Lead Scoring Case Study

Building Logistic Regression model to assign a lead score between 0 and 100 to each of the leads

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead.
- As you can see, 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 us to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Business Expectation

There are quite a few goals for this case study:

1. Handling of data to clean and structure it before building the model.
2. 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.
3. It is expected to have a target lead conversion rate to be around 80%, i.e. model should be around 80% accuracy.

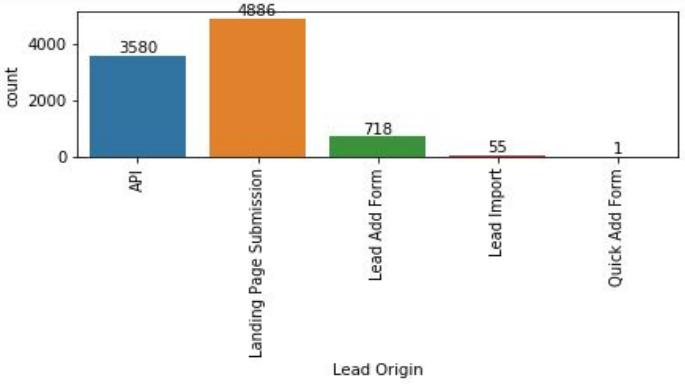
Data Handling and EDA

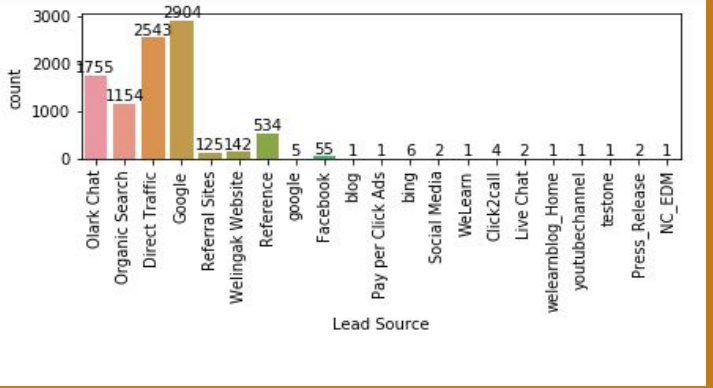
1. Data staging and clean up: This is the basic data cleanup and preparation stage. Clean dataset and prepare the master dataset.
2. Sanity checks: The next step is doing a quick sanity check of the entire dataset to observe any unusual data points that should not exist.
3. Univariate Analysis: Finally we begin with the univariate analysis part. This is where visualisation tools like histograms and boxplots come in handy as they help in analysing numerical features.
4. Bivariate Analysis: Then, you go ahead and evaluate the relationship between the target variable and the rest of the features. Here plots like scatter plots, pair plots, correlation matrices come in very handy to do the analysis.

Data related information

- No of Rows in the dataset: 9240, No of columns in the dataset: 37
- Dropping Columns which has 40% null values.
- Many of the categorical variables have a level called 'Select' which needs to be handled because it is as good as a null value - ['Specialization', 'How did you hear about X Education', 'Lead Profile', 'City']
- Remove ['Prospect ID','Lead Number','Last Notable Activity'] which is not required for analysis.
- Columns with one unique value whose count and frequency are same - ['Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque']
- Highly Skewed columns - ['Do Not Call', 'Search', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations']
- Grouping low frequency value levels to Others: Lead Source, Last Activity.
- In Lead Source column change google to Google.
- Change column name 'A free copy of Mastering The Interview' to 'Free_copy'
- Data Imbalance Ratio - 1.59 : 1

Univariate analysis



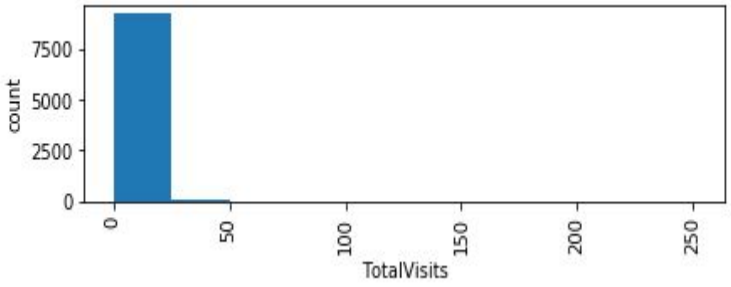


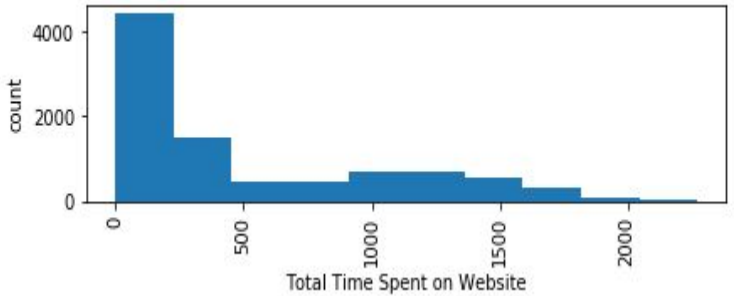
Categorical variables

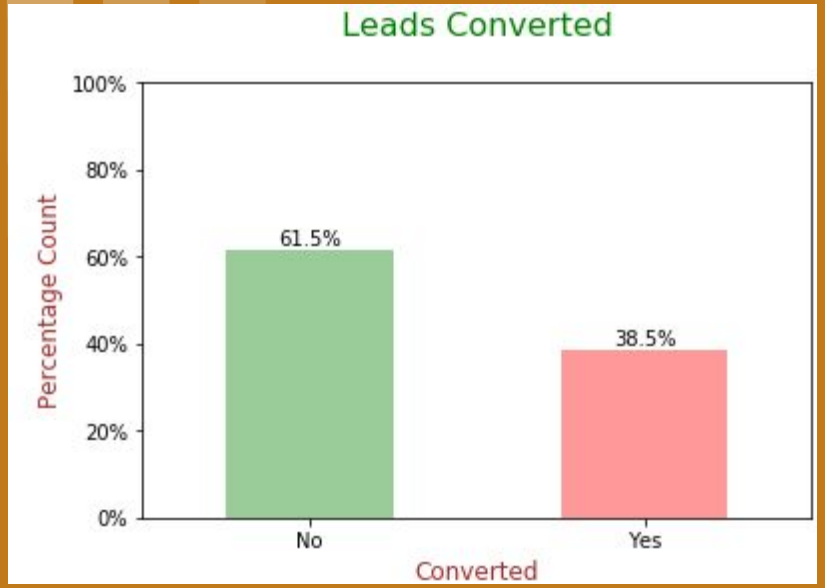
In Lead Origin, we can see that Landing Page Submission is the most leads are coming from.
In Lead Source, we can see that Google is were the most leads are coming.

Numerical variables

In Total Visits we can see that initial visits are only done by the users.
In Total time spent on website is also in the initial stages itself.

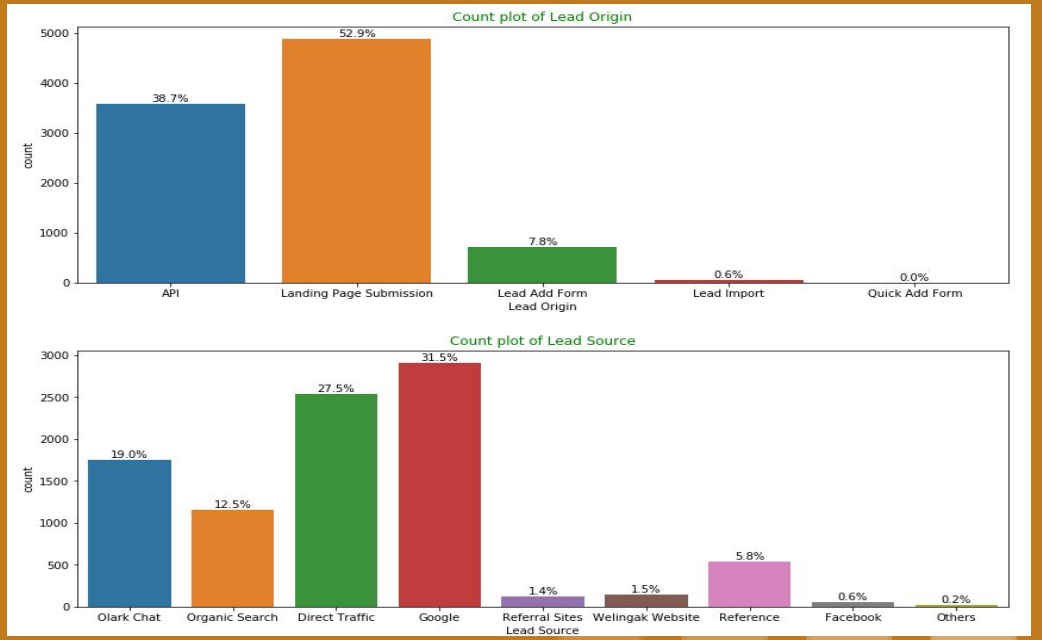






Data Imbalance

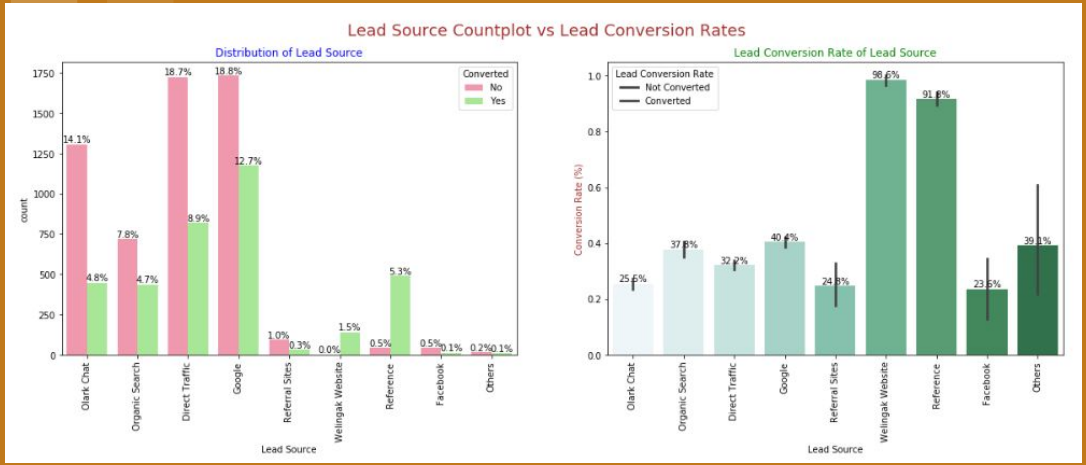
Data Imbalance Ratio : 1.59 : 1



Count Plots on Lead Origin and Lead Sources

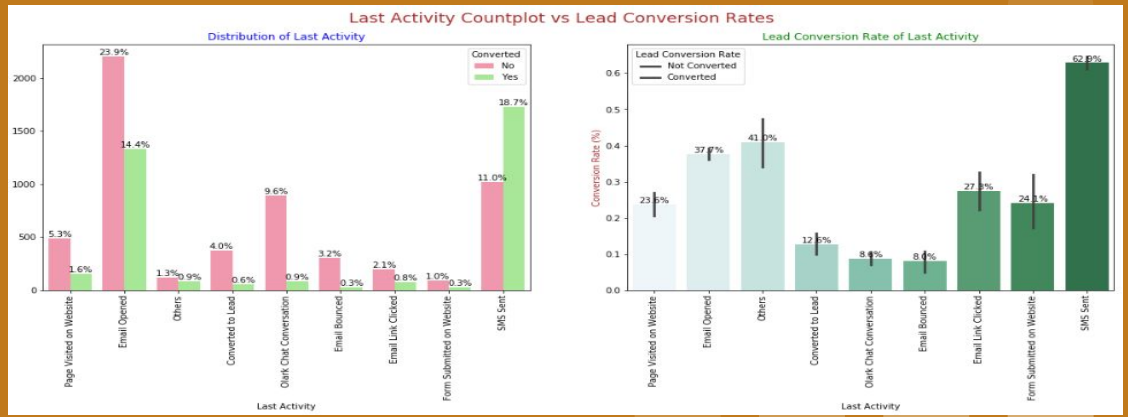
We can see that 52.9% of leads are from Landing Page Submission.
We can see that 31.5% of the leads are from Google.

Bivariate analysis

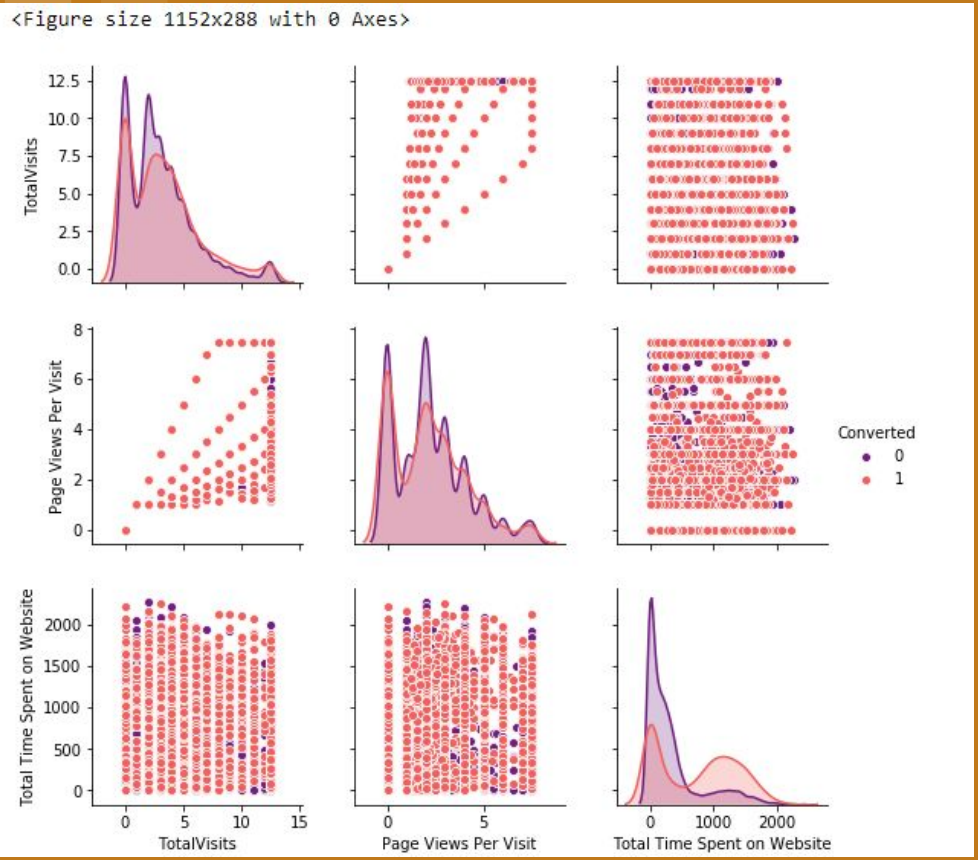


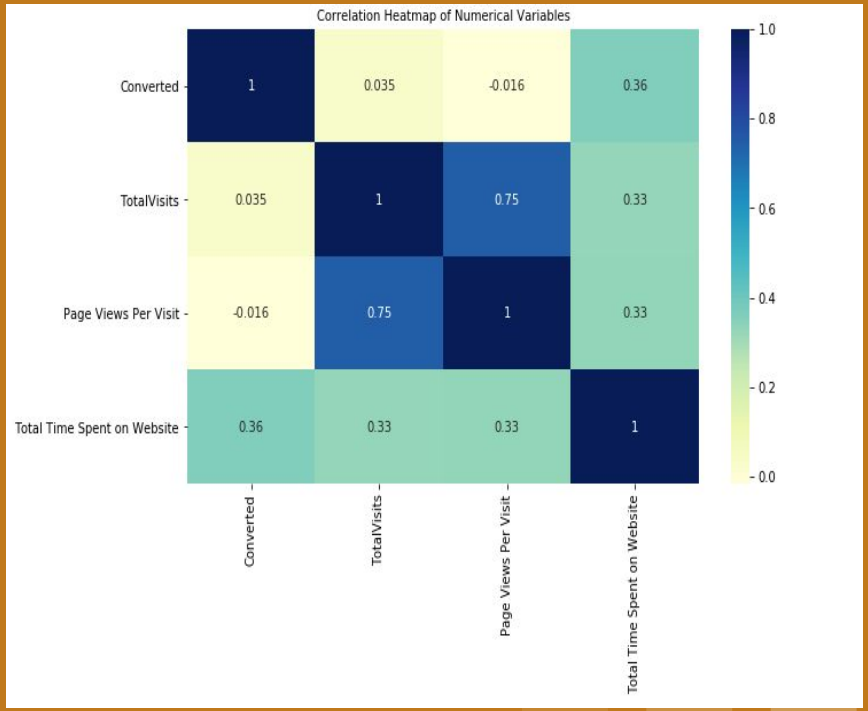
We can see that Welingak website and Reference are having a high lead conversion rate compared to other Lead Sources

We can see that if last activity is SMS sent then leads conversion rate is more. So this can be taken in consideration for converting them into leads.



Pair plot and Heatmaps





We can see from both Pairplot and Heat map that there is a correlation between Total Visits and Page Views Per Visit. And the relation is of 0.75 as we see that in heatmap.

Model Building and Model Evaluation

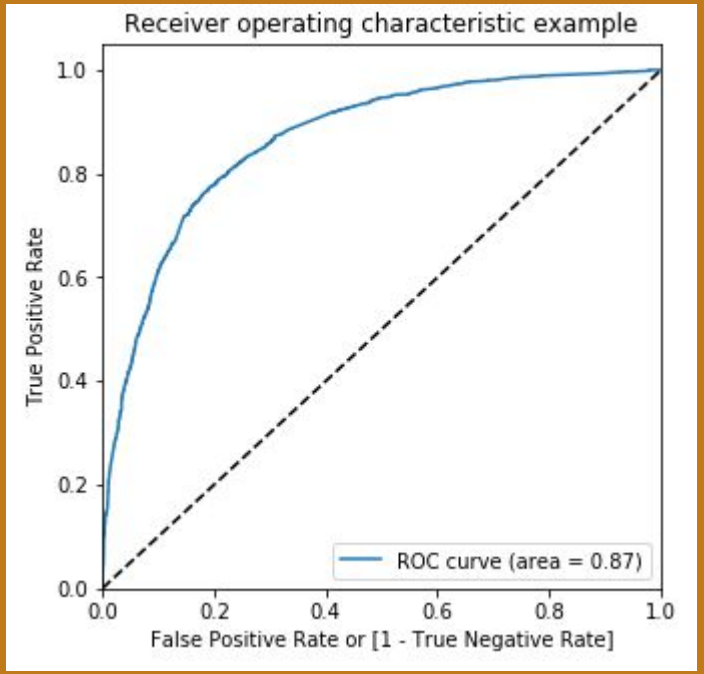
- After missing value imputation and outlier treatment, we will get the master dataset
- Dummy variable creation for categorical variables
- Test-train split of the data
- Standardisation of the scales of continuous variables
- Logistic regression model was built in Python using the function GLM() under statsmodel library.
- Some of these variables were removed first based on an automated approach, i.e. RFE and then a manual approach based on VIF and p-value.
- Model Evaluation was done using:
  > Accuracy
  > Sensitivity and Specificity
  > Optimal cut-off using ROC curve
  > Precision and Recall
- Then Predictions were made on the test set

Final Model

Generalized Linear Model Regression Results

Dep. Variable:	Converted	No. Observations:	6468
Model:	GLM	Df Residuals:	6455
Model Family:	Binomial	Df Model:	12
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-2888.4
Date:	Mon, 17 Jun 2024	Deviance:	5776.7
Time:	21:02:42	Pearson chi2:	6.65e+03
No. Iterations:	7	Covariance Type:	nonrobust

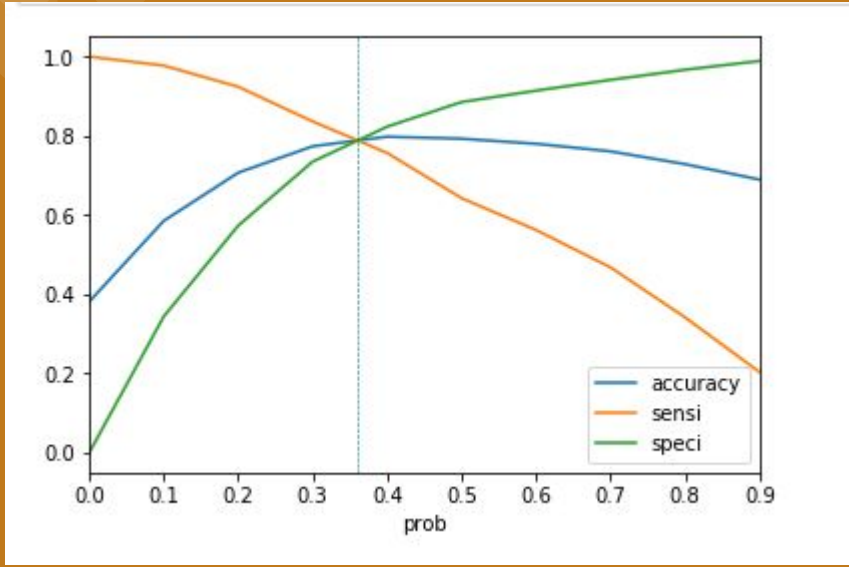
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6554	0.136	-4.805	0.000	-0.923	-0.388
Total Time Spent on Website	1.0684	0.038	28.269	0.000	0.994	1.142
Lead Origin_Landing Page Submission	-1.4068	0.121	-11.635	0.000	-1.644	-1.170
Lead Source_Olark Chat	0.9306	0.116	8.032	0.000	0.703	1.158
Lead Source_Reference	3.0364	0.211	14.413	0.000	2.623	3.449
Lead Source_Welingak Website	5.3805	0.728	7.388	0.000	3.953	6.808
Last Activity_Email Opened	0.9166	0.100	9.147	0.000	0.720	1.113
Last Activity_Olark Chat Conversation	-0.5734	0.181	-3.165	0.002	-0.929	-0.218
Last Activity_Others	1.2859	0.229	5.611	0.000	0.837	1.735
Last Activity_SMS Sent	2.0317	0.103	19.694	0.000	1.830	2.234
Specialization_Hospitality Management	-1.0198	0.303	-3.368	0.001	-1.613	-0.426
Specialization_International Business	-0.5404	0.244	-2.216	0.027	-1.018	-0.062
Specialization_Unknown	-1.5388	0.118	-13.069	0.000	-1.770	-1.308



Model Accuracy : 0.7927
Area under the curve of the ROC is 0.87 , which is good model

Whichever is having positive value in coef fields(in the left screenshot) which provides top leads whereas negative values in coef fields(in the left screenshot) which should be improved.

Finding Optimal Cutoff Point

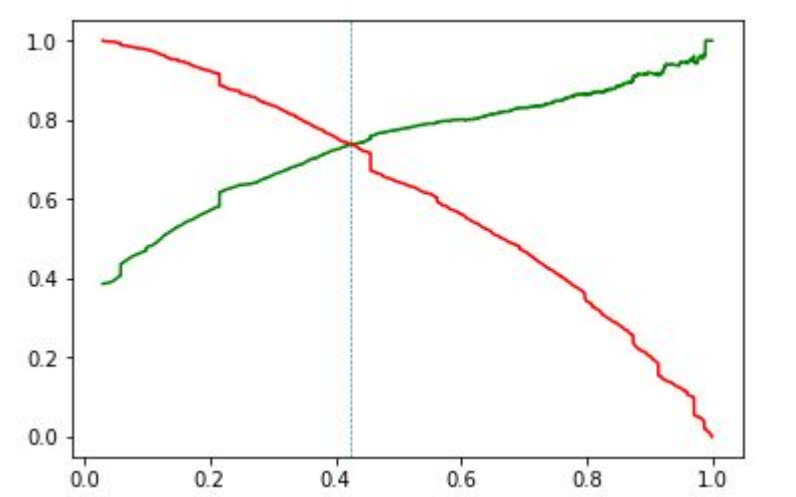


Around 0.36,we get the optimal values of the three metrics. So let's choose 0.36 as cutoff now.

Train Data Model

True Negative	: 3169
True Positive	: 1944
False Negative	: 522
False Positive	: 833
Model Accuracy	: 0.7905
Model Sensitivity	: 0.7883
Model Specificity	: 0.7919
Model Precision	: 0.7
Model Recall	: 0.7883
Model True Positive Rate (TPR)	: 0.7883
Model True Negative Rate (TPR)	: 0.7919
Model False Positive Rate (FPR)	: 0.2081
Model False Negative Rate (FPR)	: 0.2117

Final Prediction on Test Data



Precision Recall curve: we get value of 0.425.
After which we substitute the value of threshold to Test Data, we get the metrics as below:

True Negative	: 3353
True Positive	: 1823
False Negative	: 643
False Positive	: 649
Model Accuracy	: 0.8002
Model Sensitivity	: 0.7393
Model Specificity	: 0.8378
Model Precision	: 0.7375
Model Recall	: 0.7393
Model True Positive Rate (TPR)	: 0.7393
Model True Negative Rate (TPR)	: 0.8378
Model False Positive Rate (FPR)	: 0.1622
Model False Negative Rate (FPR)	: 0.2607

After which if we predict for the test data prediction:

True Negative	: 1406
True Positive	: 796
False Negative	: 299
False Positive	: 271
Model Accuracy	: 0.7944
Model Sensitivity	: 0.7269
Model Specificity	: 0.8384
Model Precision	: 0.746
Model Recall	: 0.7269

So the Model accuracy is 79% which is expected as per the problem statement which should be around 80%.

Recommendation:

- Can be more focused on positive coefficients which gives more on marketing strategies.
- Top Lead sources can make positive impact on the leading score.
- Last activities are acting as both positive and negative so if it focus in proper way then it can improve the performance.
- Make advancement in the website so if the person stays for a long time then there is chances of them converting into Leads which we can see in total time spent in Website
- Advertisement can be improved so that there is positive impact.
- Monitoring the Last activity can also help in contacting the users whenever possible.
- Improvement is required in few specialization as there is low coefficient rates or can concentrate on them later.
- Olark Chat Conversation should be improved or it will impact on lead scoring.
- Improve Landing page submission or get to know what is causing the leads to stop using the website.

Thank you!!