

LEAD SCORING CASE STUDY

By

Nikhil Jojen & Manas Shirsat

Problem & Approach

- X- Education Business is to sell online course to Industrial Professionals
- Numerous leads are generated, still the lead conversion rate is insignificant.
- Company wants to identify the most potential leads
- This will narrow the leads and help the sales team focus on potential leads

Our Goals for the Case Study:

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

In the future, the company's requirements might change, so you will have to adjust.

Methodology

- Data Cleaning
- 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 Loading & Cleaning

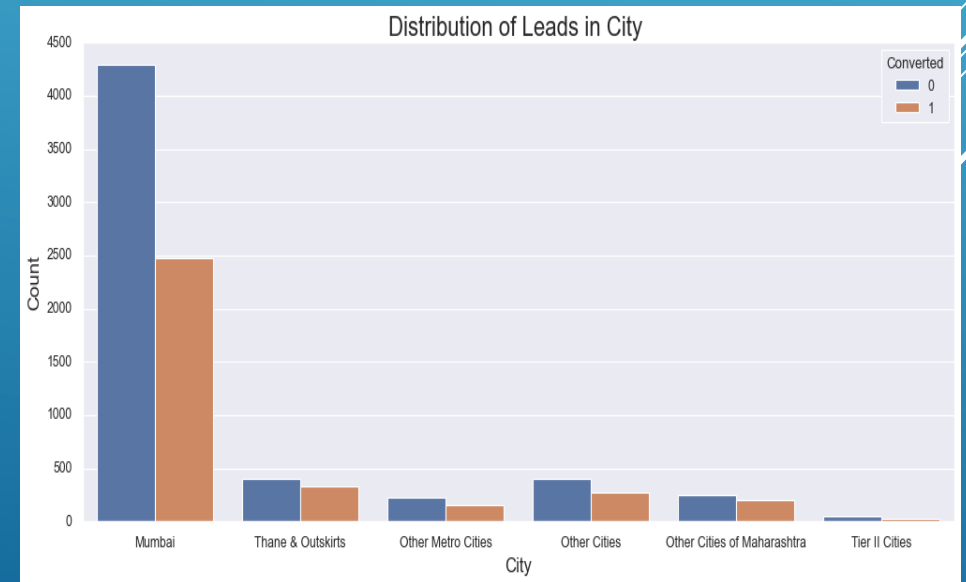
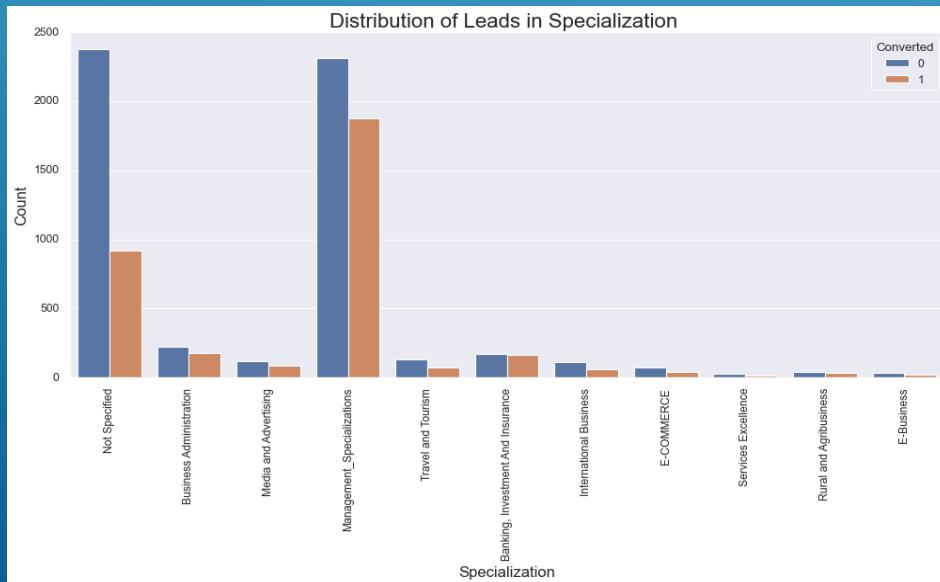
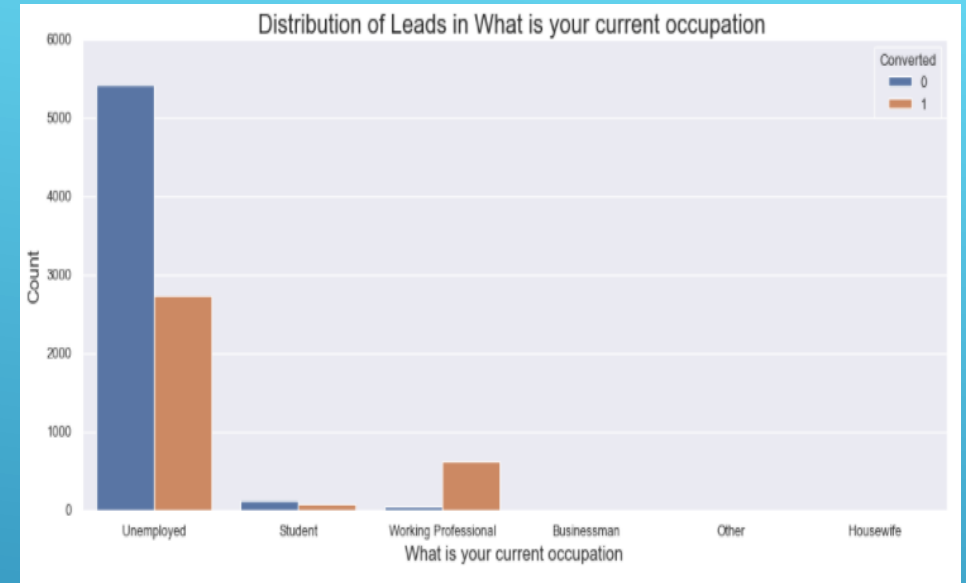
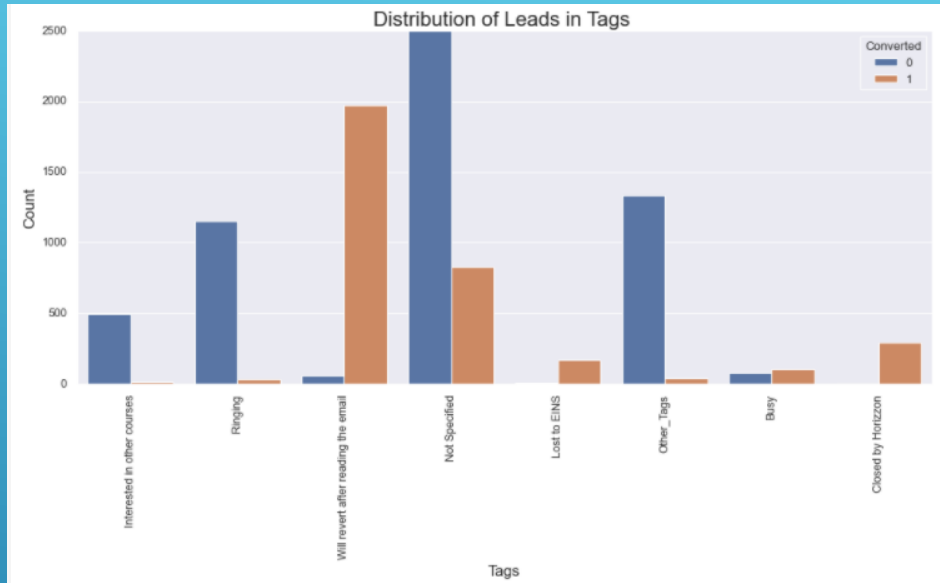
Initial Data : 37 Columns 9240 Rows

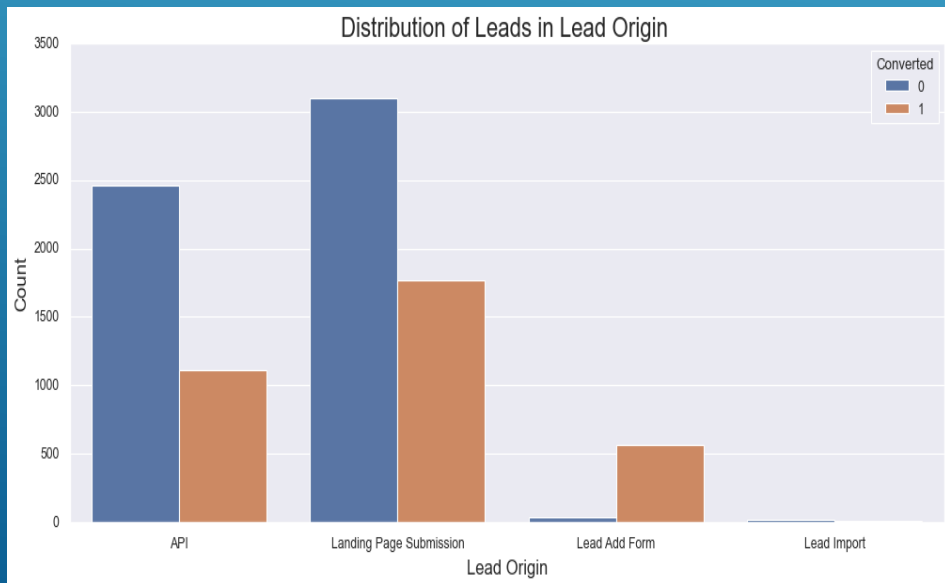
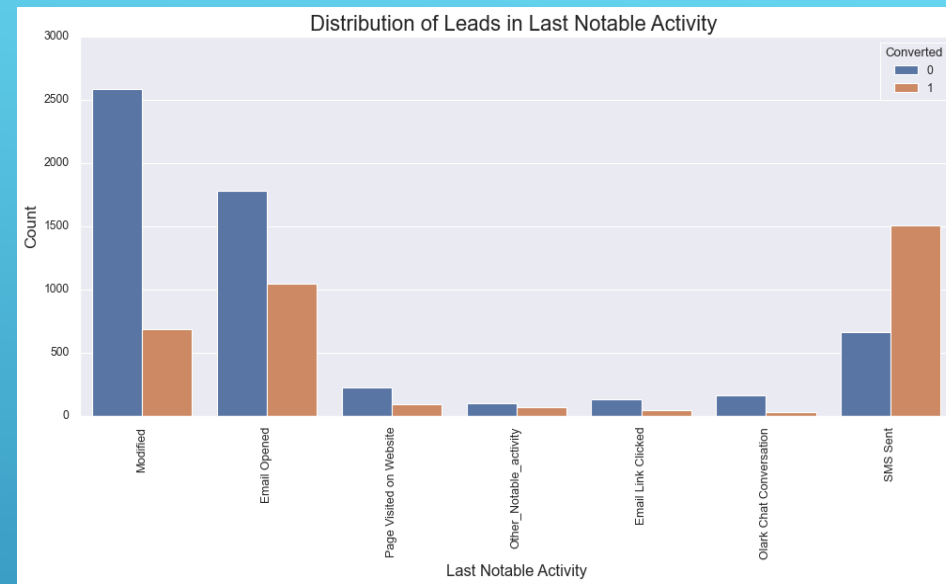
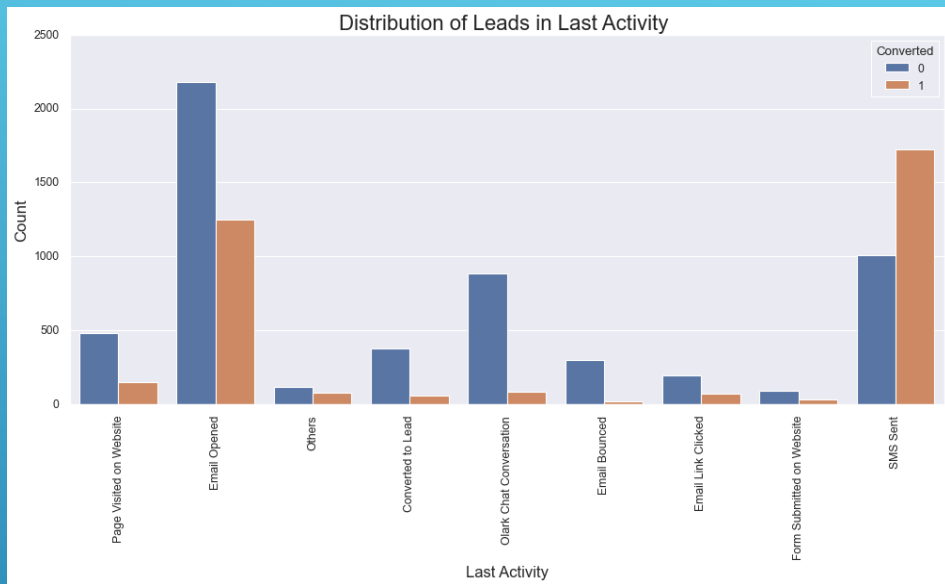
After Understanding the Data Cleaning Process started

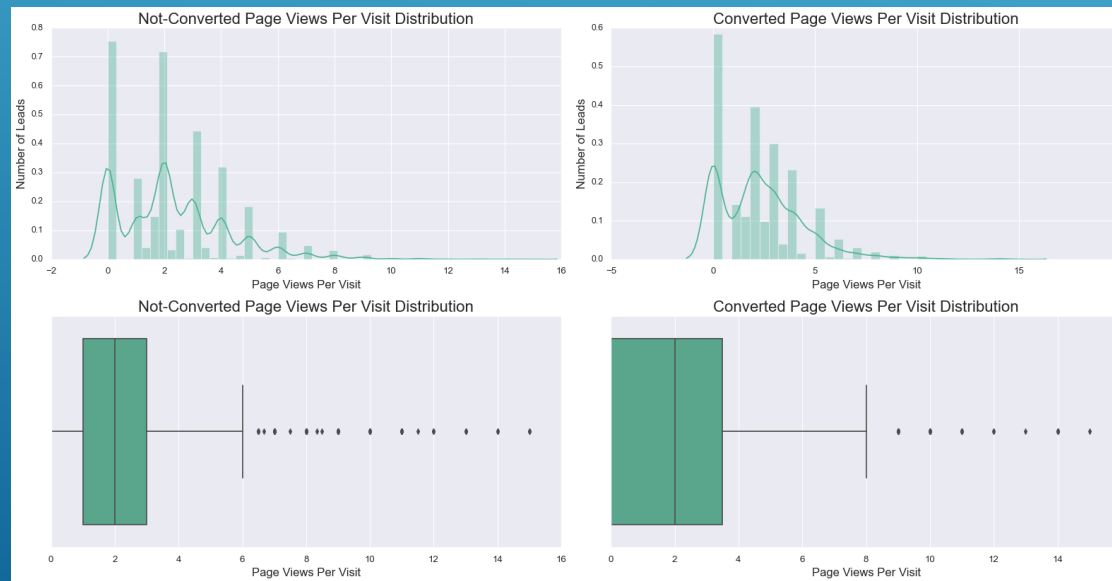
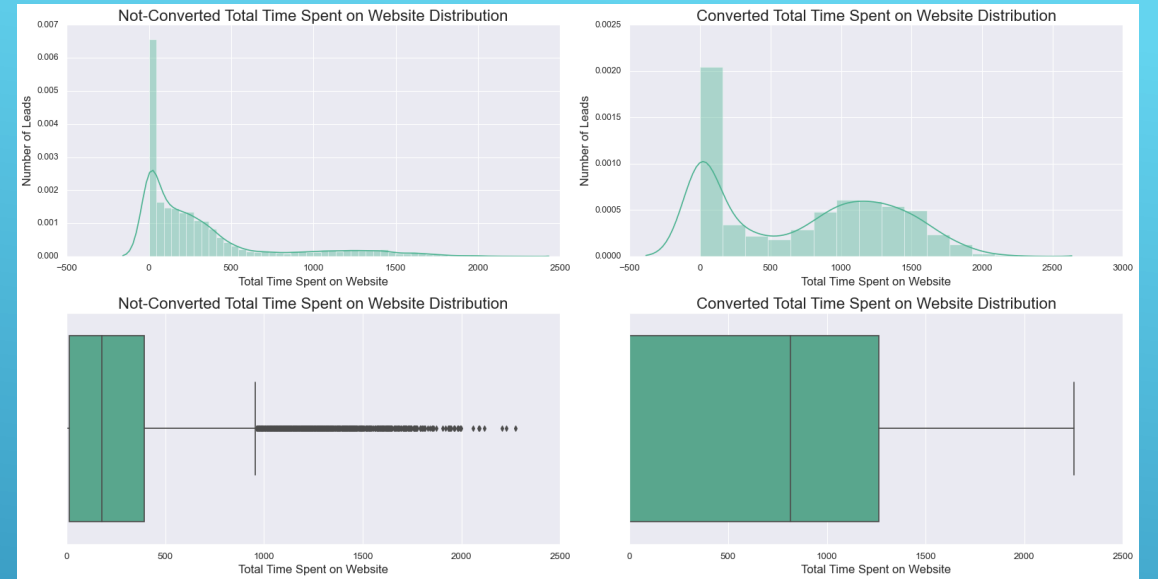
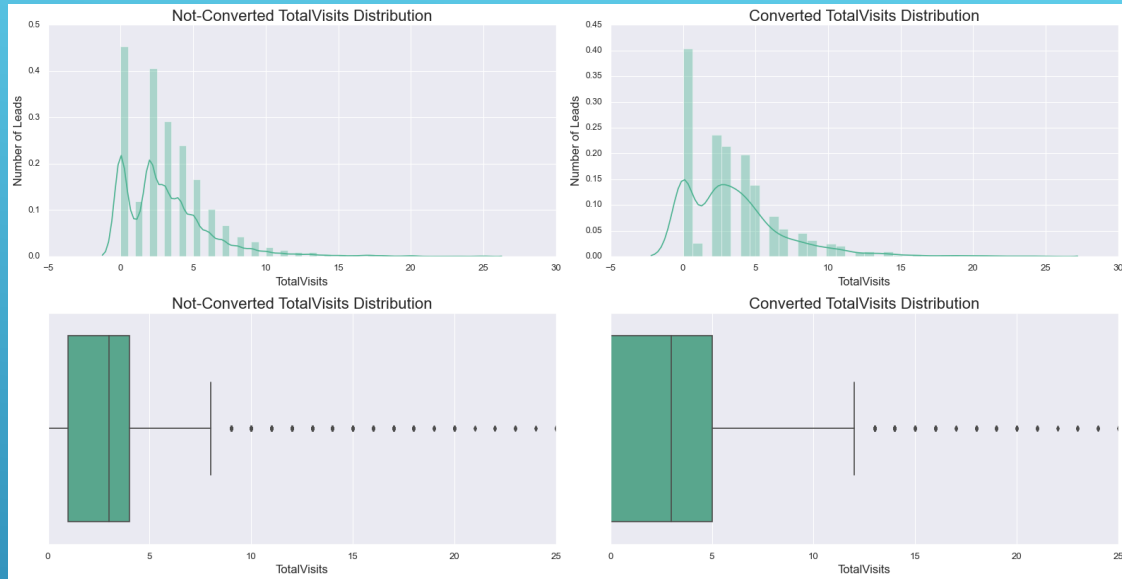
Following Columns were removed after scrutiny

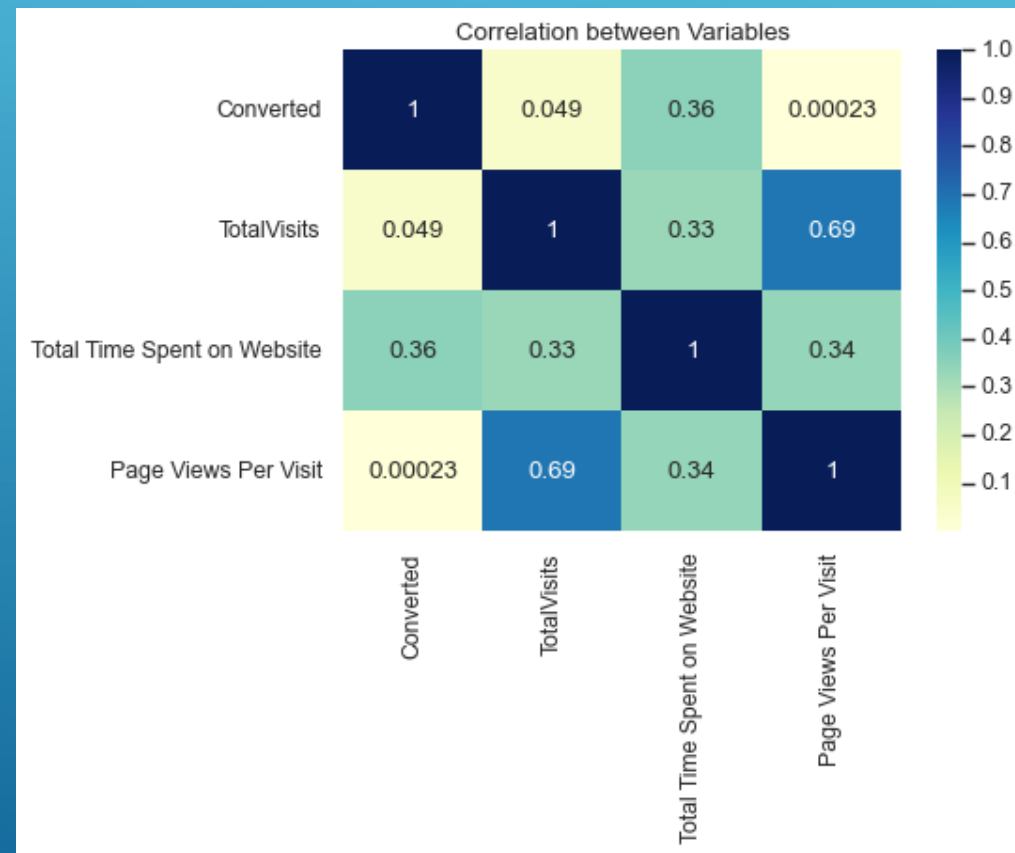
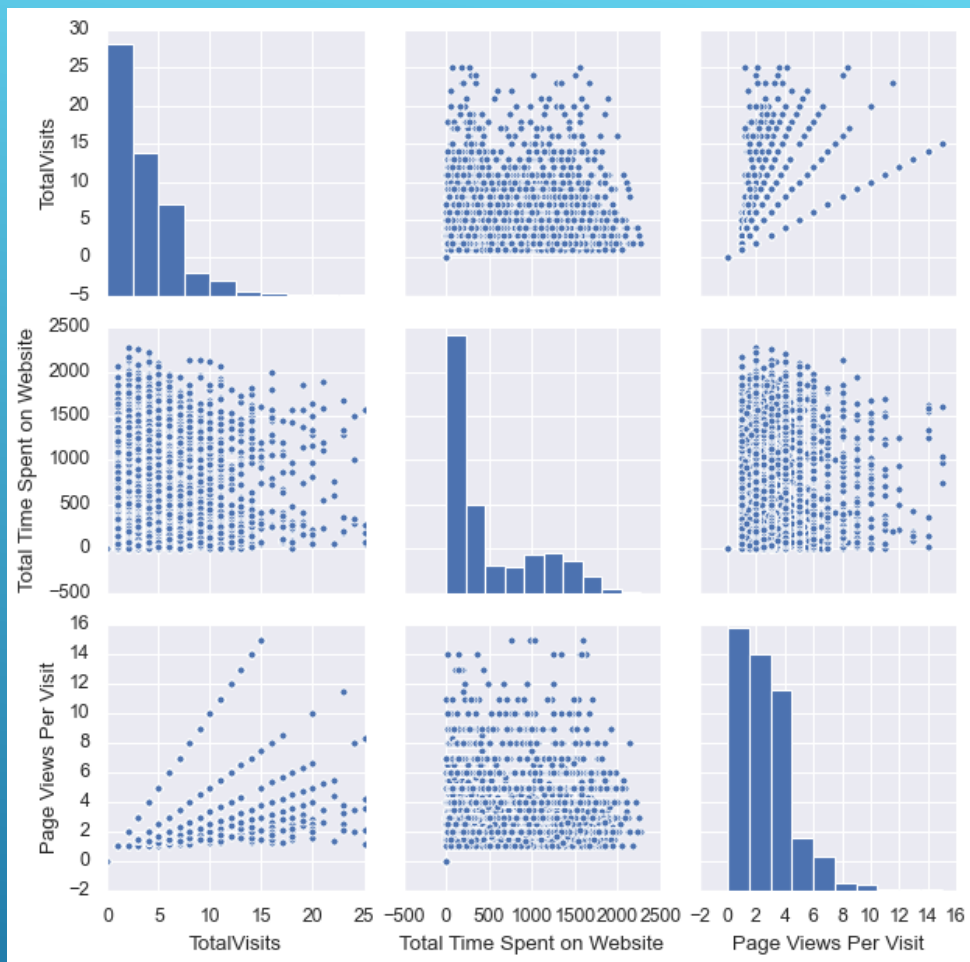
Prospect ID	How did you hear about X Education	Newspaper	Receive More Updates About Our Courses	Asymmetrique Activity Index
Lead Number	Asymmetrique Profile Score	Do Not Call	Get updates on DM Content	Through Recommendations
Lead Profile	Asymmetrique Activity Score	Magazine	Update me on Supply Chain Content	Digital Advertisement
Lead Quality	Asymmetrique Profile Index	X Education Forums	What matters most to you in choosing a course	
Country	Newspaper Article	Search	'I agree to pay the amount through cheque	

Exploratory Data Analysis










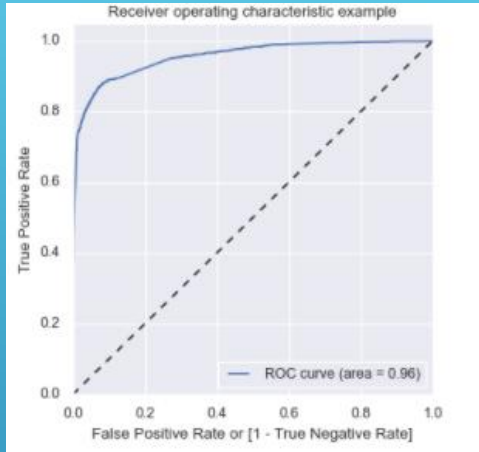
Data Preparation

- *Columns with more than two levels were converted to numerical values by creating the Dummy variables (with dropping the first variable)*
 - *The whole data was then Split into two Data-frames, Train and Test Data, which would*
 - *then be used to build regression model.*
 - *We used StandardScaler to standardize the numerical values.*
- 
- A series of three parallel white diagonal lines extending from the bottom right corner towards the center of the slide.

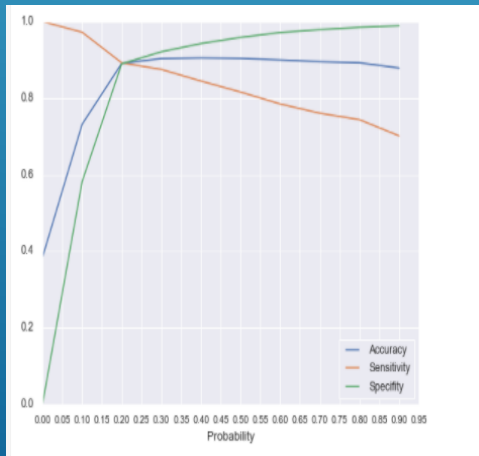
Model Building

- We first used RFE (Recursive Feature Elimination) to obtain top 20 relevant variables.
- The next step was to make the model more stable, by checking the p-values and VIF (Variance Inflation Factor).
- Once the model was stable, we predicted probabilities on the Train data and created new Data-Frame, with prediction 1 if 'Converted_Prob' was greater than 0.5 else 0.
- We then calculated the Confusion matrix on this new Data-Frame and calculated the Accuracy (90.38%), Sensitivity (81.6%), Specificity (95.9%), also plotted the ROC curve to find the area under the curve, which came out to be 0.2.

Model Building

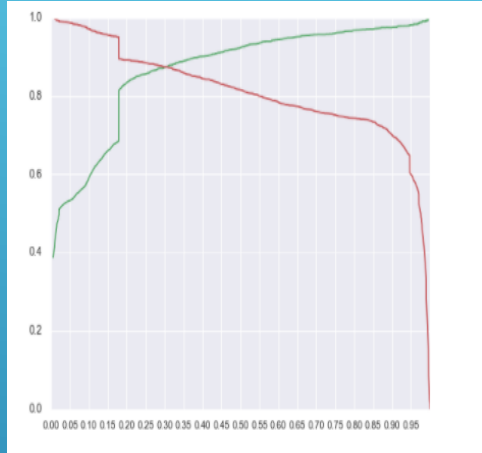


- Found the optimum Cut-Off by comparing the Accuracy, Sensitivity and Specificity for probabilities [0.0 to 0.9], which came out to be 0.3.



- Plotted a graph with Accuracy, Sensitivity and Specificity in the Y-axis and Probability in the X-axis, to find out the intersection point of Accuracy, Sensitivity and Specificity to find the optimum Cut-Off, that came to be at 0.3.

Model Building



- Also plotted a graph with Precision and Recall in Y-axis and Probability on the X-axis, this also came out to be intersecting at 0.3.
- A new column was added 'Final_Predicted', and the values were recorded.

Model Evaluation

- In this step, predictions were made on the Test data and the values were recorded.
- We checked the Confusion matrix and calculated Accuracy (91.15%), Sensitivity (86.8%), Specificity (93.74%), Precision (89.17%) and Recall (86.8%) of the final predicted model.
- We created a Lead_Score [$\text{'Converted_Prob'} * 100$] to provide a score between 0 to 100 where higher the value of Lead_Score means the Lead is 'Hot' and there is a high possibility that the lead can be converted.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- When 'Tag' was:

- Closed by Horizzon
- Lost to EINS
- Will revert after reading the email

- When 'Lead Source' was:

- Welingak Website
- Reference

- When Occupation was:

- Working Professional

- When Specialization was :

- Travel and Tourism

Let us compare the values obtained for Train & Test:

Train Data:

Accuracy : 90.33%

Sensitivity : 87.50%

Specificity : 92.10%

Test Data:

Accuracy : 91.15%

Sensitivity : 86.8%

Specificity : 93.74%

THANK YOU

