Lead Scoring Case Study

Amit Mishra

Problem Statement

Develop a logistic regression model with a conversion rate of at least 80 percent

- 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.

- Assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

Tasks performed

- Understanding the Data
  - Carefully analysed all the columns and their relevance, to make meaningful insights

- Identifying and handling missing values
  - Used techniques like Mean and Median for continuous numerical columns
  - Used Mode for categorical column
  - Dropped columns with high missing percentage

- Identifying and handling outliers
  - Used techniques like Boxplot to identify outliers in continuous numerical column

- Creating Dummies for categorical columns

Tasks performed

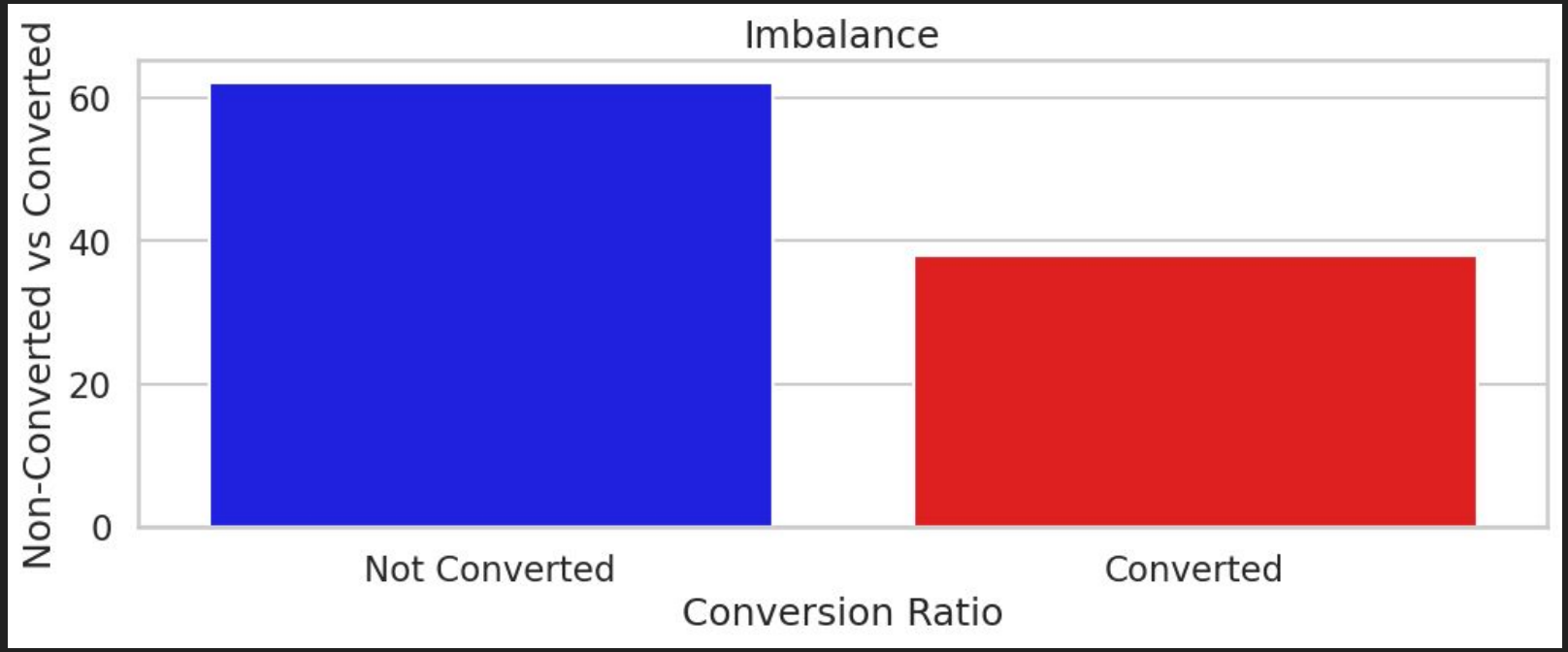
- Performed Univariate, Bivariate and Multivariate analysis
  - Analysed the relationship between the variables using different graphs and make inferences using them

- Feature selection using RFE
  - Used RFE to identify the feature that are more relevant.

- Building the logistic regression model
  - Manual Feature elimination using p-value and VIF
  - Split data into train-test sets
  - Model performance using metrics like accuracy, sensitivity, specificity etc.

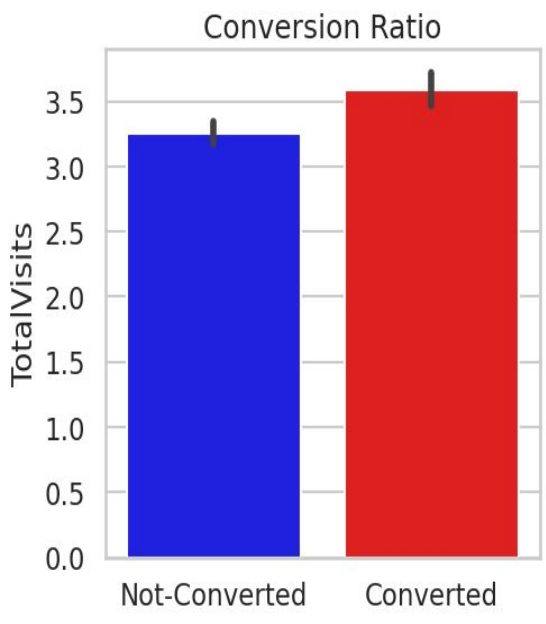
- Making inferences
  - Depending on the prediction make inferences on the features used in the model

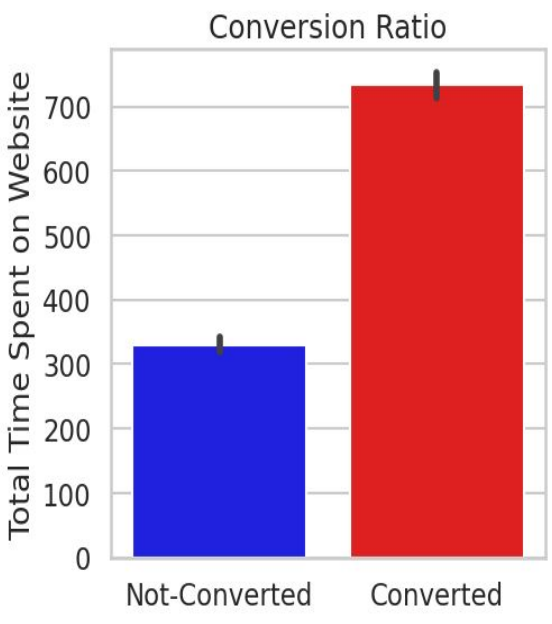
Imbalance percentage

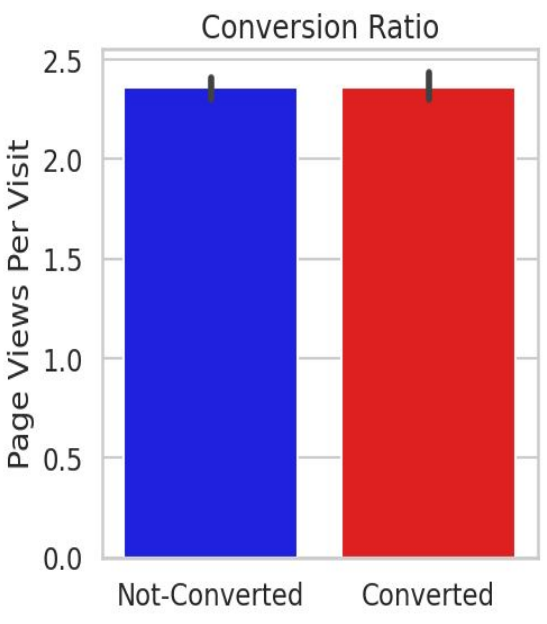


- The conversion rate is close to 40 percent.

Numerical Features

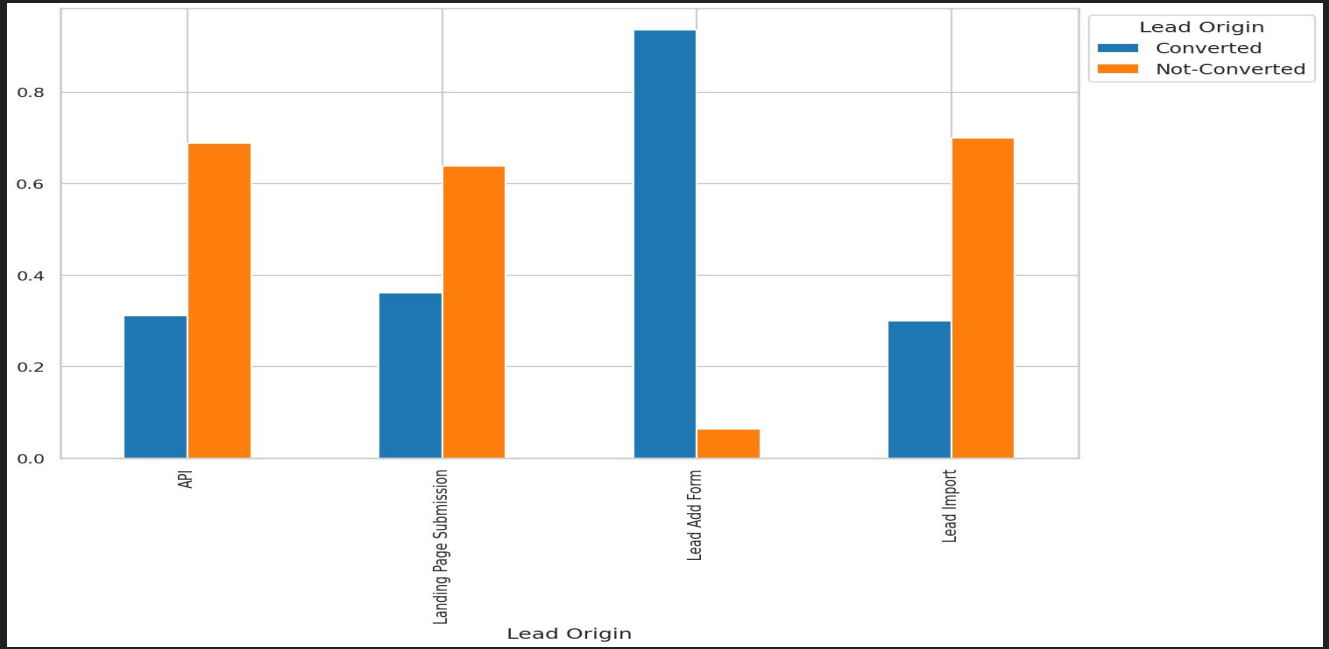






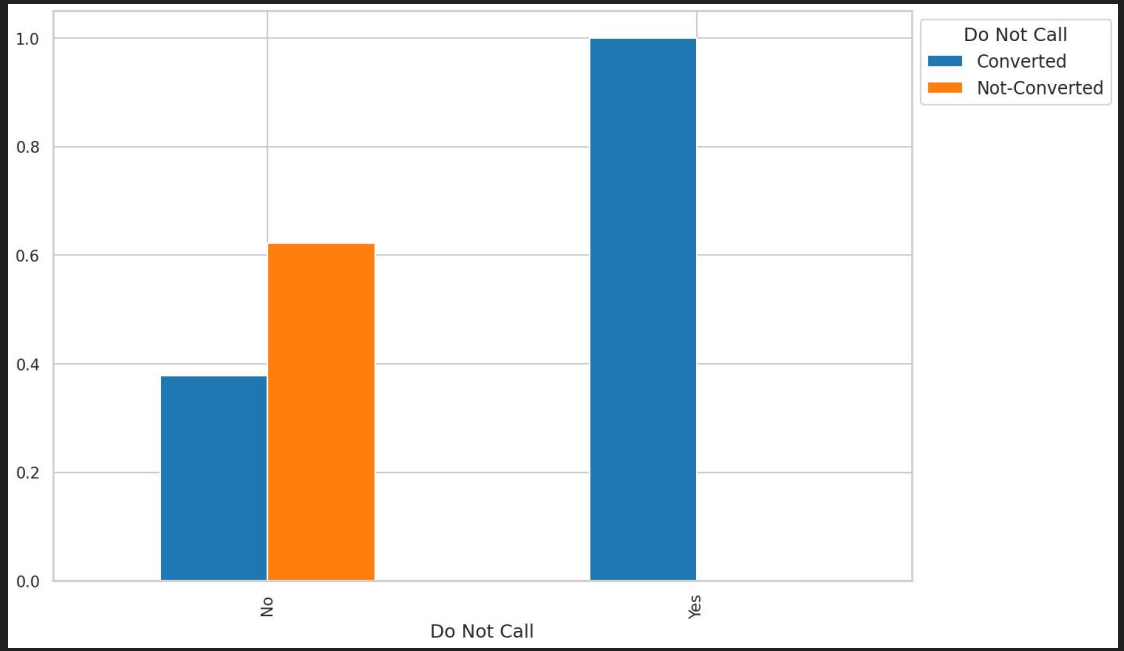
- Higher the time spent on website higher the conversion ratio.
- Page views per visit has no significant impact on conversion rate

Lead Origin



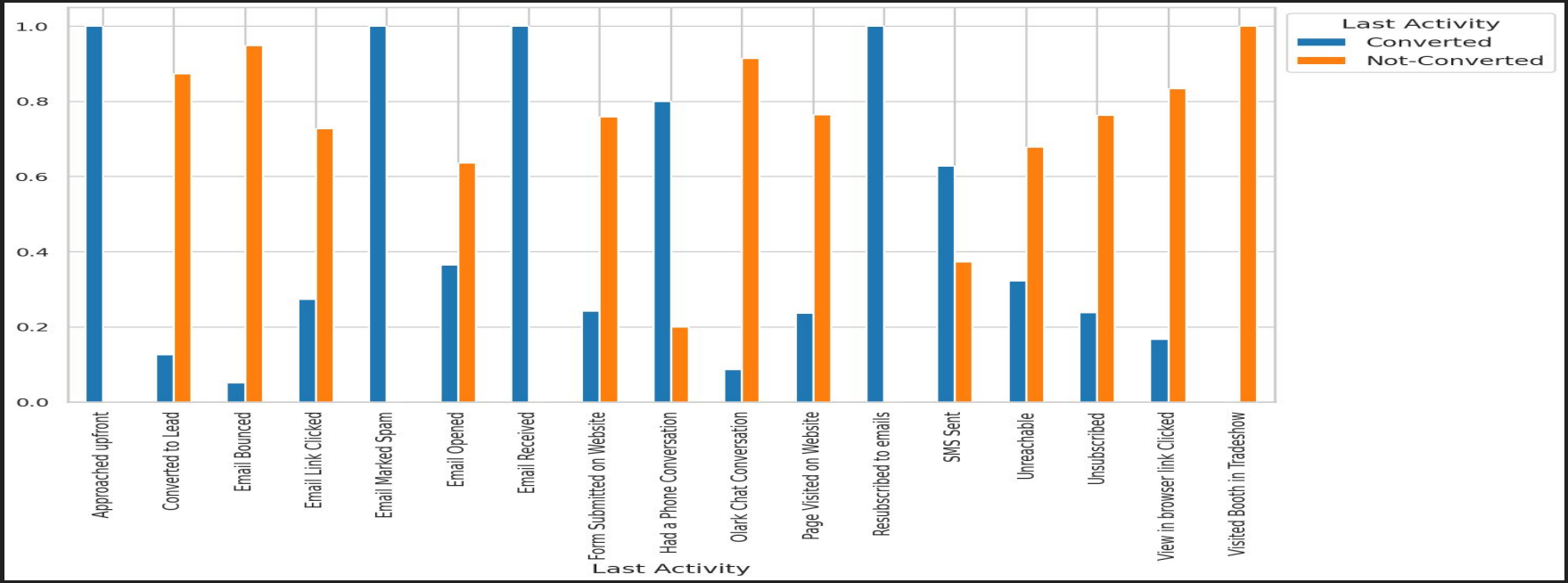
- Lead originating from Add form have a very high conversion rate

Donot Call



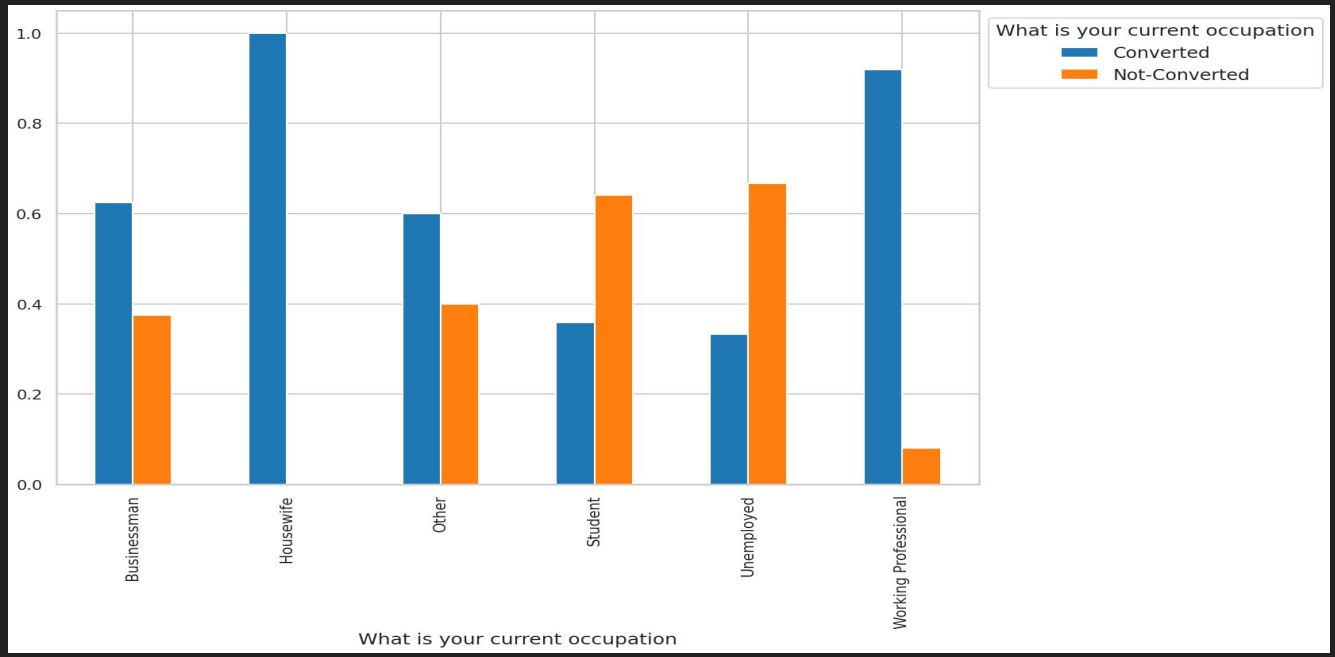
- People who have opted for no calls are more likely to be converted

Last Activity



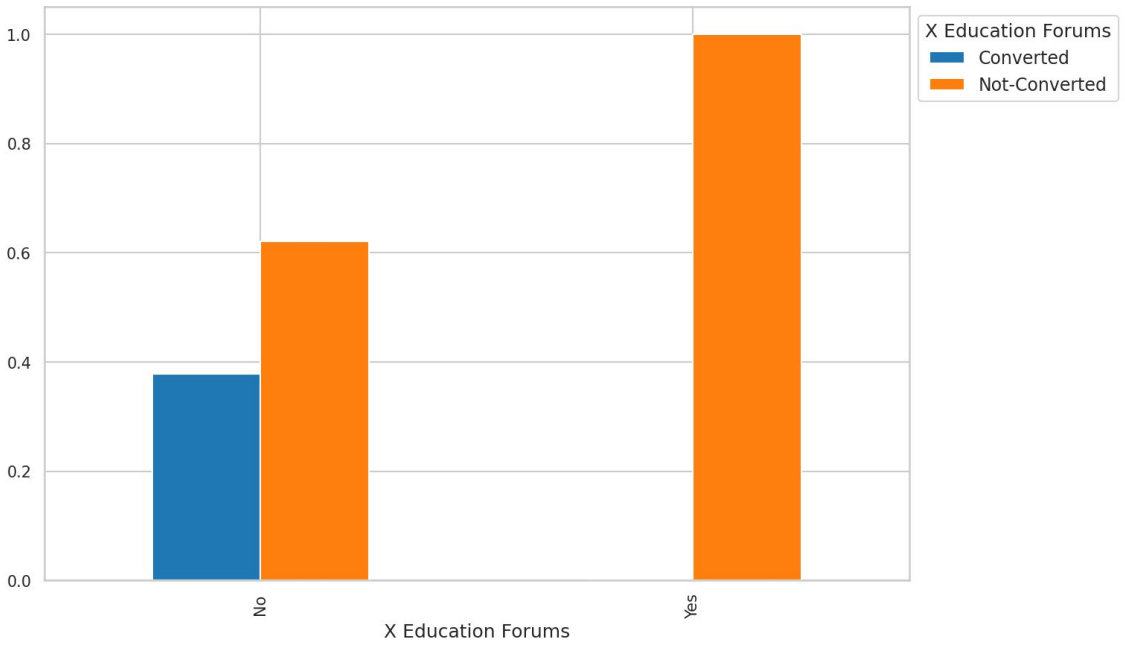
- Last activity like Approached upfront, Email Marked Spam, Email Received and Subscribed to emails have good conversion rate
- Last activity like Email bounced, Olark chat conversation and visited booth in trade show have poor conversion rate

Occupation



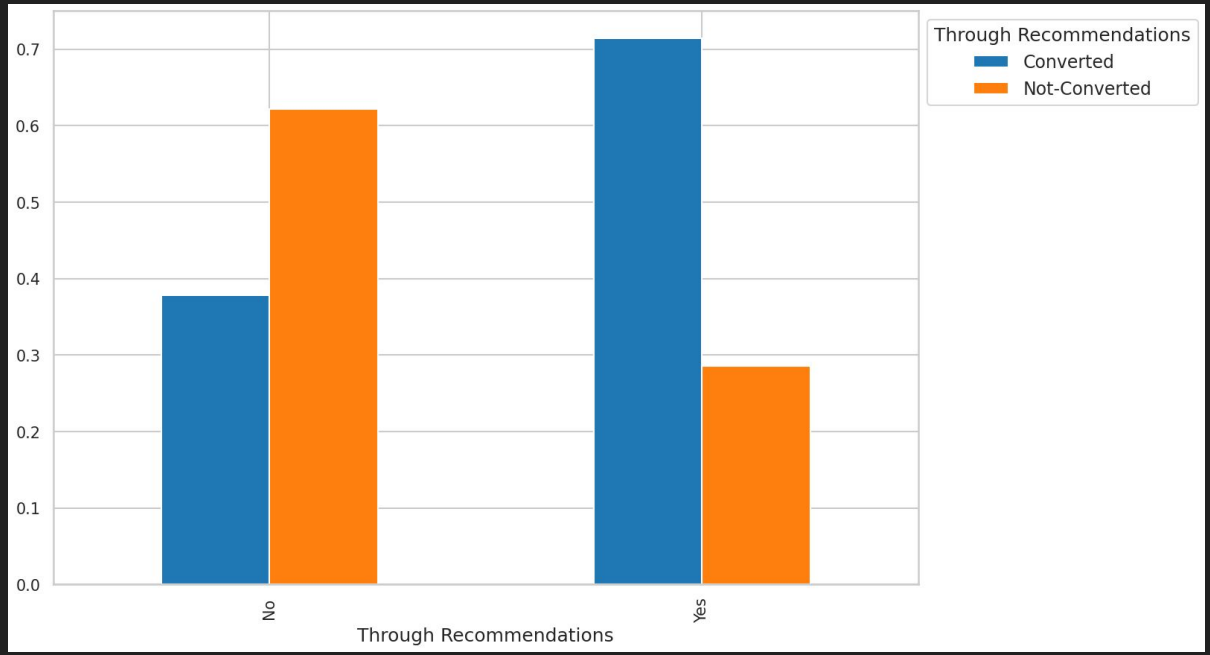
- Occupation like Housewife, working professional has higher conversion rates.

X Education



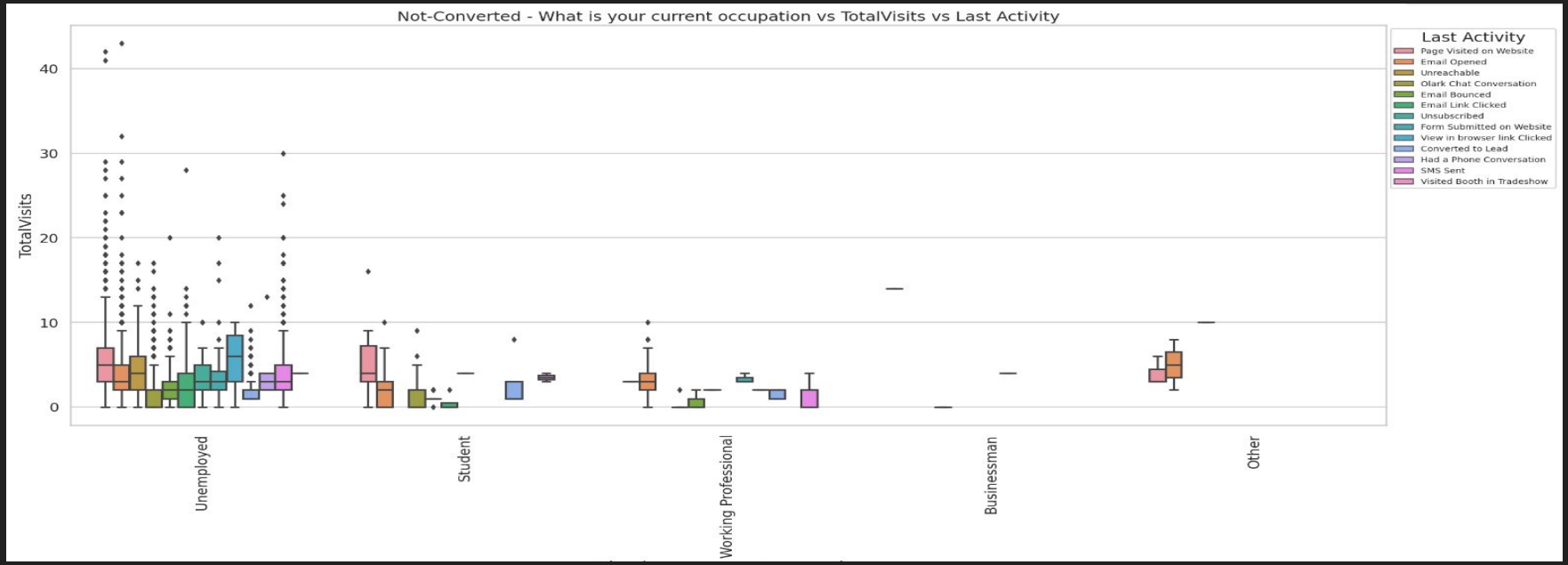
- Customers from X Education forums have poor conversion rate.

Recommendation



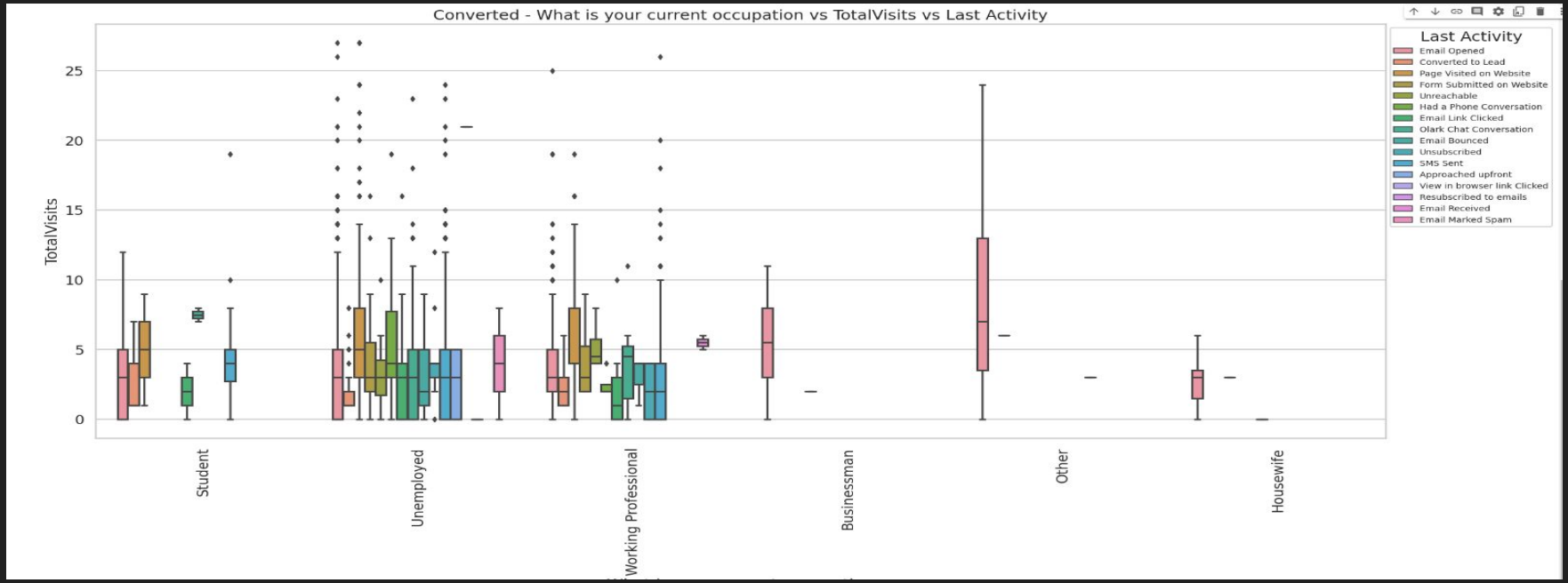
- Customers Through Recommendations are more likely to be converted.

Occupation vs Total visits vs Last activity



- For leads that were not converted we see that house wife has no activity, working professional and others have lower total visits

Occupation vs Total visits vs Last activity



- For leads that were converted we see house wife, businessman and others have opened emails and have higher total visits.

Inferences from Analysis

- After our analysis we found several important feature listed below
  - 'Lead Origin', 'Lead Source', 'Do Not Email', 'Do Not Call', 'Converted', 'TotalVisits', 'TotalTime Spent on Website','Page Views Per Visit', 'Last Activity', 'Specialization', 'What is your current occupation', 'Through Recommendations', 'Tags', 'City', 'Last Notable Activity'
  - Next step would be to create dummies for all the categorical variables and feature scaling of all numerical variables so that we can build our model

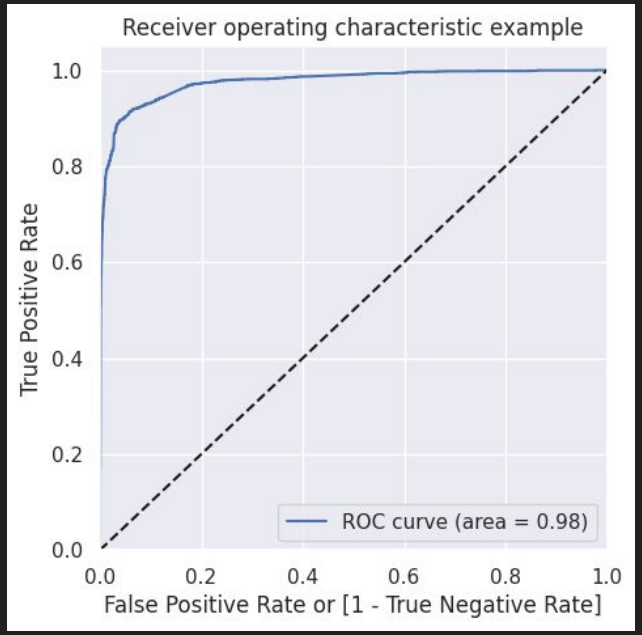
Model Building

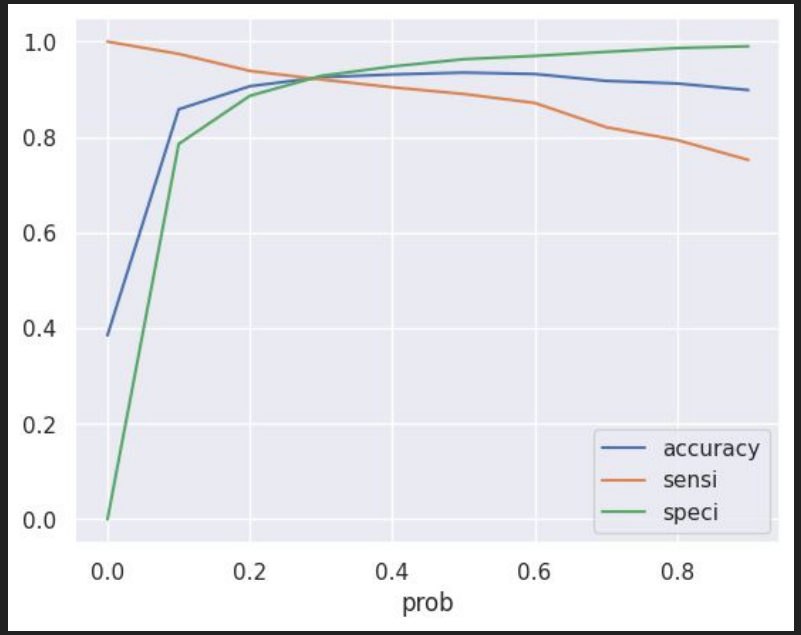
- We need to remove all the irrelevant feature using RFE and restrict the feature count to 20
- After RFE we will do manual feature elimination of column with high p-value and VIF
  - Iteratively drop columns with p-value greater than 0.05
  - Rebuild the model after dropping all such features
  - Check if any feature has VID value greater than 5.
  - Drop such features
- Train the model on Test set

Training Model

- With an arbitrary cut off of 0.5 percent below is the model performance.
- Model accuracy is 93 percent which is pretty good
- Some other metric to decide if the model is performing good
  - Sensitivity / Recall : 0.89
  - Specificity : 0.96
  - False Positive Rate : 0.03
  - Positive Predictive Value / Precision : 0.93
  - Negative predictive value : 0.93
- The model has Sensitivity and Specificity which is greater than 85 percent which means our model is performing good

Training Model





- We can see the ROC curve is curved in the top left corner with area under curve 0.98
- And as per the accuracy, sensitivity and specificity we can take the cutoff of 0.36

Final Features

- Tags_Closed by Horizzon
- Tags_Lost to EINS
- Tags_Will revert after reading the email
- Lead Source_Welingak Website
- Tags_Busy
- Last Activity_SMS Sent
- Tags_Missing
- Total Time Spent on Website
- Do Not Email
- Lead Origin_Landing Page Submission
- Tags_switched off
- Last Notable Activity_Olark Chat Conversation
- Tags_Ringing
- Last Notable Activity_Modified

Analysis Verdict

- The company should reach out “Tags_Closed by Horizzon”, “Tags_Lost to EINS”, “Tags_Will revert after reading the email” and “Lead Source_Welingak” Website as these customers have high conversion rate
- The company should reach out “Tags_Busy”, “Last Activity_SMS Sent” only when the have additional resource to make calls as these feature don’t have very high conversion rate
- The company should focus on customer with high “Total Time Spent on Website” as they have high conversion rate and should make calls to customers with low “Total Time Spent on Website” only when there are ample resources to make calls

- The company should not contact customers “Do Not Email”, “Tags_switched off” and “Tags_Ringing” as they have high probability of being not converted
- The company should not contact customers “Last Notable Activity_Olark Chat Conversation” and “Last Notable Activity_Modified” as they have high probability of being not converted