Lead Scoring Case Study

-Atul Jagtap

Problem Statement

- An education company named X Education sells online courses to industry professionals
- Although X Education gets a lot of leads, its lead conversion rate is very poor around 30%
- The company requires to build a model wherein assigning 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
- The company has a target lead conversion rate to be around 80%

Goals of Case Study

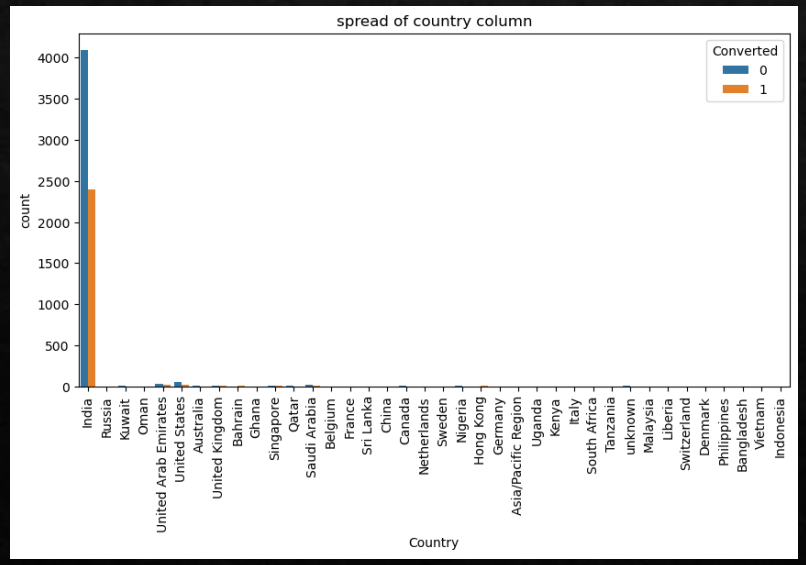
- Build a logistic regression model.
- Assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

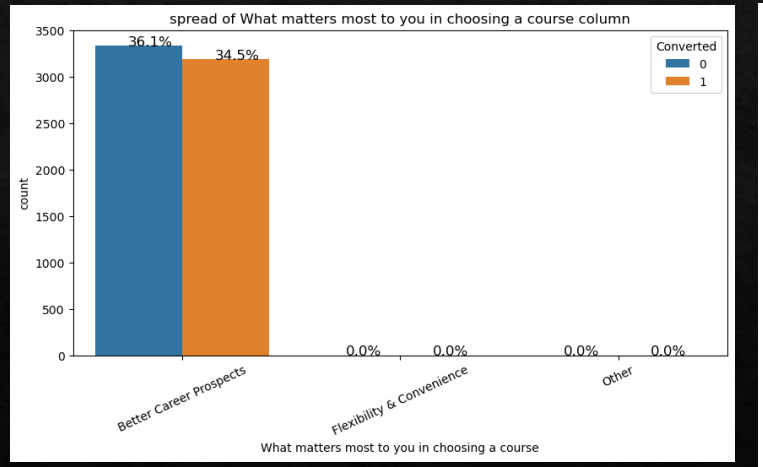
Steps Followed

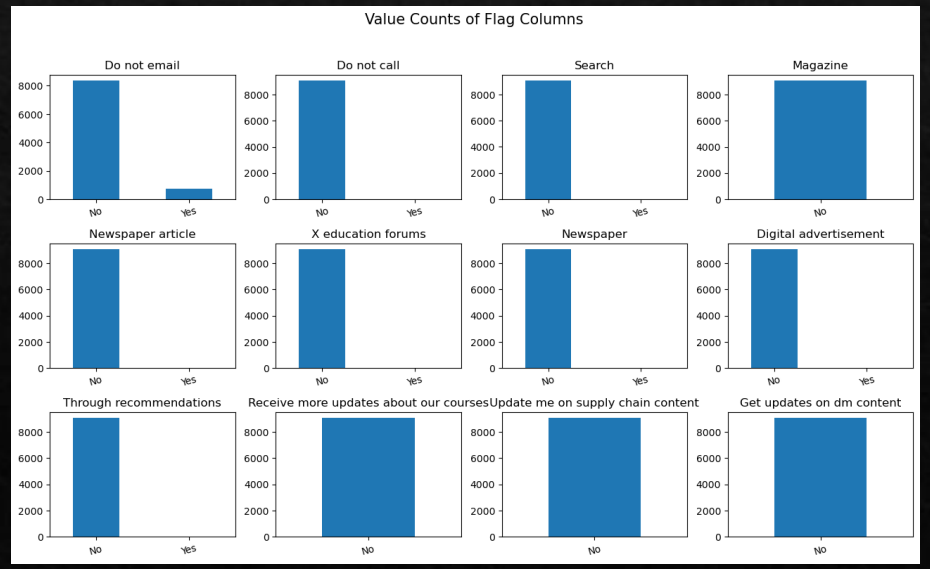
- Data Reading and Understanding
- Data Cleaning
  - Handling select values in the data
  - Handling Null values
  - Handling highly skewed values
- Exploratory Data Analysis
  - Univariate Analysis
  - Bivariate Analysis
- Data Preparation
  - Creating dummy variables of categorical columns
  - Scaling numerical columns
  - Splitting data into train and test sets
- Model Building and Model Metrics Validation
- Final Conclusions and Recommendations

Data Cleaning

- ‘Select’ values are replaced by null values as the level 'Select Specialization' means the customer had not selected this option while filling the form.
- The columns which have >30% of missing values are dropped.
- If missing values are <30%, data imputation by other category or mode of that categorical variable has been performed.
- Some of the columns which have highly skewed data, i.e. above 90% of the data in one category, are dropped.
- Some columns which are irrelevant from the perspective of the business objective are dropped.



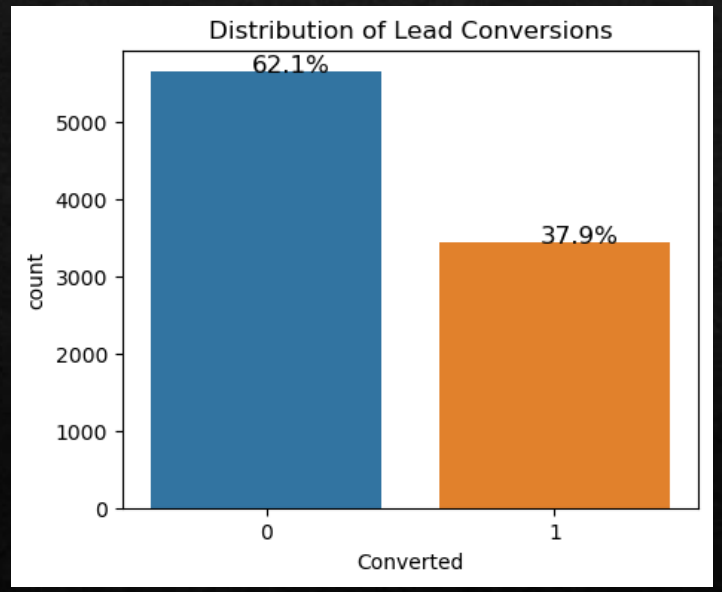




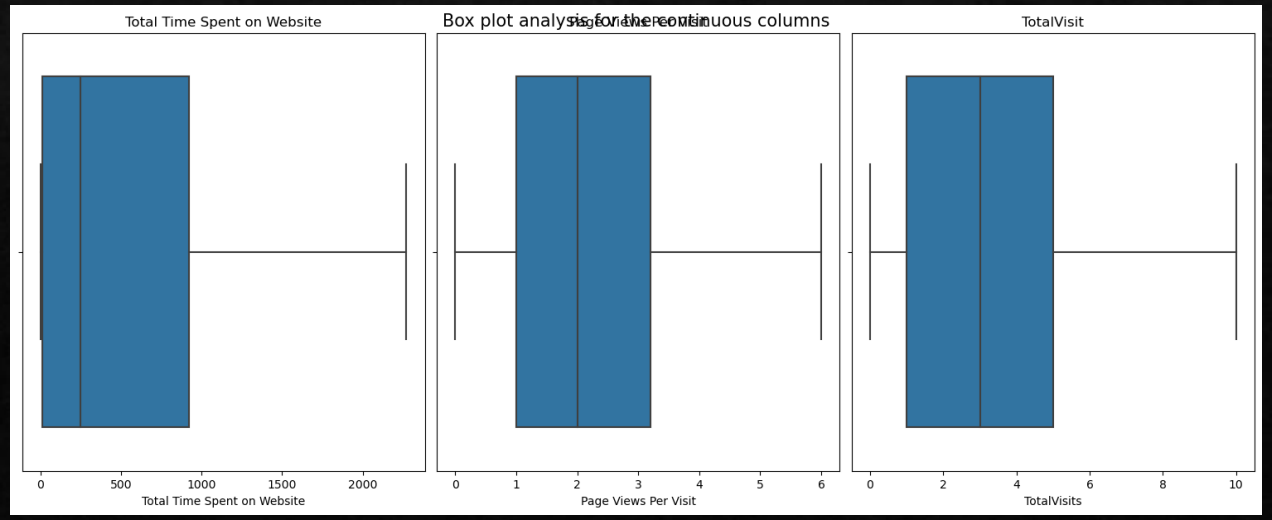
These columns are highly skewed and having one major value. Including these columns won’t be helpful for analysis

Exploratory Data Analysis

- Lead conversion rate at X Education company is around 38%.

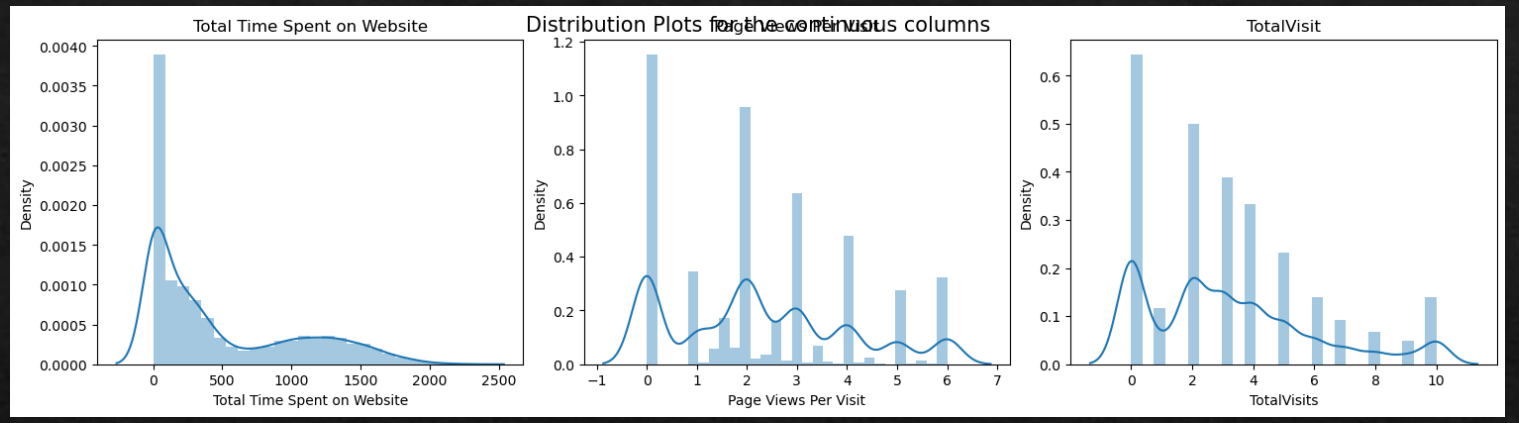


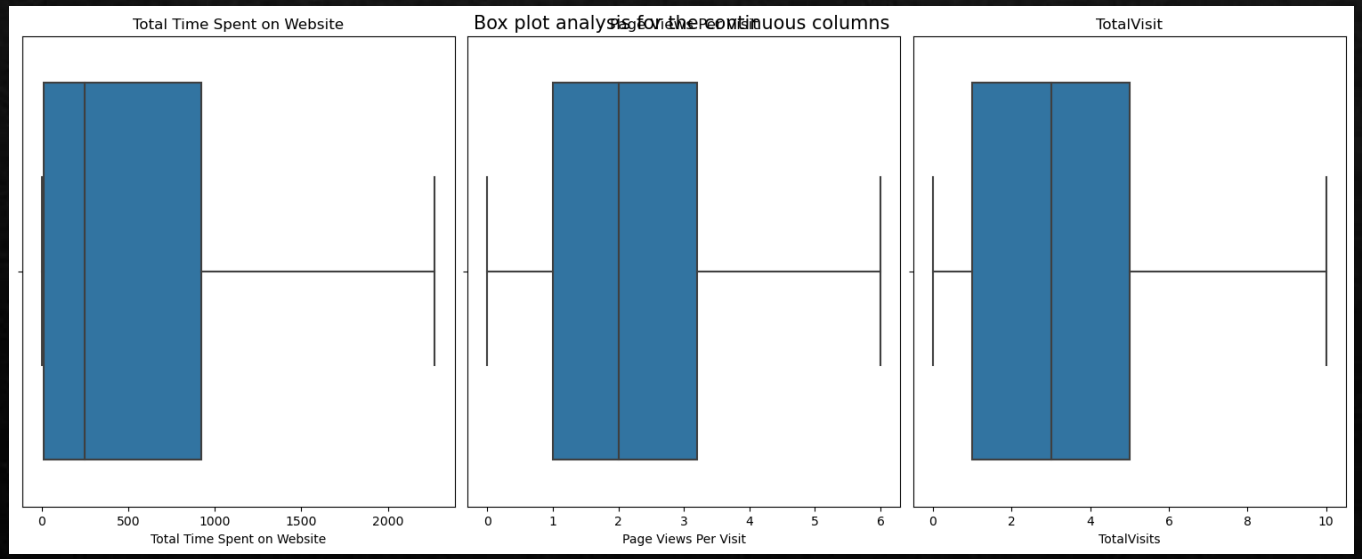
- For the students who have converted to hot leads have higher median value overall in 'Total Time pent on website' than those who are not converted
- That means those who are spending more time on websites are likely to get converted to Hot leads So website can be more attractive to increase the conversion rate



Exploratory Data Analysis

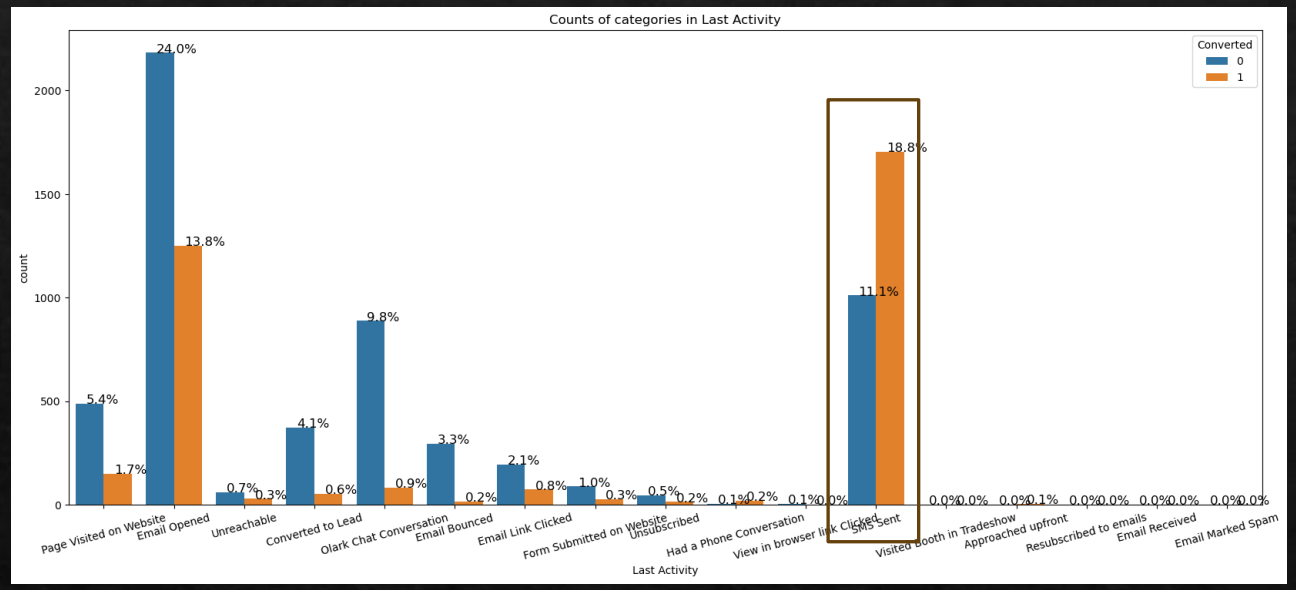
- Distribution plot gives overall idea about the distribution of Total time spent, Page Views per Visit and Total Visit
- Although, the distribution is not normal and we have to scale the data afterwards, we can see the higher values tends to appear less for all the 3 columns and there are local maxima and minima values till some range





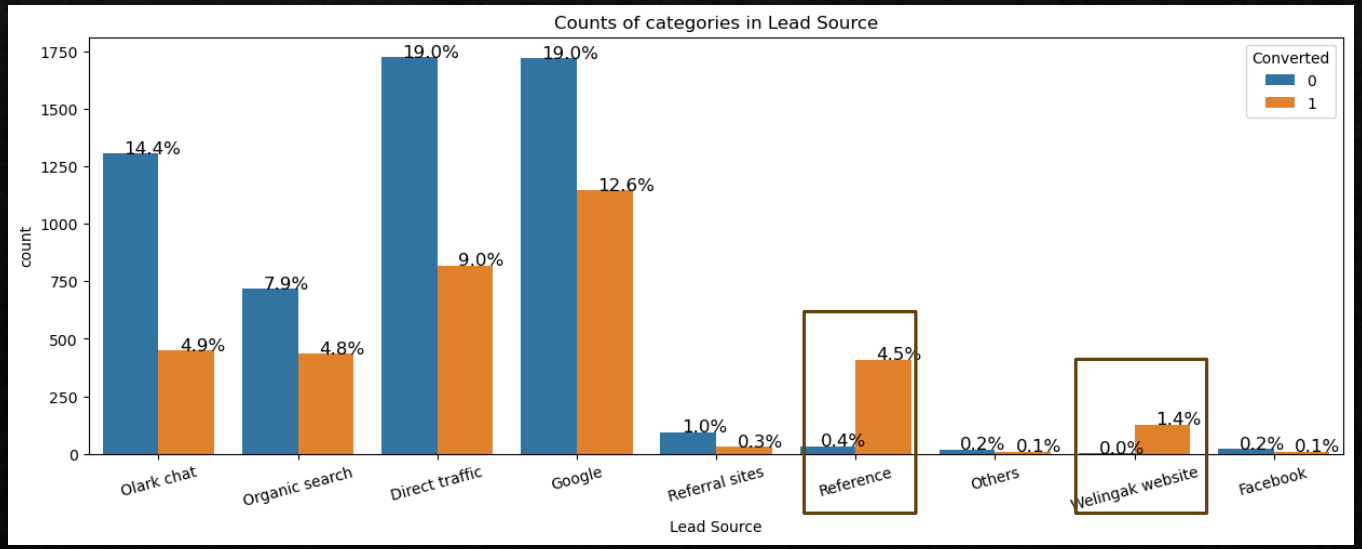
- But for 'Page Views Per Visit' and Total Visits there is not much significant difference between the leads who are converted and not converted

Exploratory Data Analysis

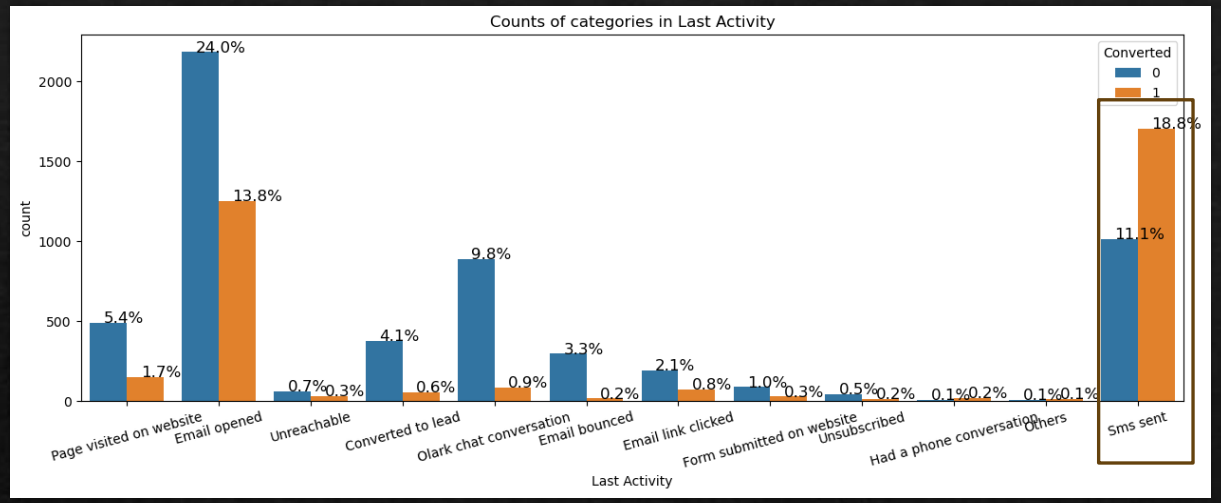


Those who have Last Notable Activity as SMS sent, have higher conversion of hot leads.
Keeping track of this one would e beneficial.

Whichever prospective leads are using Welingak website and are referred have successfully converted.
From business aspect these two lead sources need to be look after and number of leads should be increased here to increase the conversion rate

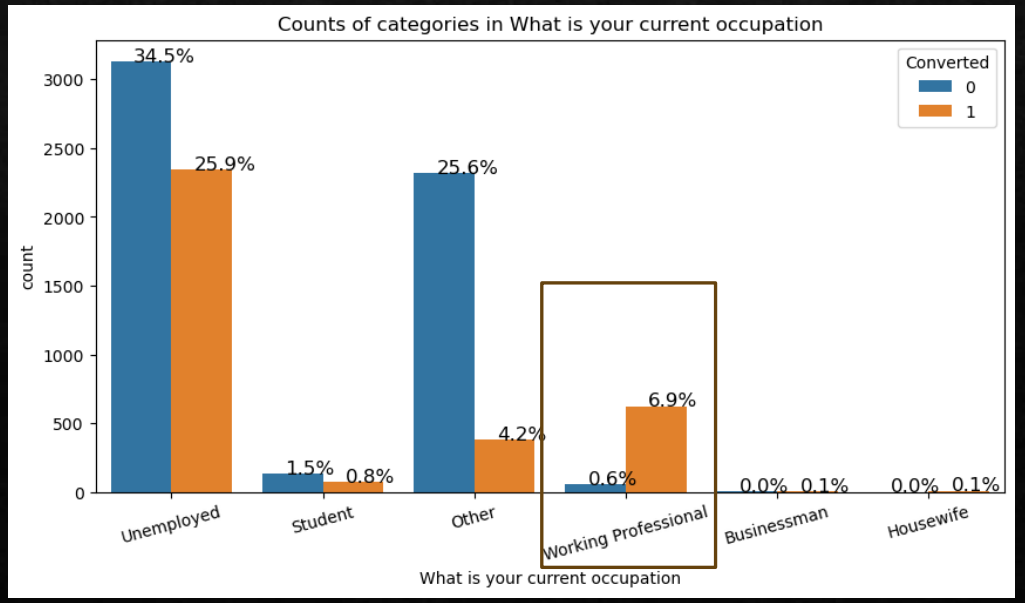


Exploratory Data Analysis

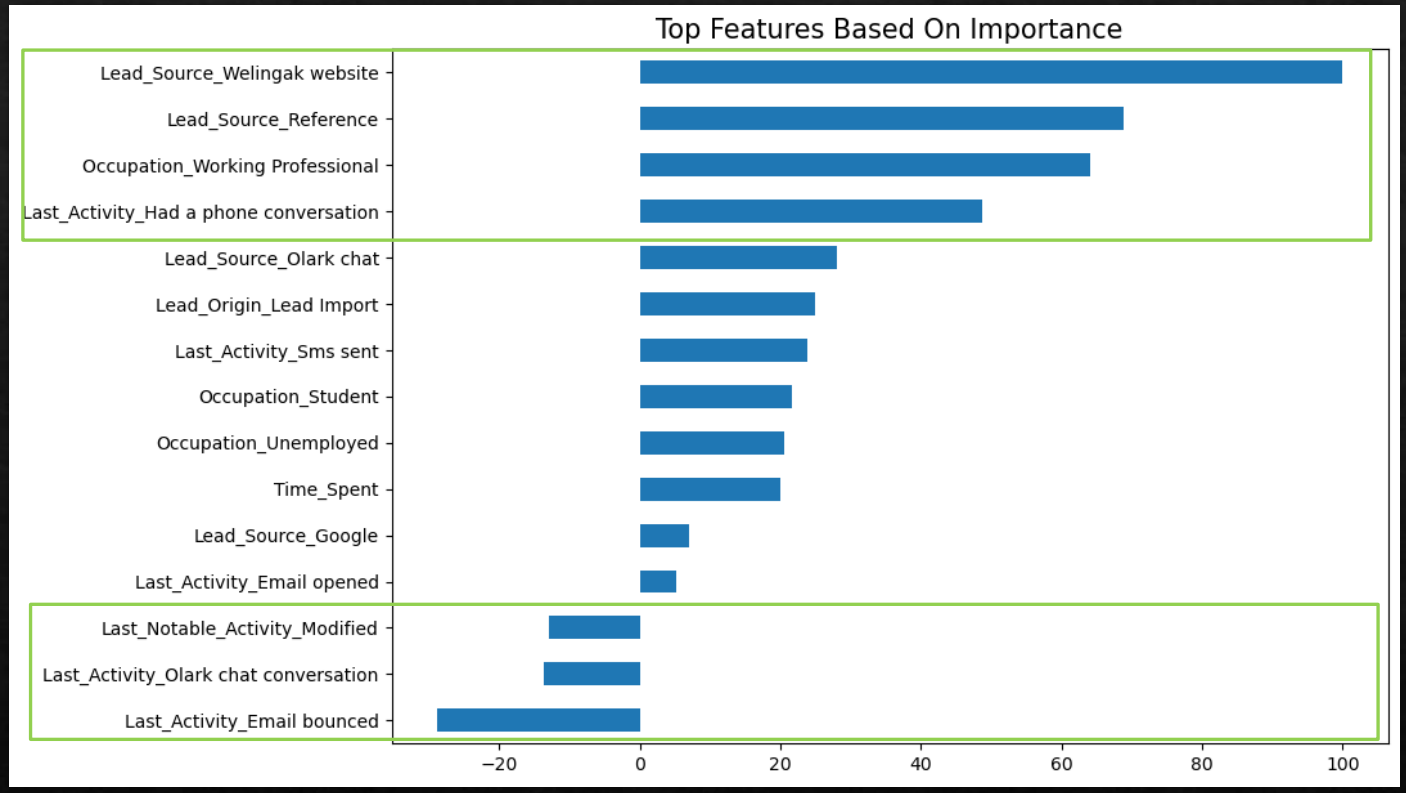


Those who have Last Activity as SMS sent, have higher conversion of hot leads.
Those who have opened Email or started Olark Chat Conversion has lesser conversion of hot leads.

Those who are unemployed have almost balanced conversion of leads into hot leads.
Those who are working professionals, have higher percentage of Conversion into Hot Leads. More focus should be given on this.



Model Building



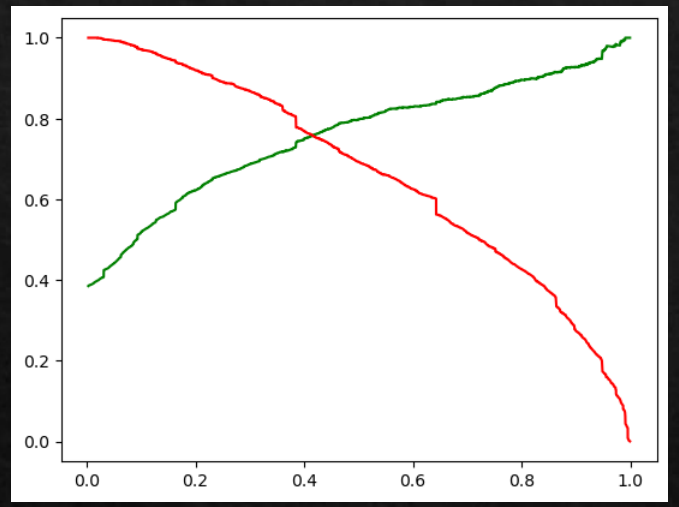
TOP FACTORS THAT IMPACT THE CONVERSION OF LEADS

Model Building

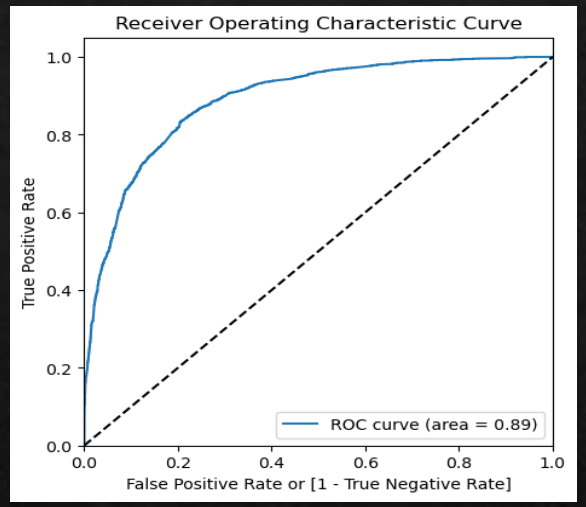
- Top 3 indicators that a lead will get converted to a hot lead:
  - Lead_Source_Welingak website: A lead sourced from Welingak Website needs to be targeted and guided properly to increase the number of Hot Leads. Website should also be made more attractive to increase the number of leads.
  - Lead_Source_Reference: A lead who referenced by another hot lead is more likely to get converted and so needs to be guided properly to increase the conversion rate.
  - Occupation_Working Professional: Working professionals are more likely to get converted.

- Major indicators that a lead will NOT get converted to a hot lead:
  - Last_Activity_Email bounced
  - Last_Activity_Olark chat conversation.
  - Last_Notable_Activity_Modified

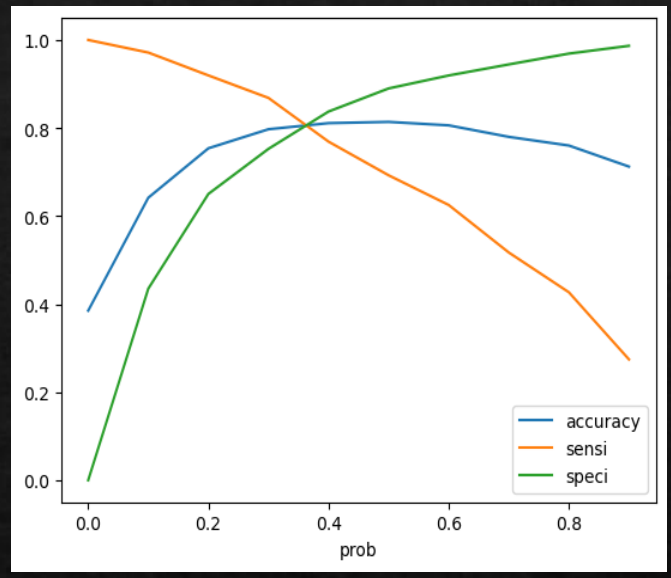
Model Summary and Performance



0.38 is also the approximate tradeoff between Precision and Recall



ROC curve showing AUC 0 89



We used Sensitivity/Specificity trade off graph to observe the optimal threshold value

	Model Accuracy	Recall/ Sensitivity	Precision	Specificity
Train	80.9%	80.7%	72.7%	81%
Test	80.7%	80.7%	70.4%	81%

~81% Sensitivity value indicates that our model is able to predict ~81% of actual conversion cases correctly

Conclusion and Recommendations

- To Achieve a conversion rate of around 80%
  - Company should not miss those prospect leads which can turn into Hot Leads which is Recall.
  - Should not overestimate a cold lead which is precision.
- With this model, we achieve a **trade off point of 0.38** . That means 38% probability of a prospect Lead is good enough to target it as HOT LEAD