Lead Add Form has Excellent Conversion Rate Lead Import Has Very Little And Data Available, Most of The Lead Converted Are From Landing Page Submission And Then API
- We observe that Google and direct traffic create the majority of leads. The maximum conversion ratio is found on the Welingak website.
- Both the Do Not Call and Do Not Email variables have a strong propensity to be no and have low lead conversion rates.
- The majority of leads are produced by SMS sent activity, followed by email opened activity; all other categories have relatively low lead conversion rates. There is extremely little or no data available for many categories, including Approached upfront, visited booth at expo, and email flagged as spam.

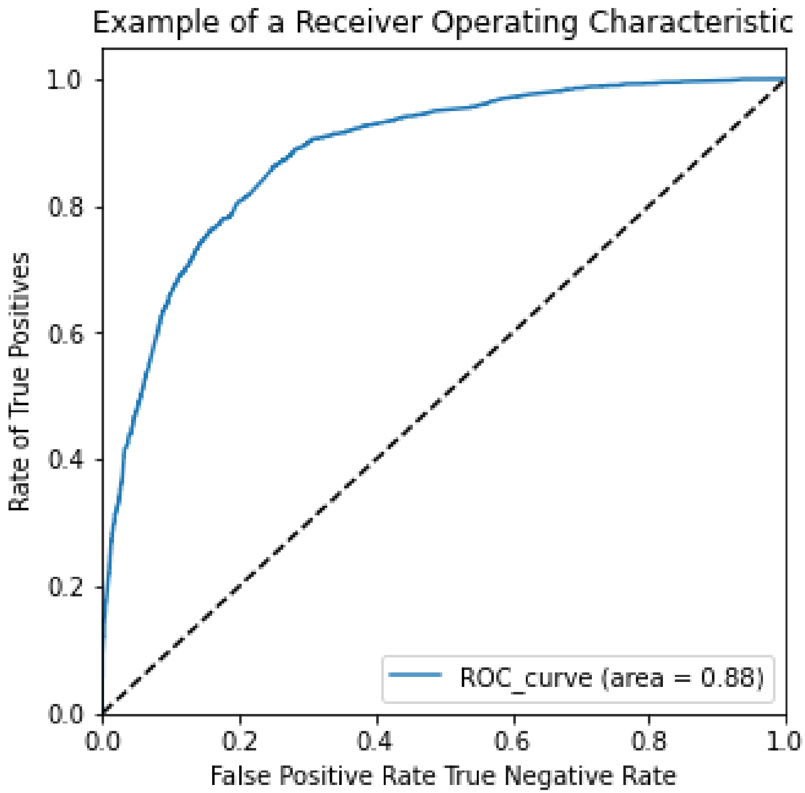
Heat map and corelation:

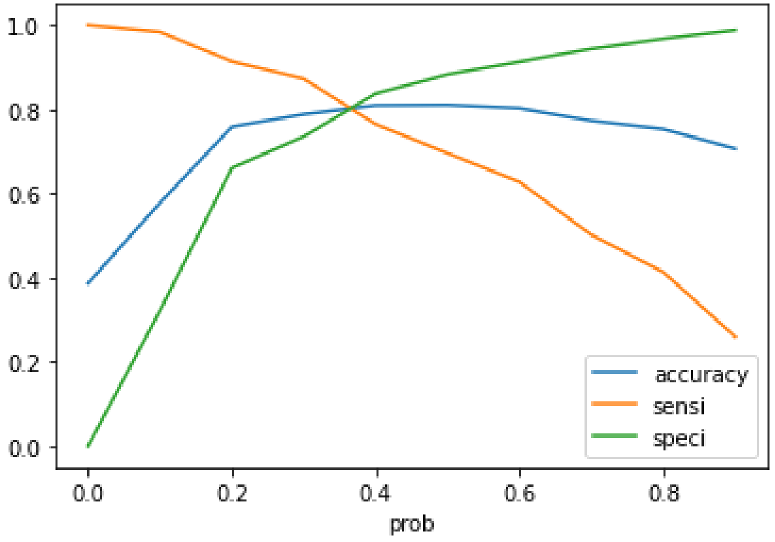


Model Building

- Splitted the Data into Training and Testing Sets
- A train-test split is the first fundamental stage in the regression process; we have selected a 70:30 split.
- Run RFE with 15 variables as output when using it for feature selection
- Removing variables from the model whose p-value is higher than 0.05 and vif value is higher than 5
- Overall accuracy of predictions on the test dataset is 81%.
- ROC optimum point finding .(ROC area 0.88)

ROC Curve:





- The probability with the best sensitivity and specificity is known as the optimal cut off probability.
- The second graph makes clear that 0.35 is the ideal cut off.

Conclusion:

- According to research, the following factors affected potential purchasers the most (in descending order):
  - The total time spend on the Website
  - When the lead source was:
    - Lead add form
    - Olark Chat
  - Total number of visits
  - while they are a working professional at the time.
  - When the last activity was
    - Olark chat conversation
    - SMS
  - We can see that the conversion rate for API and landing page submission is 30–35%, which is about typical. However, the Lead Add form and Lead import are quite low. As a result, we can intervene and say that we should pay closer attention to the leads generated by API and landing page submission.