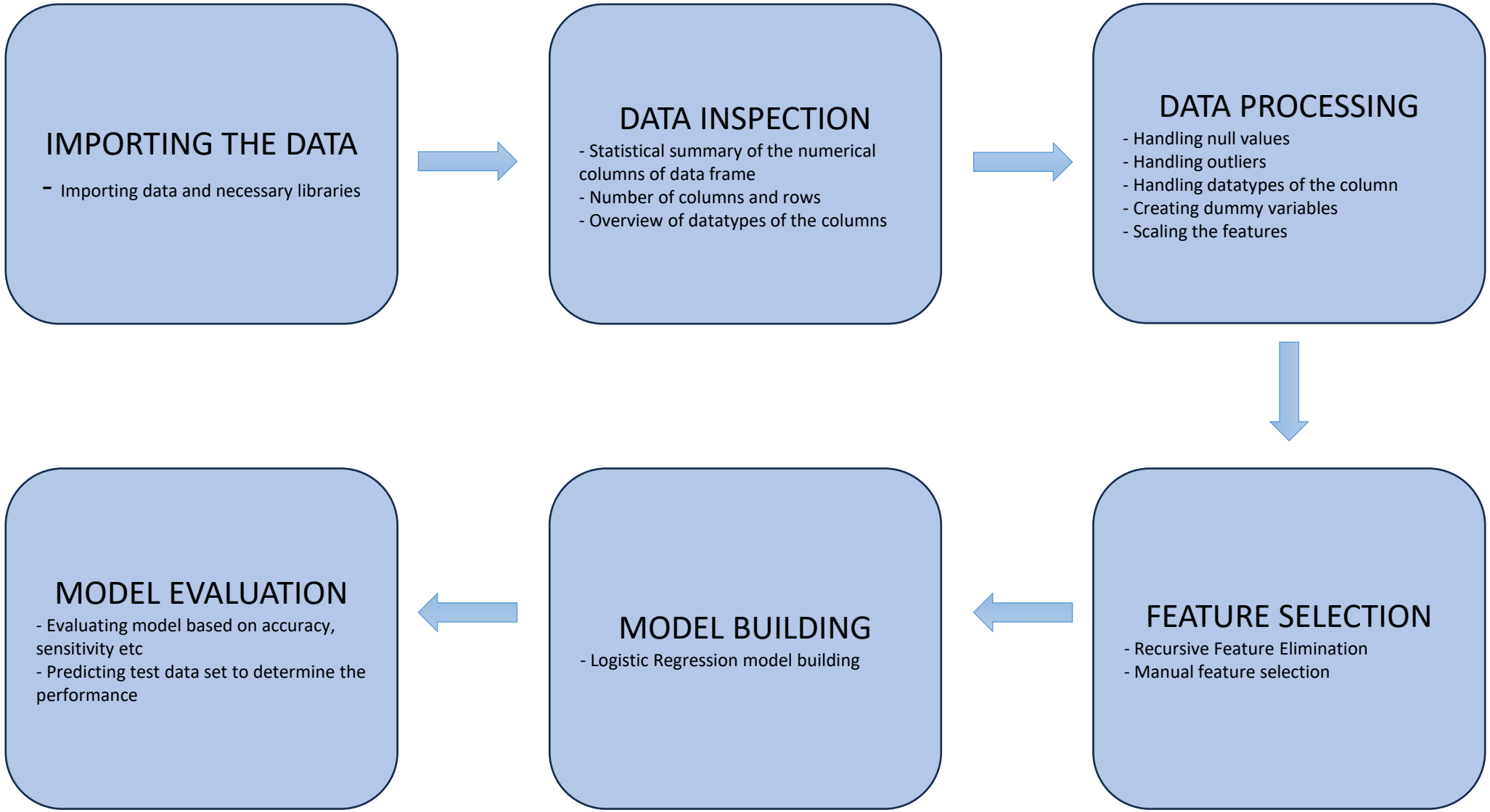
LEAD SCORING CASE STUDY

- CHETAL PATEL
- RAGUL
- RAGHAV MAHESHWARI

PROBLEM STATEMENT

X Education wants to identify the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a machine learning model wherein we need to 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.

SYSTEM DESIGN



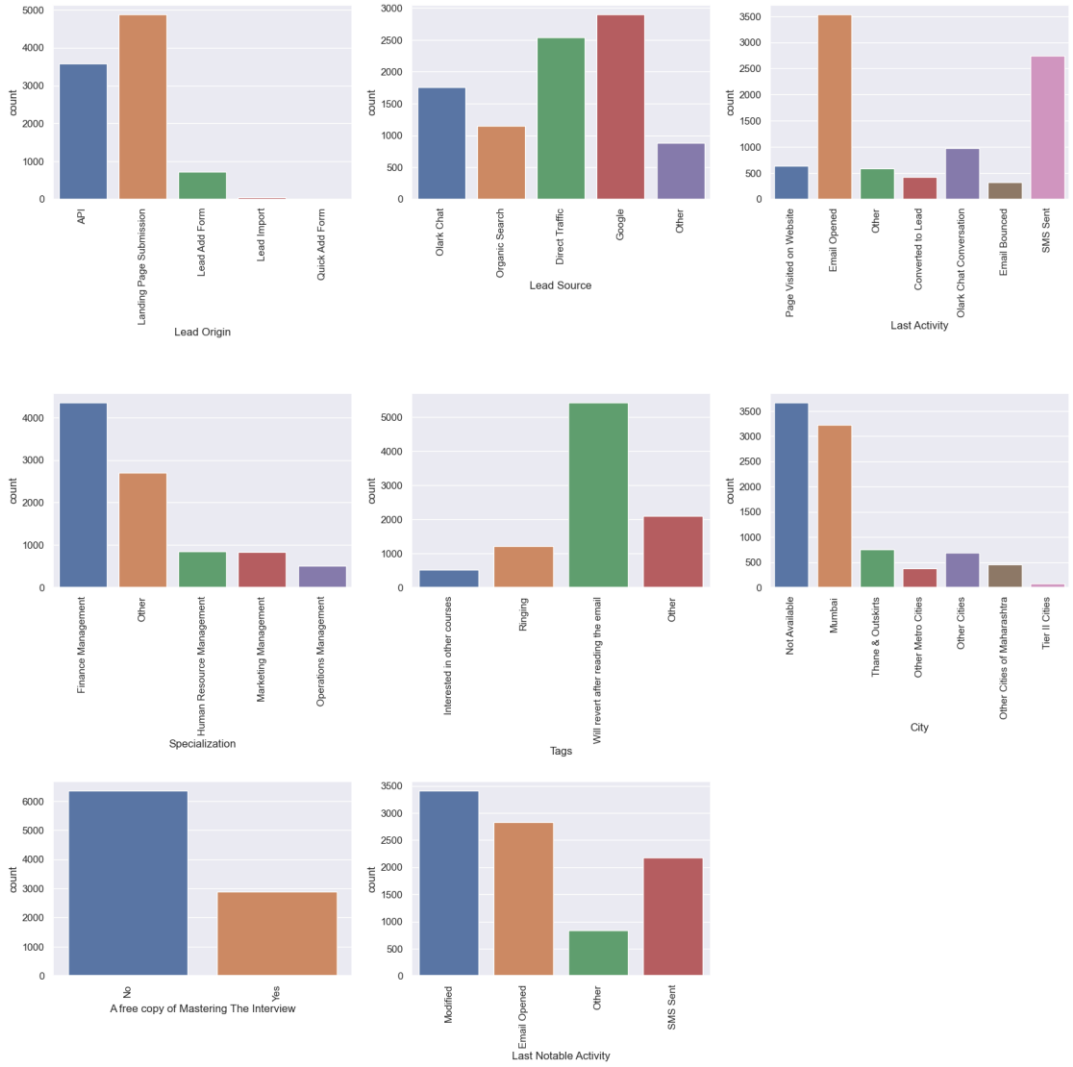
APPROACH AND METHODOLOGY

- Upon importing the necessary libraries, proceeded with data inspection where we could see that dataset contained data points of 9240 rows and 37 columns.
- Completing the initial data inspection, proceeded with data cleaning or data preprocessing which involved dropping columns with single unique value, dropping column with null values greater than 40%.
- For the column with null values less than 40%, imputed the null values with mode of the variable if it's a categorical column and imputed median for null values in case of numerical column as presence of outliers has been detected. Eg: Categorical column: Country imputed with mode of variable: India, whereas numerical column: TotalVisits imputed with median of the variable.
- Once the null values in each columns has been handled, proceeded with analysing categorical columns by plotting count plot to understand the data distribution from which we observed that there are columns which are highly skewed where one value is very high and other values in column are almost negligible. Hence, dropping those columns.
- Further proceeded wit Exploratory Data Analysis which involved univariate analysis on the categorical columns, univariate analysis on numerical columns trough distribution plots, bivariate analysis on categorical columns.
- Upon completing EDA, proceeded with data preparation and feature selection, which involved creation of dummy variables for columns: 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'Tags' & 'City'.
- Further proceeded with train test split with training set size of 70% of master dataset and 30% for test set. Continued with scaling the features with MinMaxScaler for columns: 'TotalVisits', 'Total Time Spent on Website', 'Page Views Per Visit'.
- To check if any strong correlation exists between the variables, created a heatmap with correlation data frame of training set.
- Continued with feature elimination through recursive feature elimination to extract the top 10 variables.

APPROACH AND METHODOLOGY

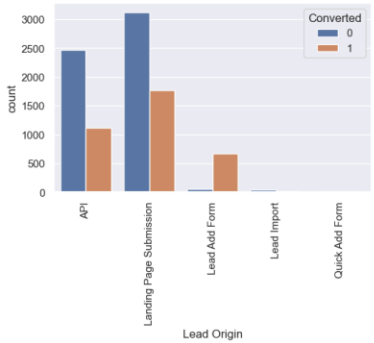
- With the extracted variables, proceeded with logistic regression model building through GLM (Generalized Linear Model). Went on to create new models by dropping few columns as p-value for few columns exceeded 5% limit.
- Created a data frame for Variance Inflation Factor to check the collinearity between the variables.
- After getting the final model, created a confusion matrix to evaluate the model based on the overall accuracy, specificity, sensitivity, positive predictive value, negative predictive value.
- Created a data frame for Variance Inflation Factor to check the collinearity between the variables.
- After getting the final model, created a confusion matrix to evaluate the model based on the overall accuracy, specificity, sensitivity, positive predictive value, negative predictive value.
- Proceeded with plotting a ROC curve between false positive rate and true positive rate.
- To determine the optimal cutoff, created a data frame with probability ranging from 0.1 to 0.9 and values of accuracy, sensitivity and specificity in columns.
- By plotting the curve for the above data frame, determined a optimal cutoff for probability to predict the conversion rate
- Calculated the precision and recall for the model created. Upon which proceeded with plotting precision recall curve to find optimum cutoff.
- Based on the optimal cutoff, made predictions on the test data set.
- Finally created a data frame with lead number, probability, actual converted column, predicted converted column and lead score based on the probability.

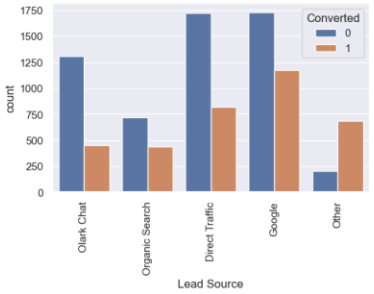
GRAPHS & INSIGHTS

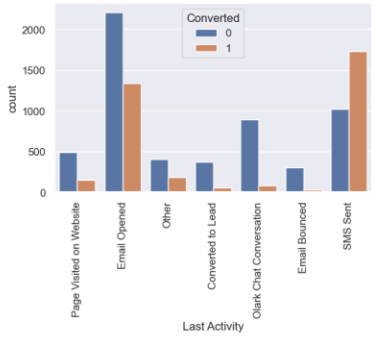


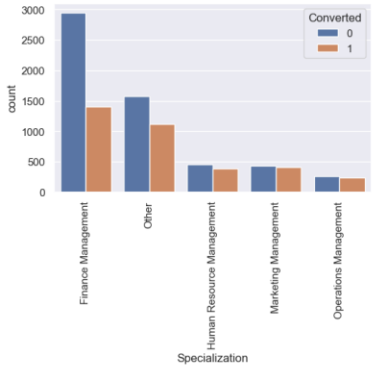
1. Lead Origin - Landing page submission and API are the main lead origins.
2. Lead Source - Direct Traffic and Google are the main lead sources.
3. Last Activity - Number of emails opened and SMS sent has more number of last activity.
4. Specialization - Most of the people choose Finance management as specialization.
5. Tags - Will revert after reading the email is the highest among the column tag.
6. City - There were many null values but Mumbai city has maximum value.
7. A free copy of mastering the interview - Majority of people choose "No".
8. Last Notable Activity - Modified are high in numbers followed by Email Opened.

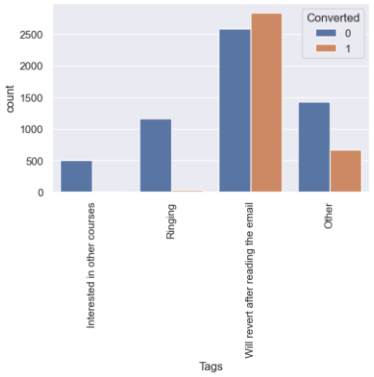
GRAPHS & INSIGHTS

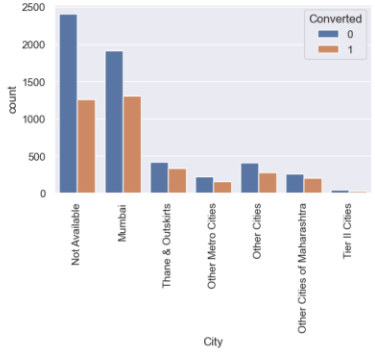


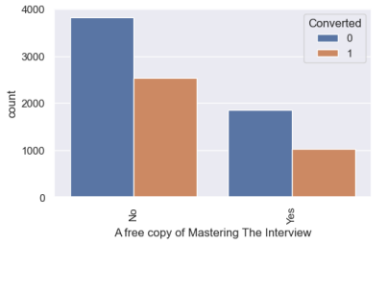


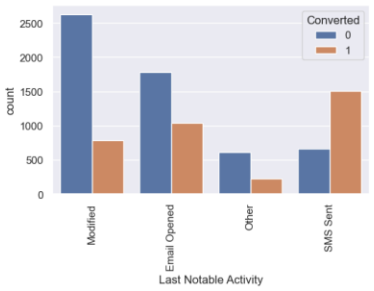






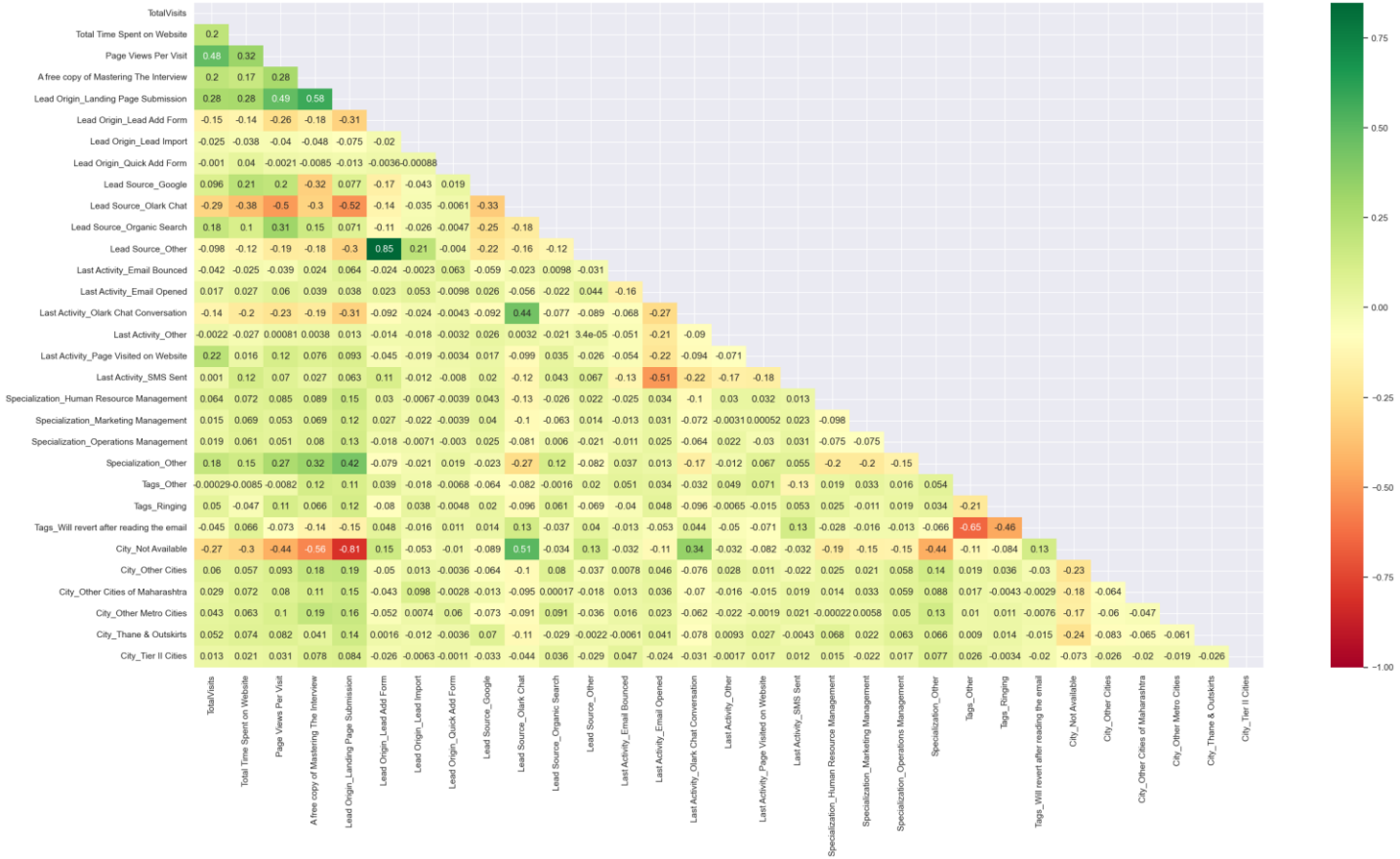






- Lead origin - API and Landing page submission has high hot leads
- Lead Source - The number of Hot leads is higher in Direct Traffic and Google as compared to other categories.
- Specialization - Most of the leads comes from Finance management but here Hot leads are lesser than Cold leads.
- Last Activity - The number of hot leads are higher in SMS Sent. In Email Opened, cold leads are higher than hot leads.
- Last Notable Activity - Same behavior as Last Activity column
- City - The response from Mumbai is high. It has maximum hot leads

GRAPHS & INSIGHTS



Correlation heatmap for the variables in training data set. We could see a strong correlation between Lead Origin_Other and Lead Origin_Lead Add Form & good correlation between A free copy of Mastering The Interview and Lead Origin_Landing Page Submission

MODEL BUILDING & FINDINGS

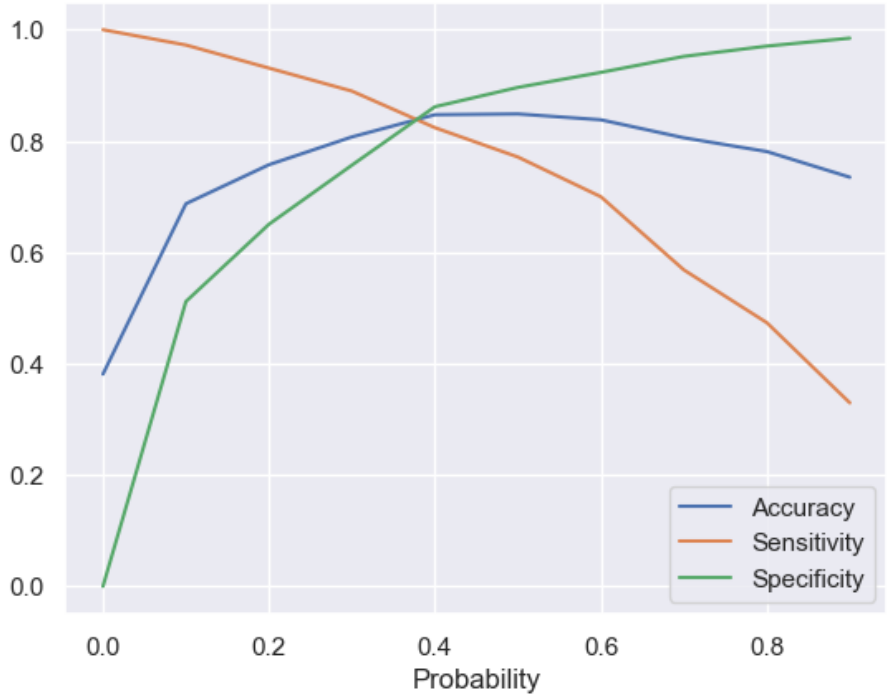
	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.8325	0.274	-21.317	0.000	-6.369	-5.296
TotalVisits	7.2324	1.666	4.341	0.000	3.967	10.498
Total Time Spent on Website	4.8553	0.185	26.290	0.000	4.493	5.217
Lead Origin_Landing Page Submission	-0.8138	0.141	-5.752	0.000	-1.091	-0.536
Lead Origin_Lead Add Form	4.2462	0.229	18.534	0.000	3.797	4.695
Lead Source_Olark Chat	0.9853	0.129	7.621	0.000	0.732	1.239
Last Activity_Email Bounced	-1.4421	0.302	-4.773	0.000	-2.034	-0.850
Last Activity_Email Opened	0.5894	0.104	5.674	0.000	0.386	0.793
Last Activity_Olark Chat Conversation	-1.2899	0.191	-6.749	0.000	-1.665	-0.915
Last Activity_SMS Sent	1.8530	0.111	16.679	0.000	1.635	2.071
Tags_Other	3.4606	0.216	16.021	0.000	3.037	3.884
Tags_Will revert after reading the email	4.4165	0.210	21.057	0.000	4.005	4.828
City_Not Available	-0.9151	0.136	-6.710	0.000	-1.182	-0.648

Summary of logistic regression final model:-

Variables included in the model: 'TotalVisits', 'Total Time Spent on Website', 'Lead Origin_Landing Page Submission', 'Lead Origin_Lead Add Form', 'Lead Source_Olark Chat', 'Last Activity_Email Bounced', 'Last Activity_Email Opened', 'Last Activity_Olark Chat Conversation', 'Last Activity_SMS Sent', 'Tags_Other', 'Tags_Will revert after reading the email', 'City_Not Available'.

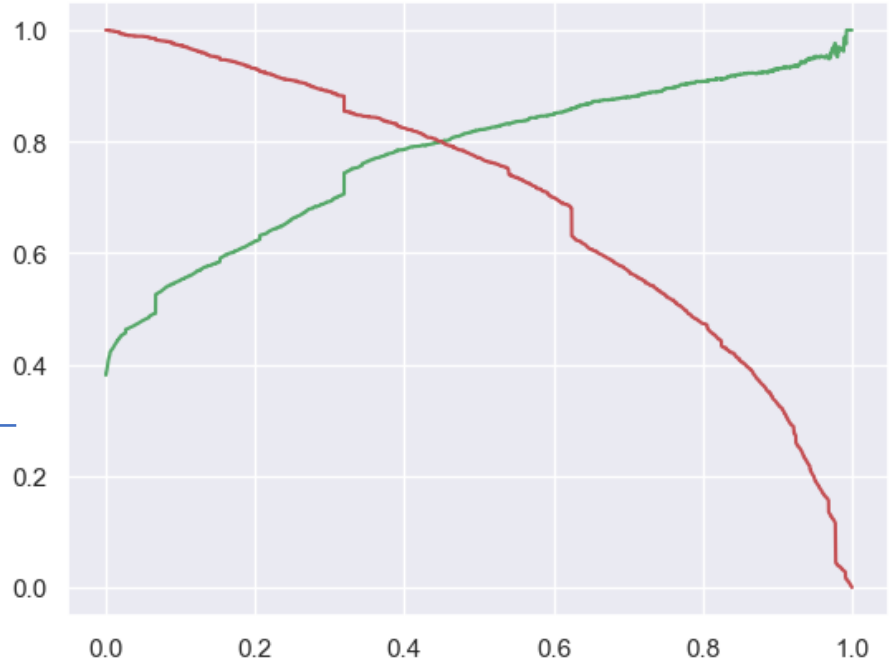
Could observe p-value for the variables are stands at 0% which states coefficients of the variables are statistically significant.

MODEL BUILDING & FINDINGS



Finding Optimal Cutoff point:

- Optimal cutoff probability is the probability point where we get balanced sensitivity and specificity.
- From the above curve, we could observe **0.38** is the optimum point to take it as a cutoff probability.



Precision Recall Curve:

The precision-recall curve shows the tradeoff between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

MODEL BUILDING & FINDINGS

Calculation of accuracy, sensitivity, specificity with probability cut-off: **0.38**

Accuracy: 84.66%
Sensitivity: 83.49%
Specificity: 85.338%
Positive Predictive Value: 77.87%
Negative Predictive Value: 89.35%

	Converted	Lead Number	Converted_Prob	final_predicted
0	1	639211	0.866962	1
1	1	590711	0.950284	1
2	1	644174	0.653498	1
3	0	641392	0.012892	0
4	1	587955	0.910468	1

	Converted	Lead Number	Converted_Prob	final_predicted	Lead Score
0	1	639211	0.866962	1	86.70
1	1	590711	0.950284	1	95.03
2	1	644174	0.653498	1	65.35
3	0	641392	0.012892	0	1.29
4	1	587955	0.910468	1	91.05

Lead Score has been assigned to the leads based on the probability.

CONCLUSION / RECOMMENDATION

- As per the logistic model, we could observe that total visits play a major role in converting the lead to hot leads. Hence, it is advised to improve the visit count of the leads for positive impact.
- Similarly, more time the lead spends on website, the more possibility that it could get converted into a hot lead.
- Lead origin from the lead add form should be focussed as it has good impact on conversion.
- Email bounced from the last activity of the customer has negative implication on the conversion and requires attention.
- Leads who does not mention cities are have low chances of getting converted to hot leads.
- Lead number has been assigned to the leads based on the probability which could be used as an indicator for company to target potential leads.