LEAD SCORING CASE STUDY

Submitted By –

Shubhangi Kalyane

Pallavi Bothra

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not.

The typical lead conversion rate at X education is around 30%. There are a lot of leads generated in the initial stage (top) but only a few of them come out as paying customers from the bottom. In the middle stage, you need to nurture the potential leads well (i.e. educating the leads about the product, constantly communicating etc.) in order to get a higher lead conversion. X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Goals

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. 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.

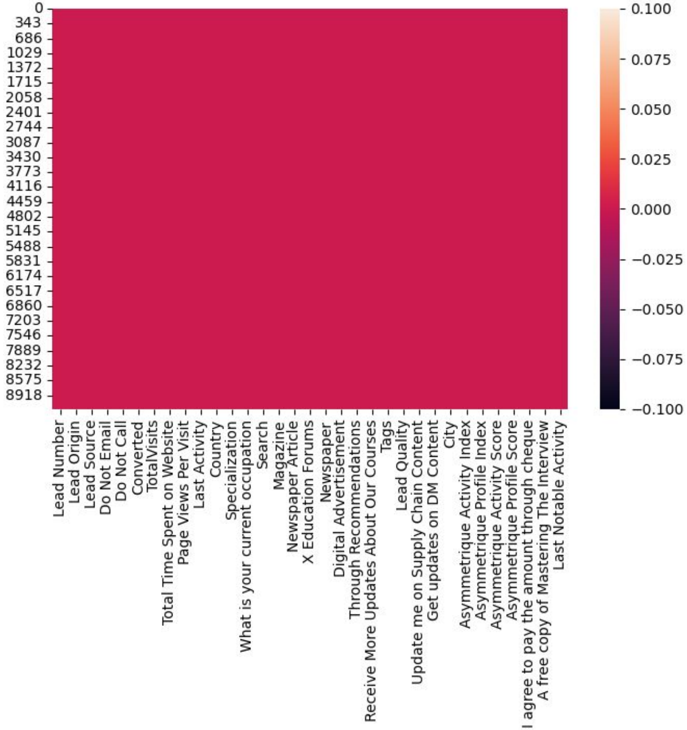
There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Strategy

- Read and understand the data
- Clean the data and Prepare the data
- Exploratory Data Analysis
- Model Building
  - Feature Scaling
  - Data Split into Train and Test
  - Build Logistic Regression Model
- Model Evaluation
  - Accuracy
  - Specificity & Sensitivity
  - Precision & Recall
- Making Predictions on the Test Set

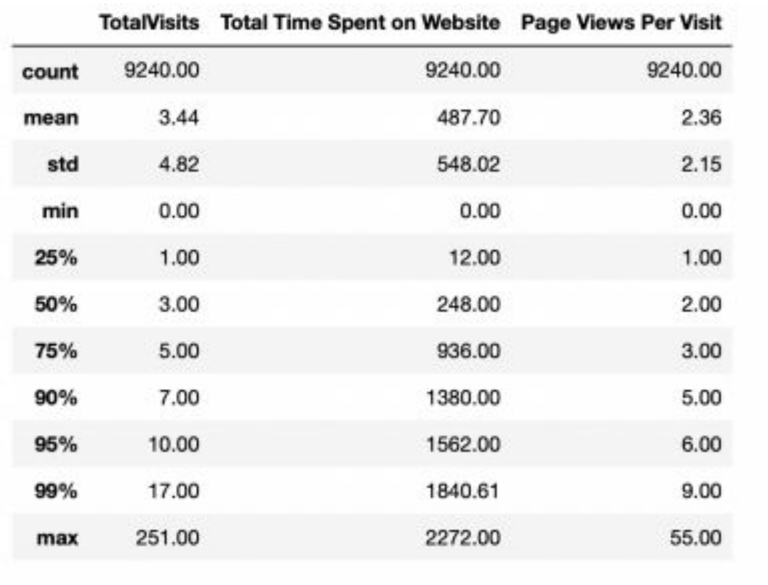
Data Cleaning and Preparation

We have cleaned the data and handled the null values as per requirement and can be seen in the below heatmap.

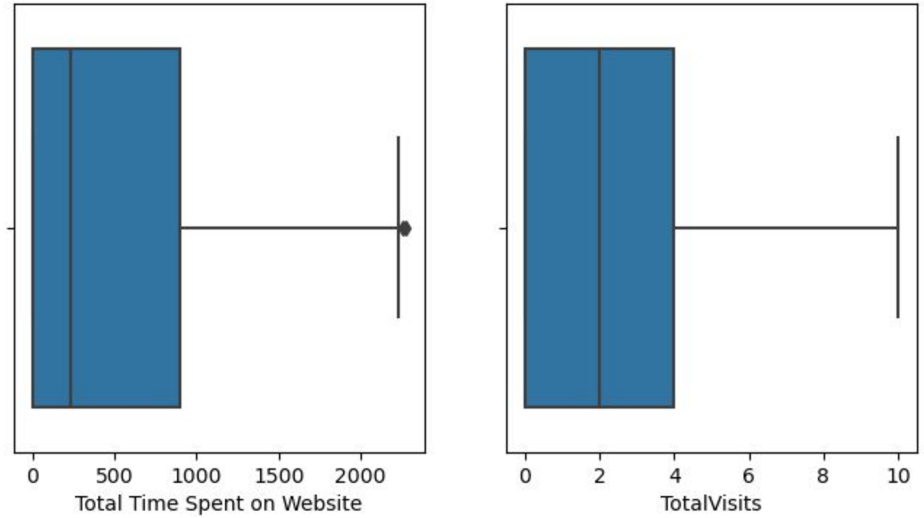


Exploratory Data Analysis

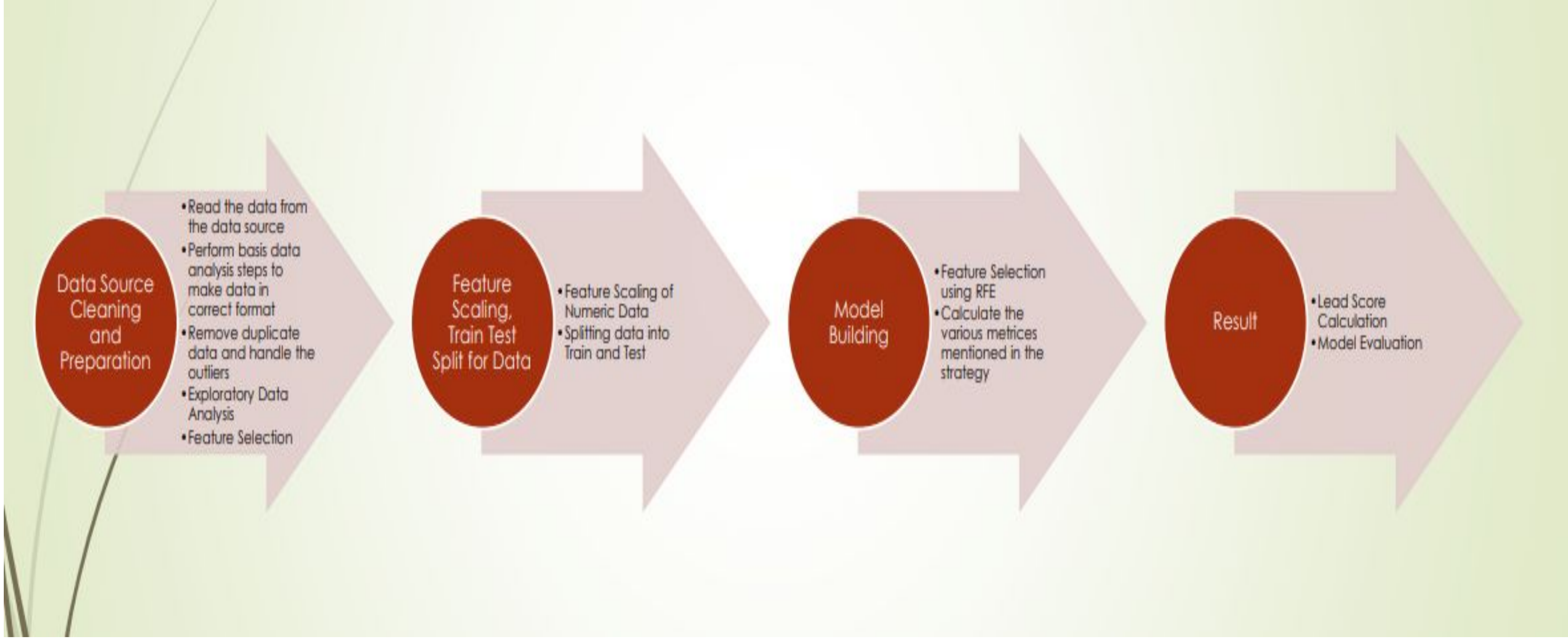
We have checked the outliers present in the numeric columns and handled them.



	TotalVisits	Total Time Spent on Website	Page Views Per Visit
count	9240.00	9240.00	9240.00
mean	3.44	487.70	2.36
std	4.82	548.02	2.15
min	0.00	0.00	0.00
25%	1.00	12.00	1.00
50%	3.00	248.00	2.00
75%	5.00	936.00	3.00
90%	7.00	1380.00	5.00
95%	10.00	1562.00	6.00
99%	17.00	1840.61	9.00
max	251.00	2272.00	55.00



Problem Solving Process



Data Source Cleaning and Preparation

- Read the data from the data source
- Perform basis data analysis steps to make data in correct format
- Remove duplicate data and handle the outliers
- Exploratory Data Analysis
- Feature Selection

Feature Scaling, Train Test Split for Data

- Feature Scaling of Numeric Data
- Splitting data into Train and Test

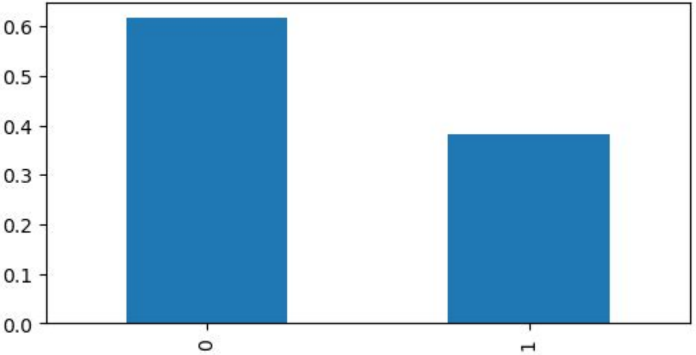
Model Building

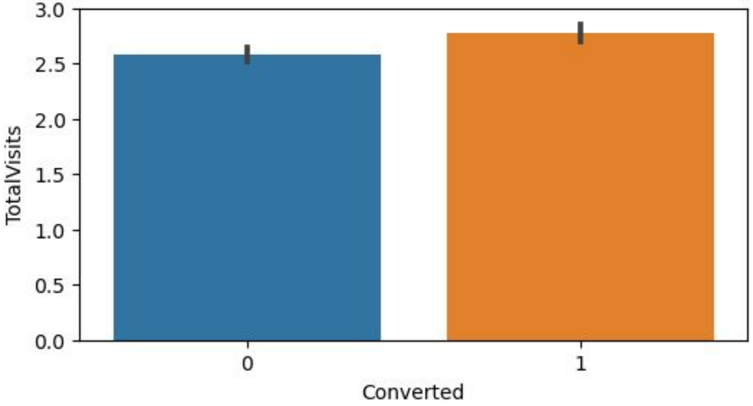
- Feature Selection using RFE
- Calculate the various metrices mentioned in the strategy

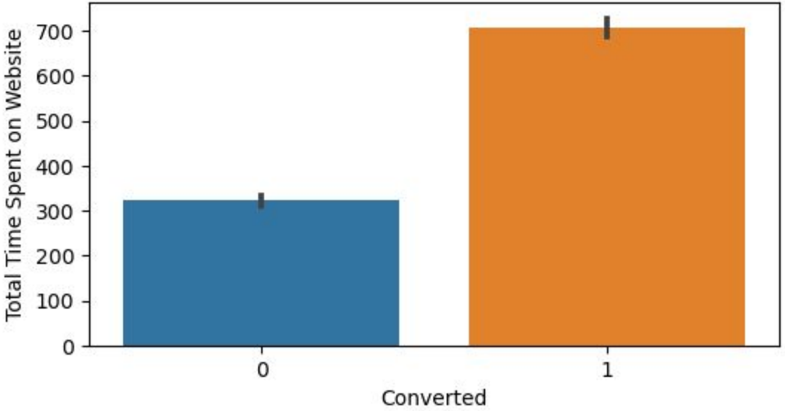
Result

- Lead Score Calculation
- Model Evaluation

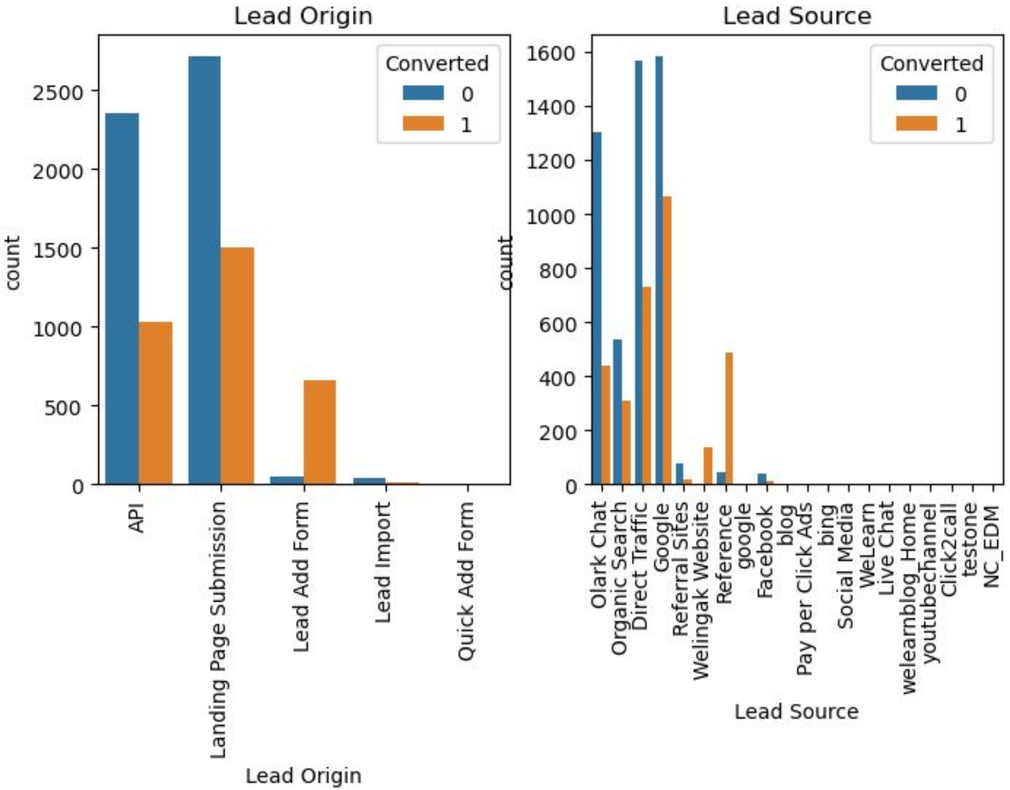
We have 38% conversion rate in total as per below plot, also have observed that total time spent on website have high conversions :



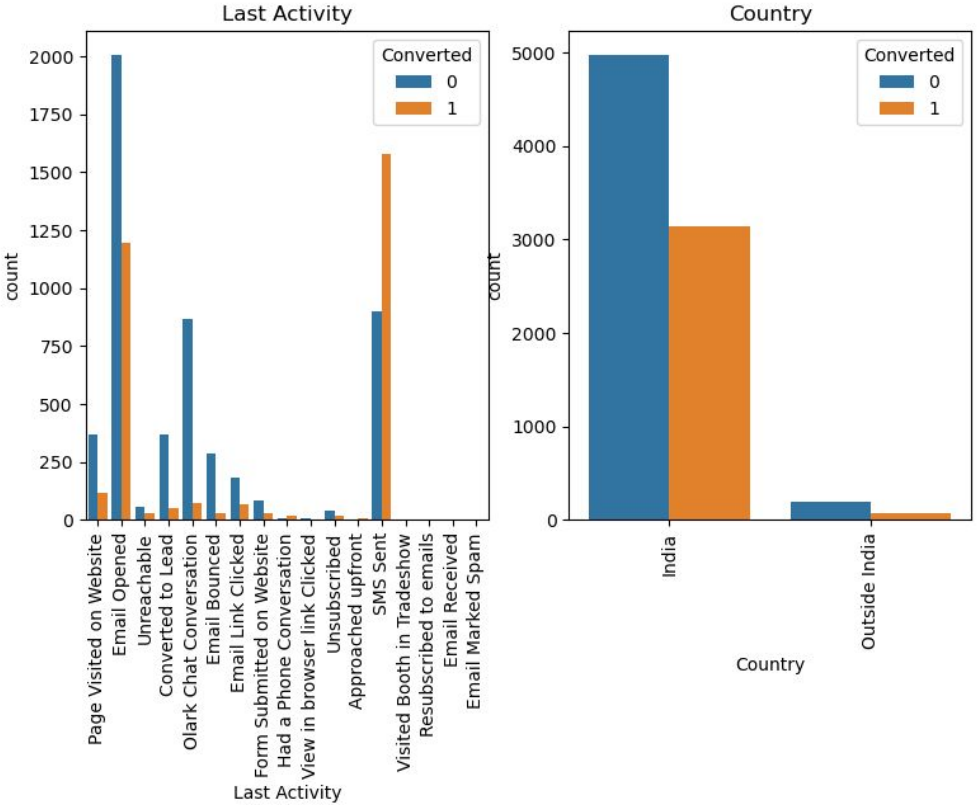




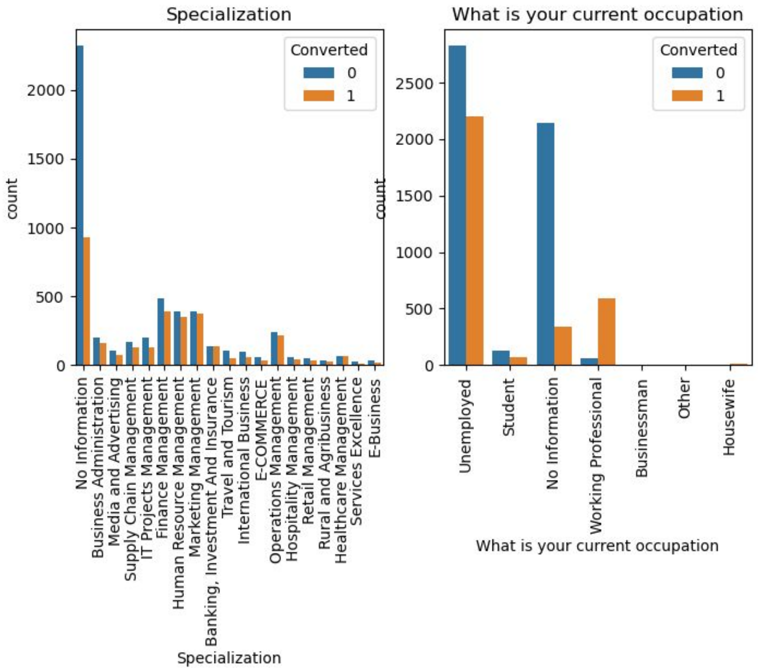
- We observe that API and Landing Page Submission has approx 30% and 35% conversion leads respectively.
- We observe that in leads source Google and Direct Traffic has more potential to convert the leads followed by Olark chat and Organic search



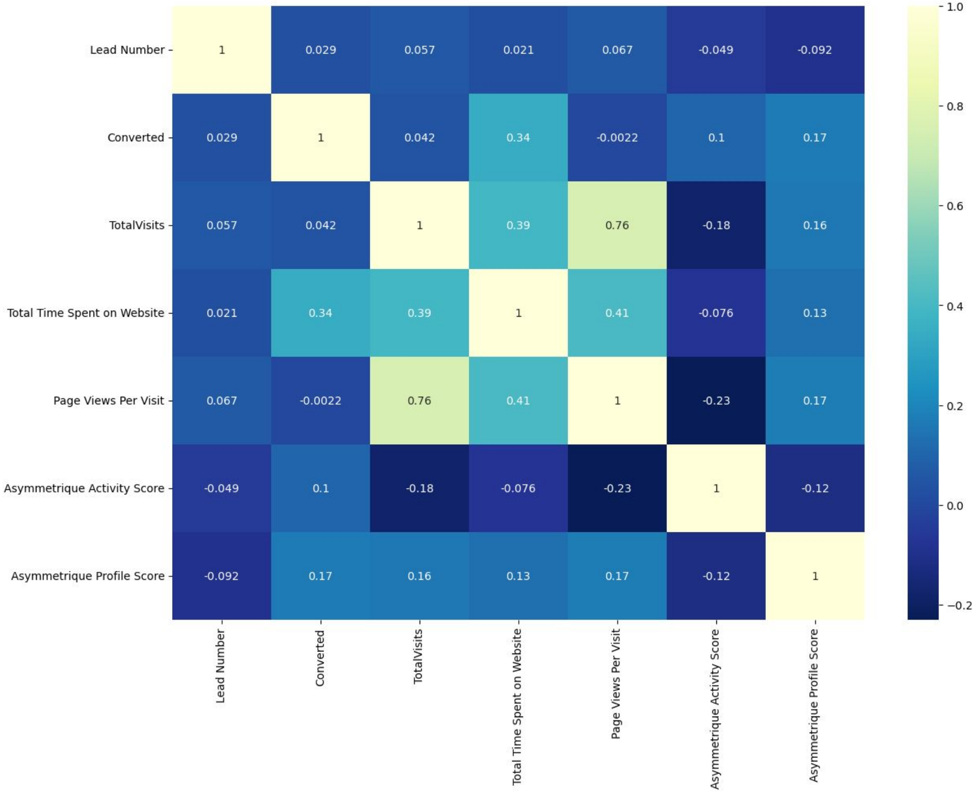
- We observe that SMS sent has high conversion leads followed by Email opened.
- india has max count, also max conversion



- We observe that the No information category has high conversion rates. Also we can see that in other specialization there are lots of opportunity to convert the leads
- We observe that unemployed leads has high conversion rates, so focus should be made to convert more positive leads in umemployed counts, as it is obvious working professional are most likely to get converted for their growth and all.

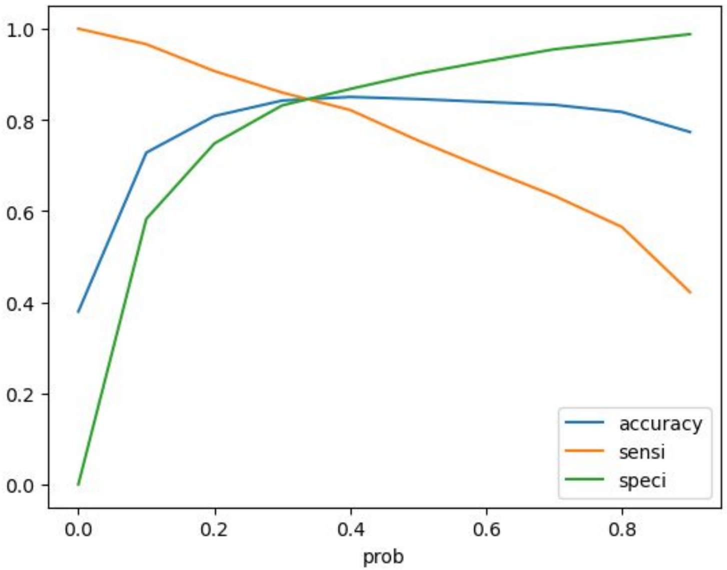


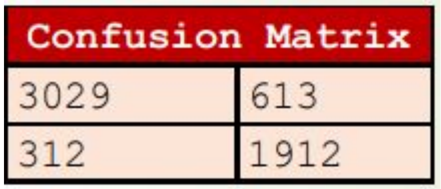
- We observe that Pages Views per visit and total visits are highly postively co related.
- Asymmetrique activity score is negatively correlated with total visit and Pages Views per visit.



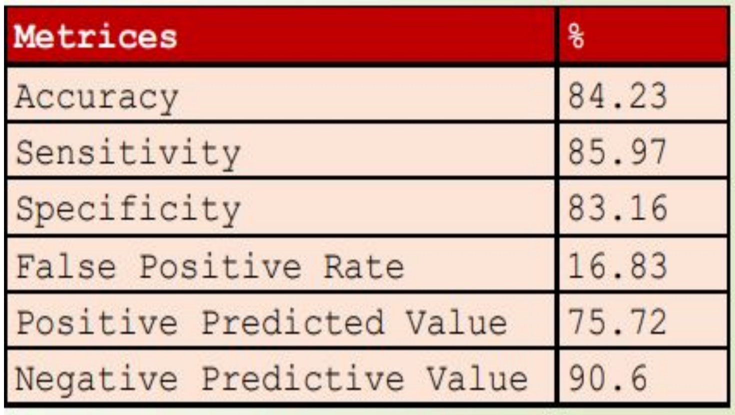
Model Evaluation – Train Dataset

The graph depicts the optimal cut off of 0.3 based on Accuracy, Sensitivity and Specificity





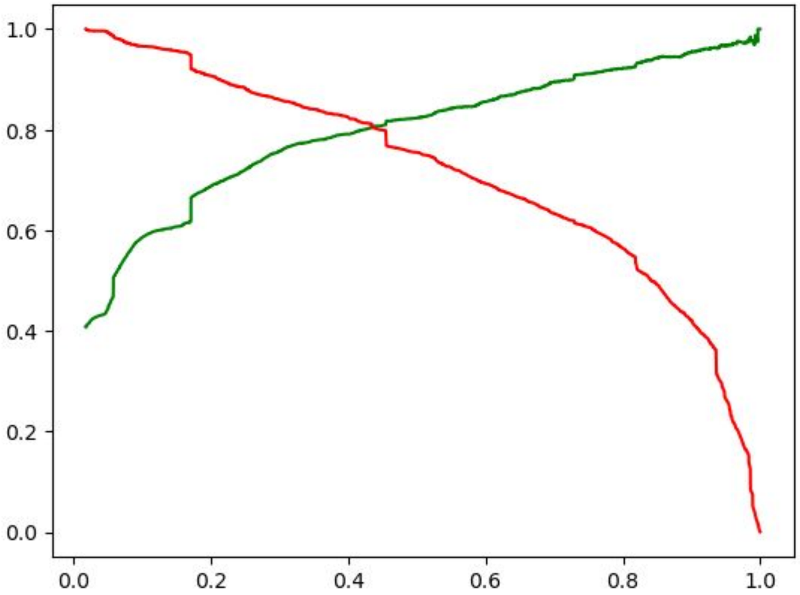
Confusion Matrix	
3029	613
312	1912

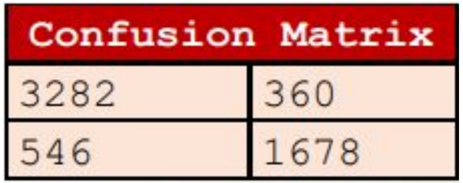


Metrices	%
Accuracy	84.23
Sensitivity	85.97
Specificity	83.16
False Positive Rate	16.83
Positive Predicted Value	75.72
Negative Predictive Value	90.6

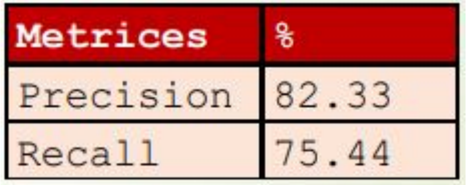
Model Evaluation – Precision & Recall

The graph depicts the optimal cut off of 0.42 based on Precision and Recall



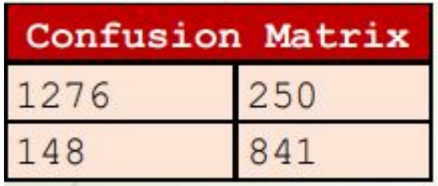


Confusion Matrix	
3282	360
546	1678

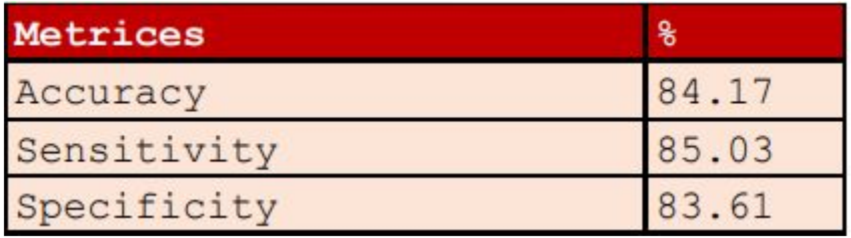


Metrices	%
Precision	82.33
Recall	75.44

Model Evaluation – Test Dataset



Confusion Matrix	
1276	250
148	841



Metrices	%
Accuracy	84.17
Sensitivity	85.03
Specificity	83.61

Conclusion

1. We have observed that the optimal cut off is at 0.3 based on the sensitivity and specificity while calculating the final prediction.
2. The top 3 variables which will help us to convert leads are:
    1. Lead Lead Source_Welingak Website
    2. Lead Orogin_Lead Add Form
    3. What is your current occupation_Working Professional
3. After RFE and VIF checks we concluded that the 3rd model will be a good model for the logistic regression. Here all the variables had low VIF and low p values.
4. Accuracy for test data set is 84%, Sensitivity is 85% and specificity is 84% - all of them are pretty close to the train data set results.
5. Seeing the overall accuracy at 84% it seems to be a pretty good model for our analysis of lead scoring.