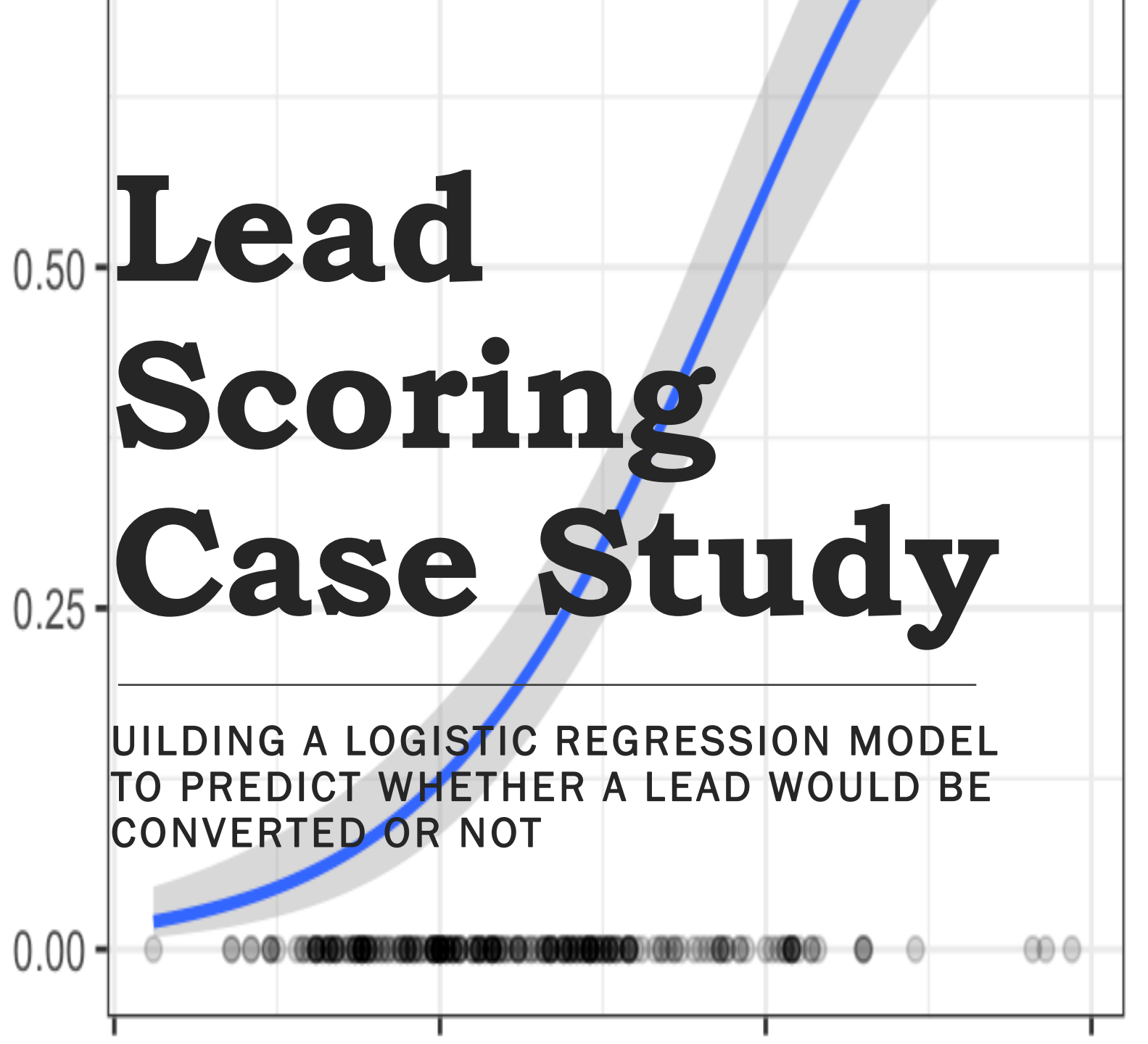


Lead Scoring Case Study

UILDING A LOGISTIC REGRESSION MODEL TO PREDICT WHETHER A LEAD WOULD BE CONVERTED OR NOT

Introduction:

X Education is an education company that sells online courses to industry professionals. Their course are marketed on various websites. People interested in such courses fill up a form, which then is classified as a lead. Few leads get converted, while most don't. We have built a logistic regression model that predicts if a lead can be successfully converted or not.

Contents:

Analysis:

The first step of the analysis is data quality check. We look for the shape of the dataset, the information, missing values and make the necessary treatment e.g. imputations using mean/median/mode

The next step of the process is exploratory data analysis and outlier treatment.

Further we move towards model building. Here we have made the additional use of woe and IV.

After model building we evaluate the model and check for accuracy, precision and recall.

The last step is predictions and calculating lead scores.

Data Quality Check – (data imbalance) & (Outlier Treatment):

Data Imbalance:

The dataset has 37 columns and 9024 rows.

The pie chart shows the data imbalance in the dataset:

Converted leads consists of Only 38.5%

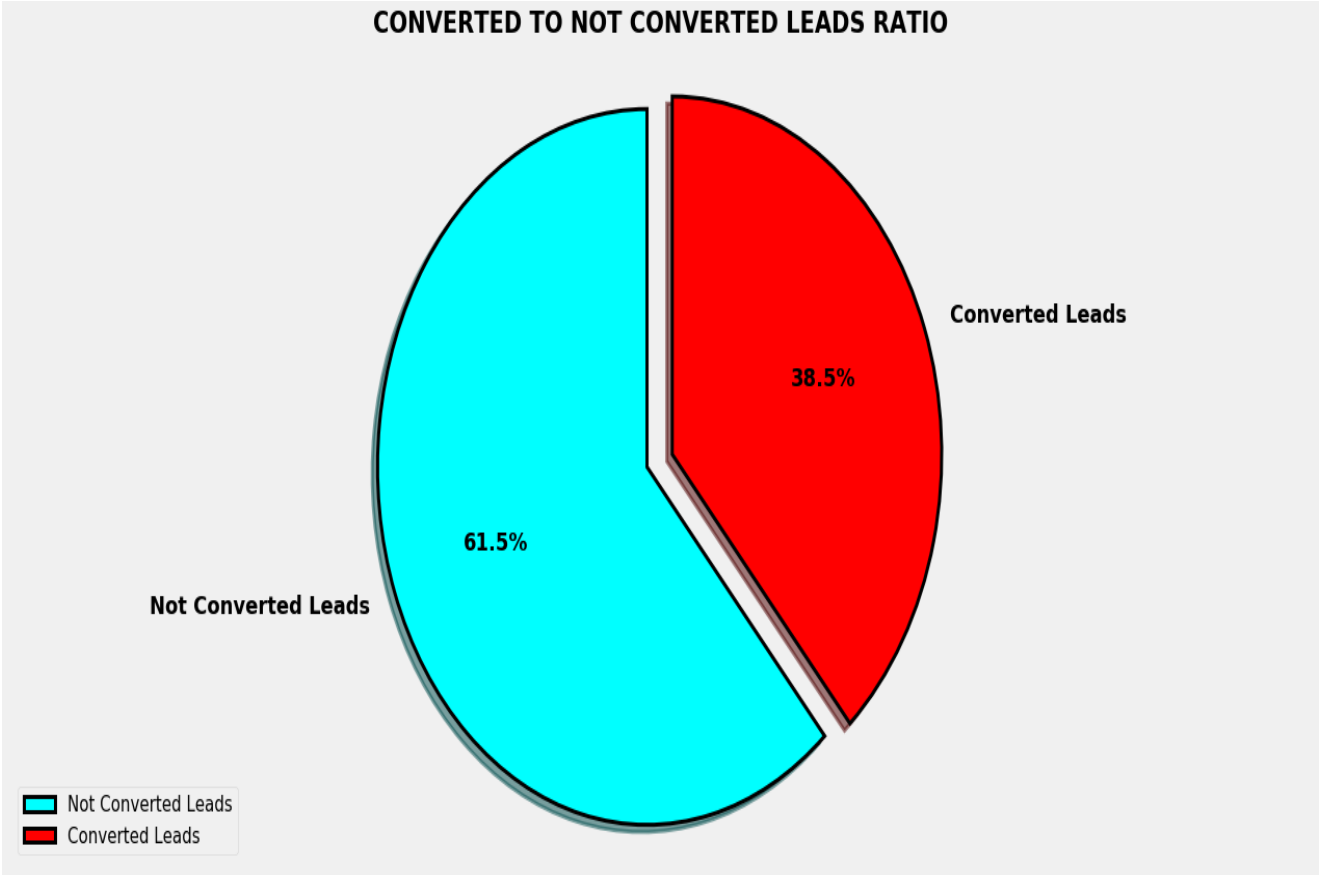
Non-converted leads consists of 61.5%

Outlier Treatment*:

Outlier Treatment is done using IQR concept.

Outliers are removed only from ‘TotalVisits’ & ‘feature Page Views per Visit’

(*NOTE: Outlier treatment is done after missing values imputation. In the ppt it is shown before only because of convenience.)



Data Quality Check – (null values treatment):

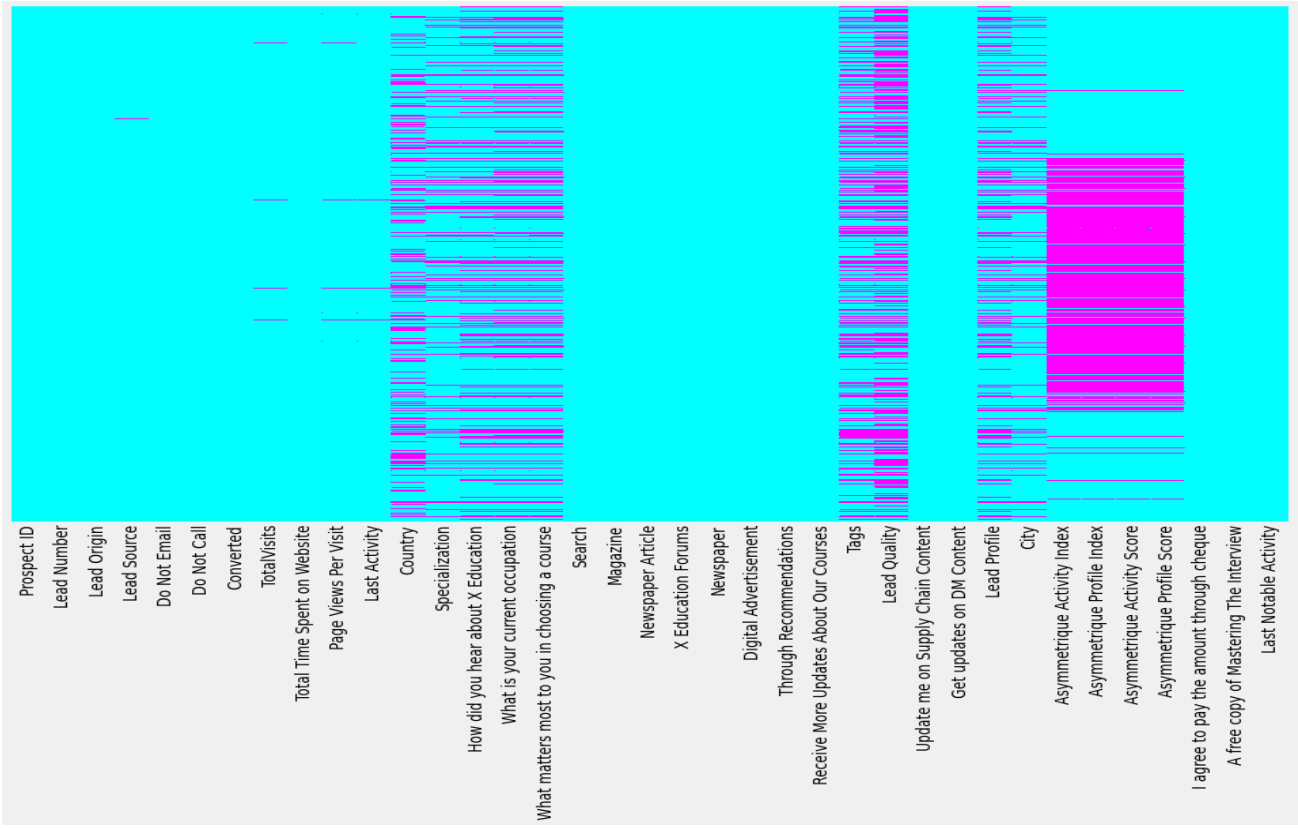
We check for null values in the dataset:

Some columns have the value ‘select’ which is converted into null values.

Columns with more than 40% null values are dropped.

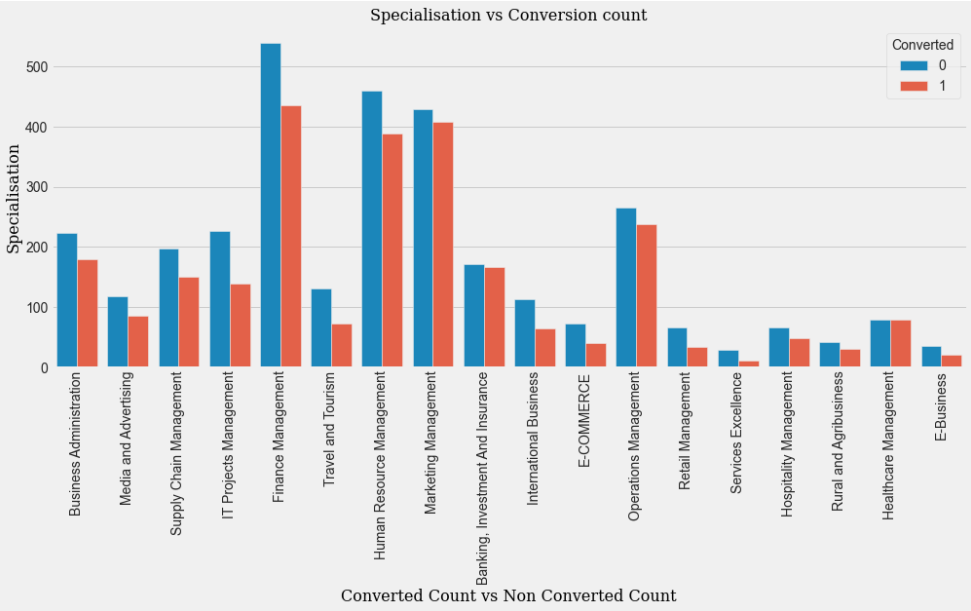
The rest of the columns containing null values are imputed using mean/median/mode.

Few columns like ‘magazine’ are highly skewed and columns like ‘Asymmetric activity Index’ are added by the sales team. Hence are dropped because they don’t influence the lead conversion.

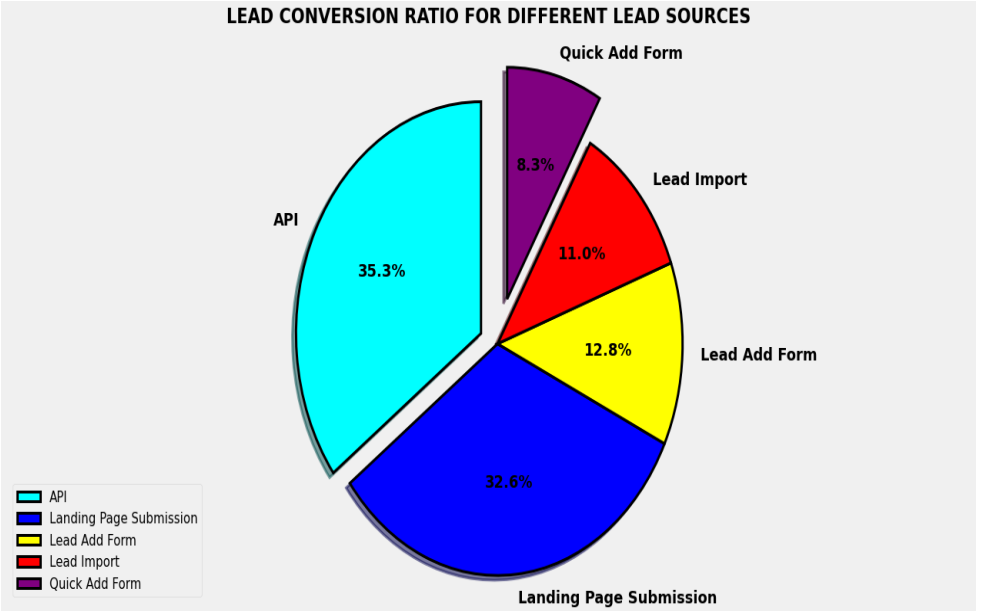


Exploratory Data Analysis:

THIS CHART SHOWS THE CONVERTED & NON-CONVERTED COUNT OF THE COLUMN ‘SPECIALIZATION’

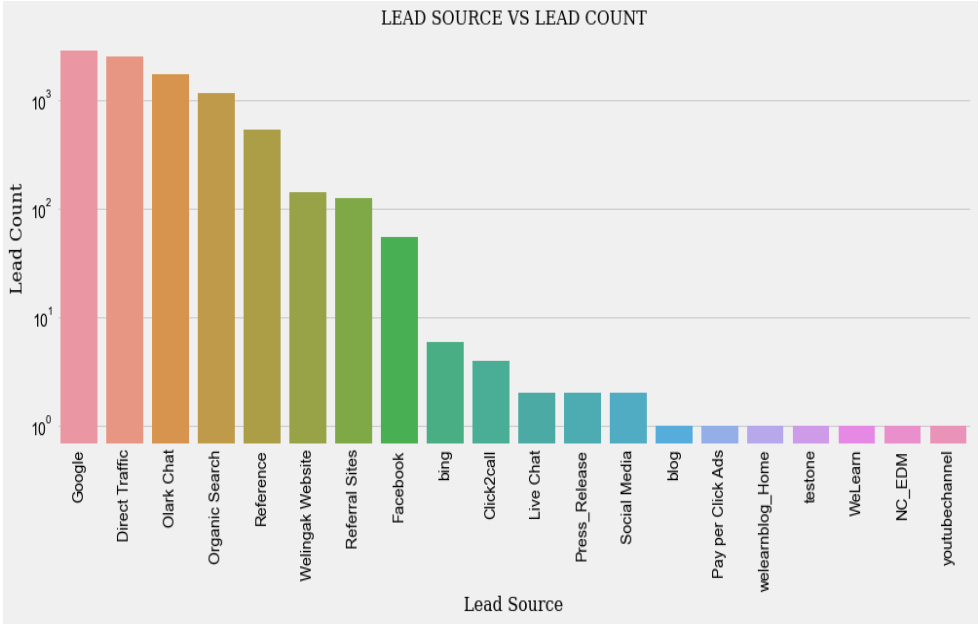


THIS PIE CHART SHOWS THE RATIO OF LEAD CONVERSION FROM DIFFERENT LEAD SOURCE

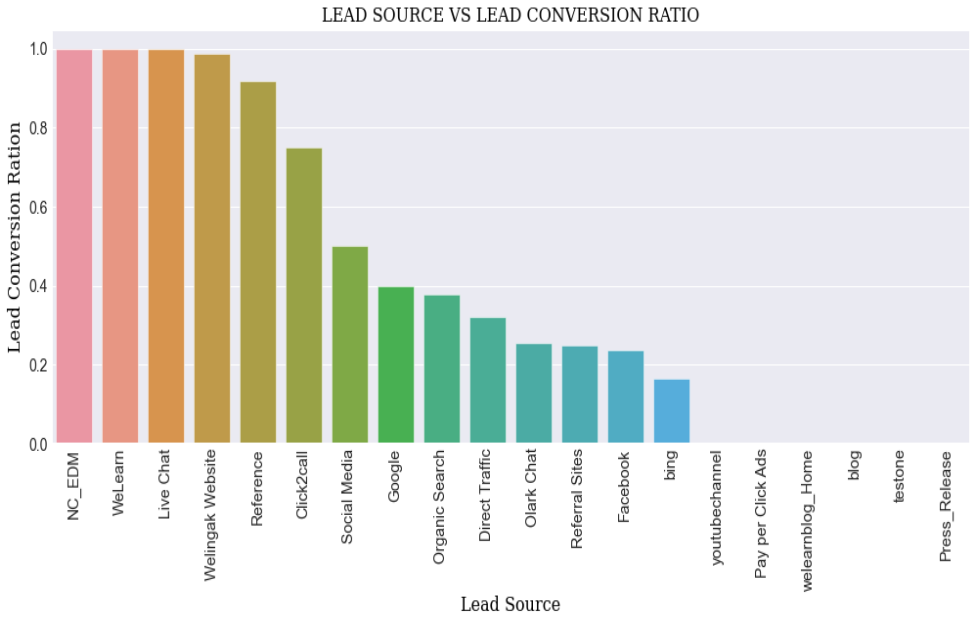


Exploratory Data Analysis:

THIS CHART SHOWS THE COUNT OF LEAD GENERATED FROM DIFFERENT LEAD SOURCE

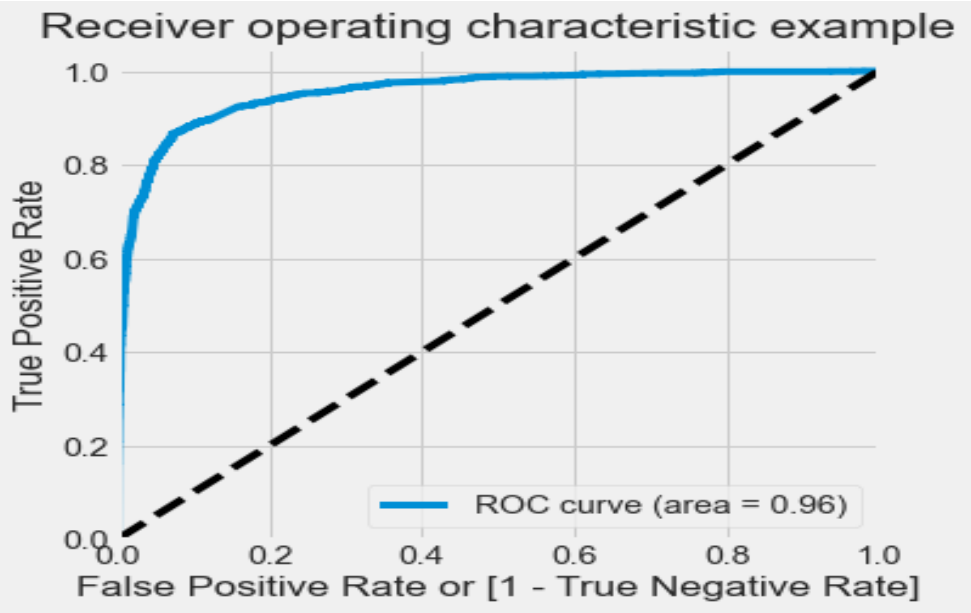


THIS CHART SHOWS THE COUNT OF LEADS CONVERTED FROM THE LEAD SOURCE

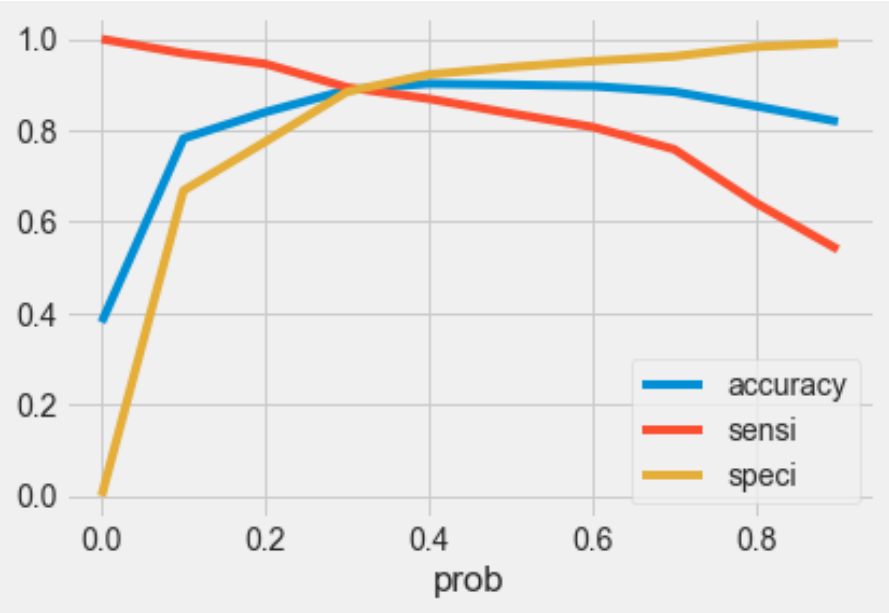


Model Evaluation:

ROC CURVE SHOWS THE TRADE-OFF BETWEEN THE SENSITIVITY AND SPECIFICITY. THE ROC CURVE IS MORE TOWARDS THE UPPER-LEFT CORNER OF THE GRAPH



AT 0.3 WE SEE THAT ACCURACY, SENSITIVITY & SPECIFICITY IS EQUAL. HENCE WE CHOOSE THAT AS THE CUT-OFF POINT



Making Predictions on the test dataset:

PROBABILITY OF THE TRAIN SET BEFORE IMPLEMENTING THE CUT OFF OF 0.3

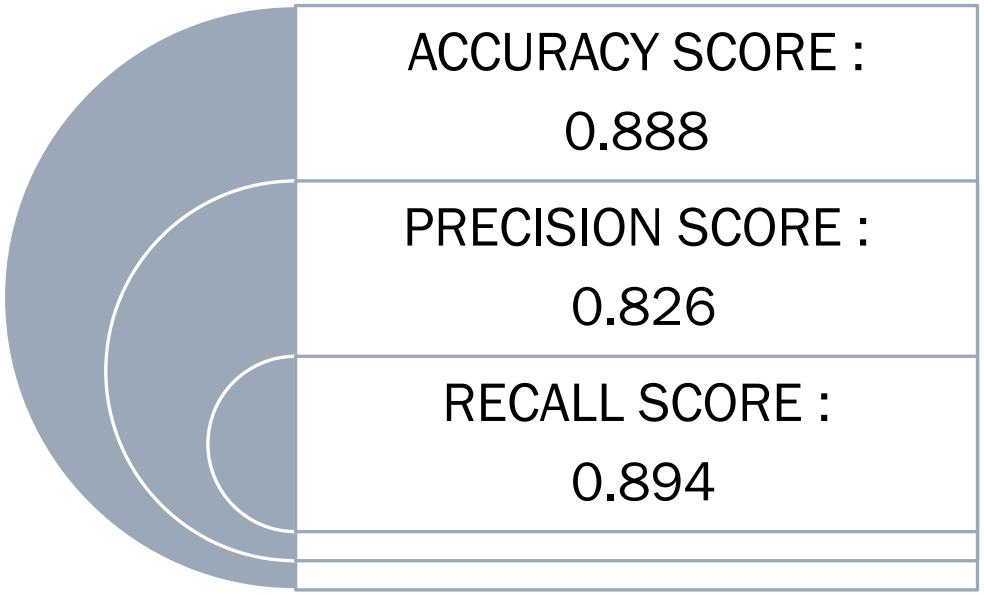
y_train_predicted_probability	y_train	
8741	0.056902	0
6337	0.017866	0
475	0.999138	1
2096	0.668929	1
7953	0.923767	1
...	...	...
350	0.883590	1
79	0.987669	1
8039	0.283058	1
6936	0.018652	0
5640	0.137745	0

MAKING PREDICTIONS ON THE TEST SET

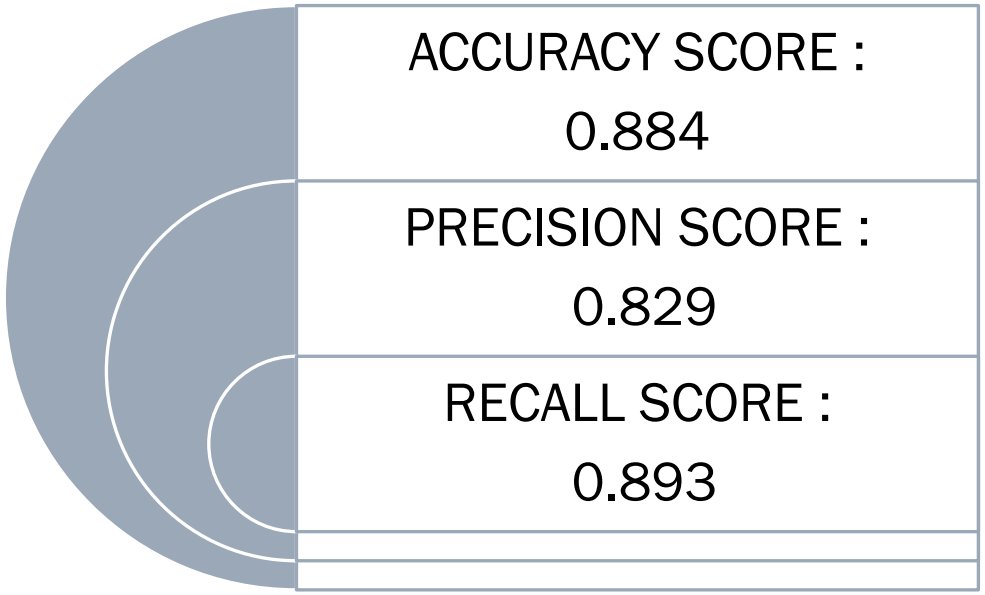
y_test	y_test_probability	y_test_final	
4269	1	0.947052	1
2376	1	0.974617	1
7766	1	0.950887	1
9199	0	0.002333	0
4359	1	0.883590	1
...	...	...	...
8924	0	0.001851	0
2601	1	0.790288	1
7180	0	0.283058	0
3141	0	0.137745	0
1006	0	0.314271	1

Score Results:

TRAIN DATASET



TEST DATASET



Final Dataset:

(note* : lead score is the last column)

	Lead_ID	Prospect_ID	Lead_Score	
0	660737	7927b2df-8bba-4d29-b9a2-b6e0beafe620	0	
1	660728	2a272436-5132-4136-86fa-dcc88c88f482	0	
2	660727	8cc8c611-a219-4f35-ad23-fdfd2656bd8a	95	
3	660719	0cc2df48-7cf4-4e39-9de9-19797f9b38cc	0	
4	660681	3256f628-e534-4826-9d63-4a8b88782852	32	
...	...	...	...	
9235	579564	19d6451e-fcd6-407c-b83b-48e1af805ea9	70	
9236	579546	82a7005b-7196-4d56-95ce-a79f937a158d	9	
9237	579545	aac550fe-a586-452d-8d3c-f1b62c94e02c	2	
9238	579538	5330a7d1-2f2b-4df4-85d6-64ca2f6b95b9	86	
9239	579533	571b5c8e-a5b2-4d57-8574-f2ffb06fdeff		