Brief Summary Report

Lead Scoring Case Study

Strategy

- Data Reading and Understanding
- Data Cleaning
- Data Transformation
- Data Analysis
- Data Preparation
- Building Logistic Regression model and calculating Lead score
- Model Evaluation

1. Data Reading and Understanding

First step was to load the given dataset to the jupyter file and analyze the data like shape of the dataset, datatype of the columns, and some statistical info about the data like mean, mode, media, outliers.

2. Data Cleaning

- It was observed that there were some redundant columns in the dataset that we decided to remove.
- There were some columns that were having a ‘Select’ label which showed that the customer didn’t select any option. It was better to put it as null value because there were no suitable options present to select for the customer searching for.
- Outliers were observed in two columns which were handled by upper capping them due to the nature of the data.
- We removed the columns having missing values more than 30%.
- For the remaining categorical columns having missing values we replaced them using the mode value.

- Two columns had identical names which were taken care of by changing the column name into one format.
- After outlier treatment and further analysis, we decided to impute missing values in the numerical columns by their respective modes due to the nature of data.

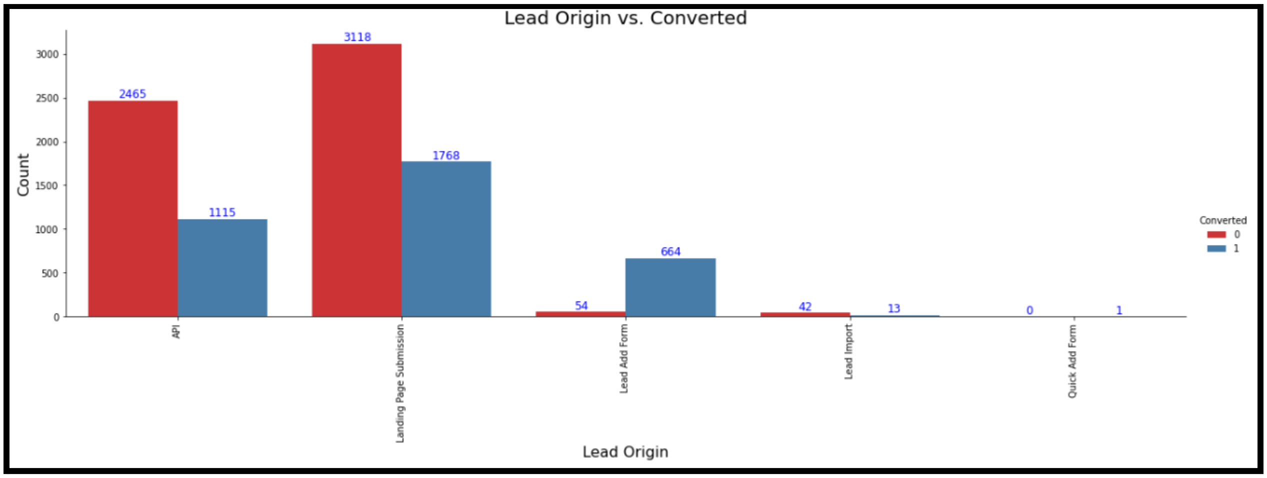
3. Data Transformation

Assigned numerical values (1 and 0) to the columns having data Yes and No.

All the columns were converted to numerical type for further analysis.

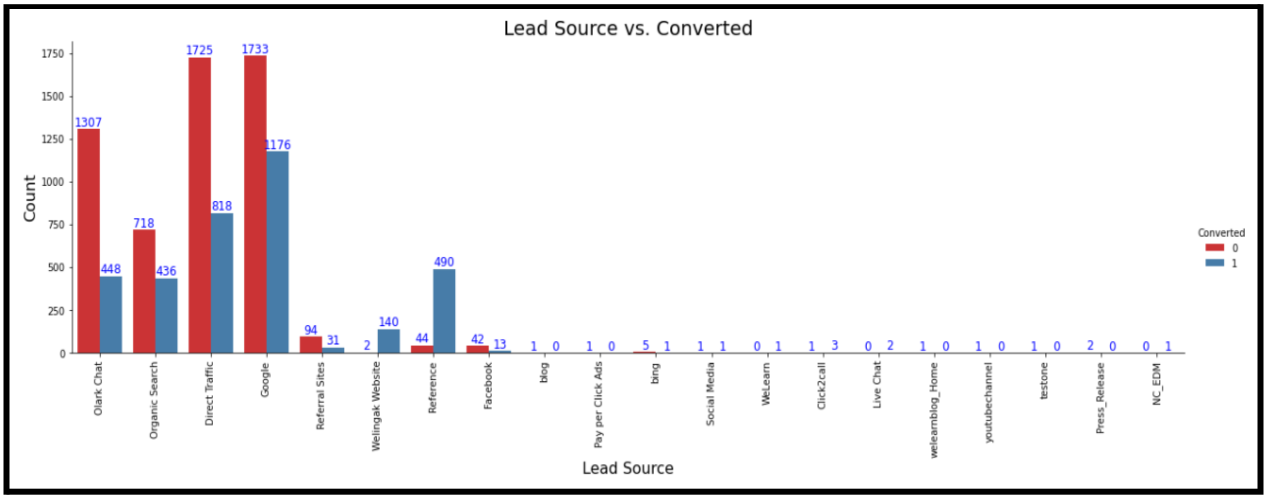
4. Data Analysis

Lead Origin



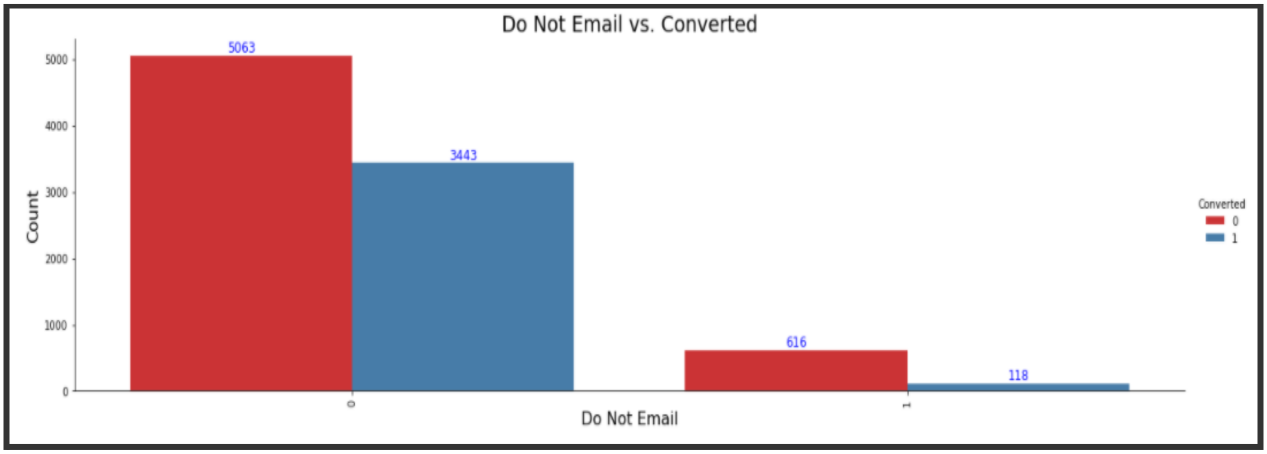
It can be seen that the maximum conversion happened from Landing Page Submission. Also there was only one request from a quick add form which got converted.

Lead Source



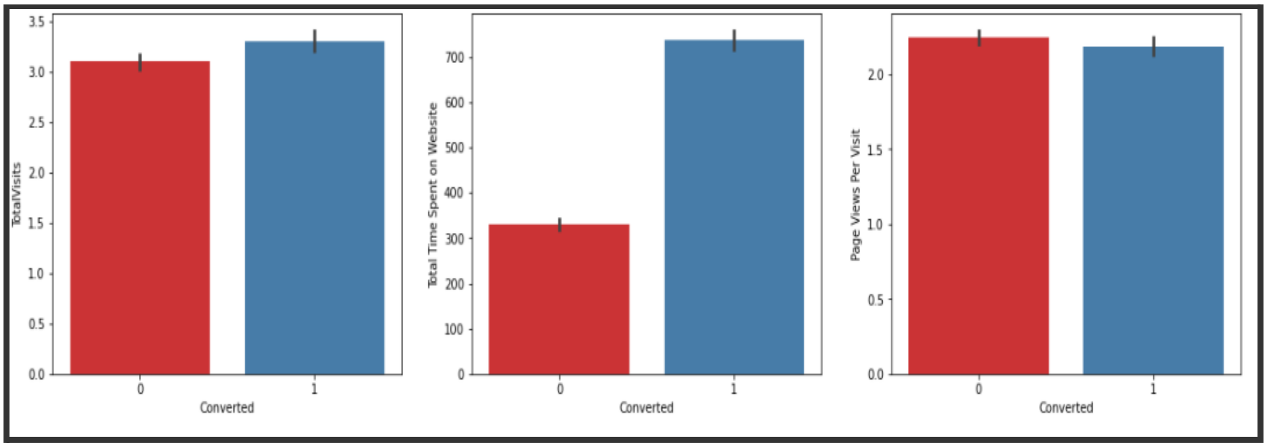
It can be seen that major conversion in the lead source is from Google.

Do Not Email



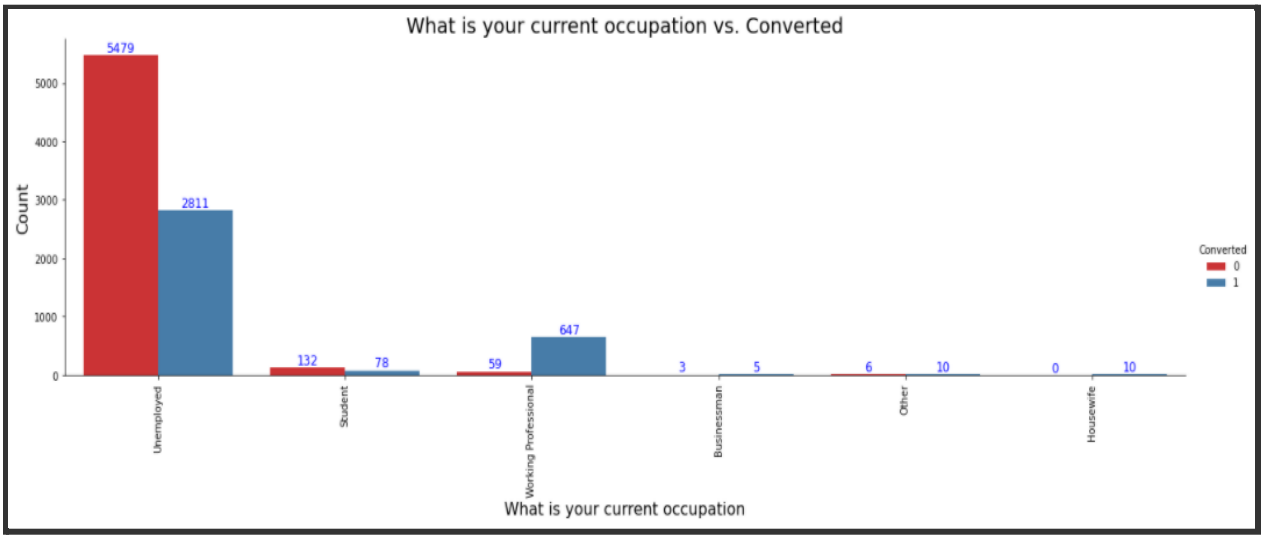
Based on the above graph, major conversion has happened from the emails that have been sent.

Numerical Columns



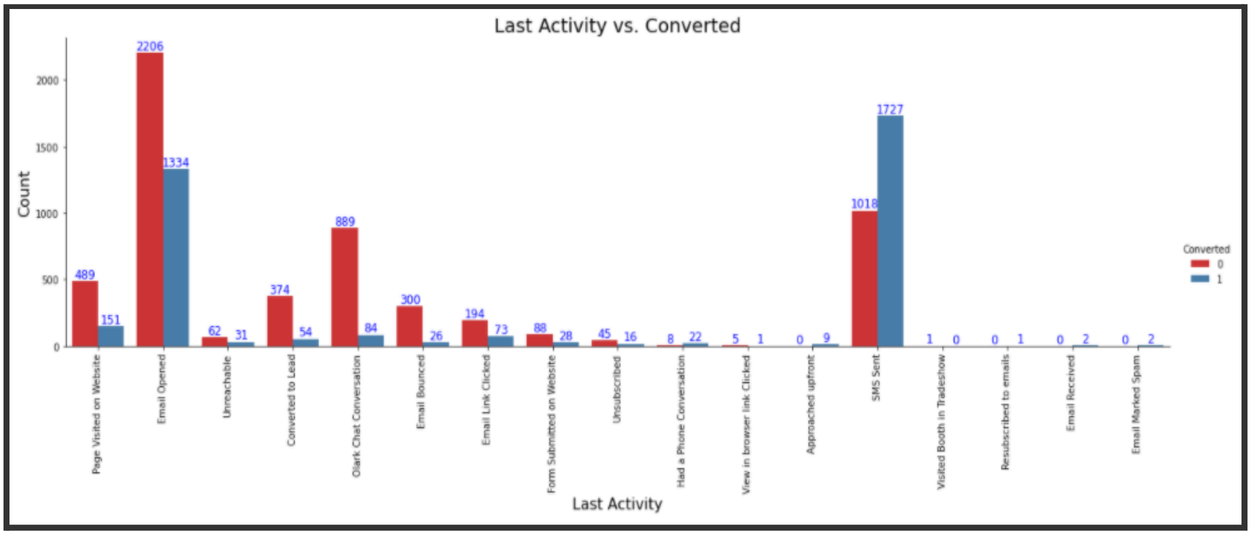
The conversion rates were high for Total Visits, Total Time Spent on Website and Page Views Per Visit.

Current Occupation



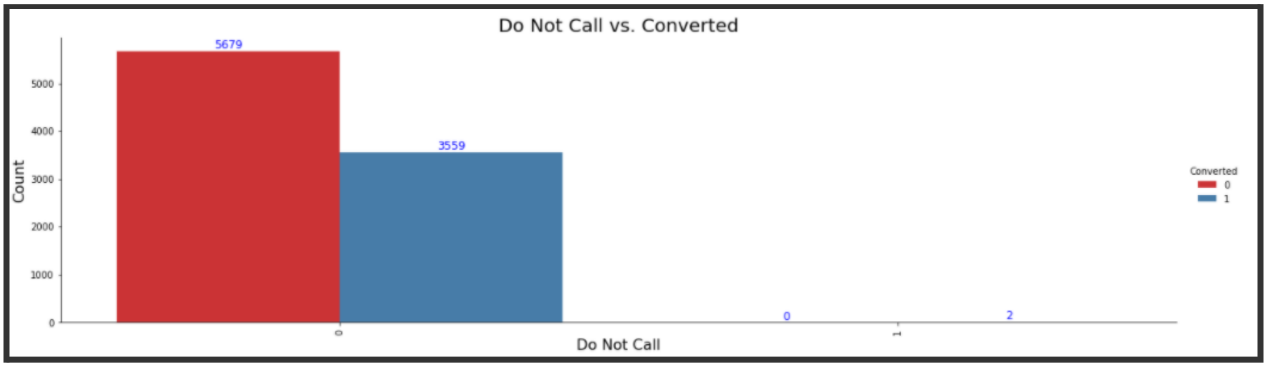
More conversion happened with people who are unemployed. It can also be noticed from the above graph that out of 7 business men, 4 got converted and out 10 of housewives, all 10 leads got converted.

Last Activity



As per the above graph, Last Activity value of 'SMS Sent' had more conversions followed by email opened.

Do Not Call



It can be noticed that major conversions happened when calls were made. However, it can also be seen that 2 leads opted for "Do Not Call", but they still got converted.

5. Data Preparation

Dataset was split into train and test data. It was observed that the overall conversion rate was around 40%.

MinMax scaling of the data was done for further modelling.

6. **Building Logistic Regression Model and calculation of Lead score**

   In the model obtained by logistic regression it was observed that many variables have high p-values. For the feature selection we used RFE as the number of variables are quite high and individually checking them was not efficient.

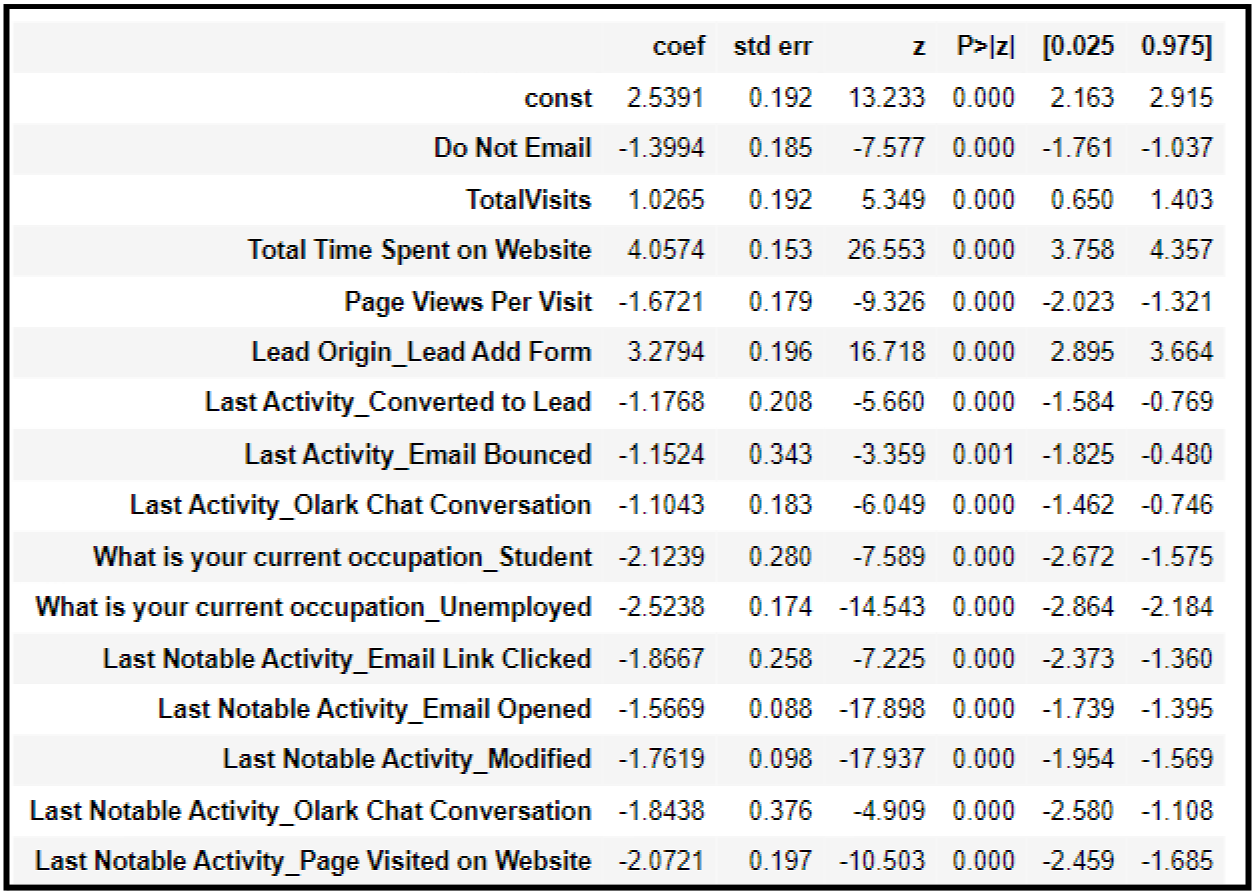
   After RFE was done all the columns based on their ranking were selected and again modelling was done.

   All the features with p-value greater than 0.05 were dropped one by one and the model was built repeatedly.

   All the mentioned features were dropped:

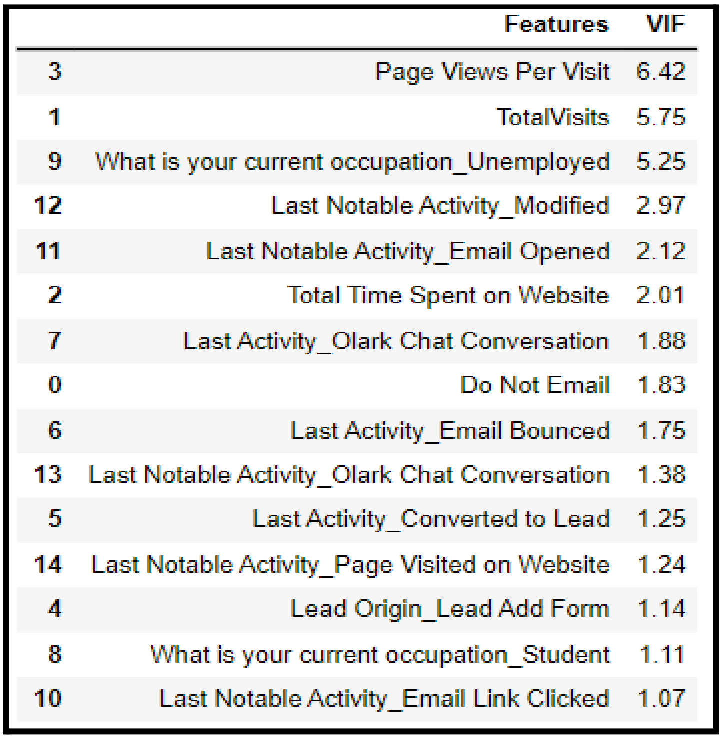
   - 'What matters most to you in choosing a course_Flexibility & Convenience'
   - 'What is your current occupation_Housewife'
   - 'Lead Source_Welingak Website'
   - 'What is your current occupation_Working Professional'
   - 'Last Notable Activity_Had a Phone Conversation’

All the features now had p-value less than 0.05 and the final model obtained was as below:



	coef	std err	z	P>|z|	[0.025	0.975]
const	2.5391	0.192	13.233	0.000	2.163	2.915
Do Not Email	-1.3994	0.185	-7.577	0.000	-1.761	-1.037
TotalVisits	1.0265	0.192	5.349	0.000	0.650	1.403
Total Time Spent on Website	4.0574	0.153	26.553	0.000	3.758	4.357
Page Views Per Visit	-1.6721	0.179	-9.326	0.000	-2.023	-1.321
Lead Origin_Lead Add Form	3.2794	0.196	16.718	0.000	2.895	3.664
Last Activity_Converted to Lead	-1.1768	0.208	-5.660	0.000	-1.584	-0.769
Last Activity_Email Bounced	-1.1524	0.343	-3.359	0.001	-1.825	-0.480
Last Activity_Olark Chat Conversation	-1.1043	0.183	-6.049	0.000	-1.462	-0.746
What is your current occupation_Student	-2.1239	0.280	-7.589	0.000	-2.672	-1.575
What is your current occupation_Unemployed	-2.5238	0.174	-14.543	0.000	-2.864	-2.184
Last Notable Activity_Email Link Clicked	-1.8667	0.258	-7.225	0.000	-2.373	-1.360
Last Notable Activity_Email Opened	-1.5669	0.088	-17.898	0.000	-1.739	-1.395
Last Notable Activity_Modified	-1.7619	0.098	-17.937	0.000	-1.954	-1.569
Last Notable Activity_Olark Chat Conversation	-1.8438	0.376	-4.909	0.000	-2.580	-1.108
Last Notable Activity_Page Visited on Website	-2.0721	0.197	-10.503	0.000	-2.459	-1.685

After this we checked the model for multicollinearity issues by VIF values.



	Features	VIF
3	Page Views Per Visit	6.42
1	TotalVisits	5.75
9	What is your current occupation_Unemployed	5.25
12	Last Notable Activity_Modified	2.97
11	Last Notable Activity_Email Opened	2.12
2	Total Time Spent on Website	2.01
7	Last Activity_Olark Chat Conversation	1.88
0	Do Not Email	1.83
6	Last Activity_Email Bounced	1.75
13	Last Notable Activity_Olark Chat Conversation	1.38
5	Last Activity_Converted to Lead	1.25
14	Last Notable Activity_Page Visited on Website	1.24
4	Lead Origin_Lead Add Form	1.14
8	What is your current occupation_Student	1.11
10	Last Notable Activity_Email Link Clicked	1.07

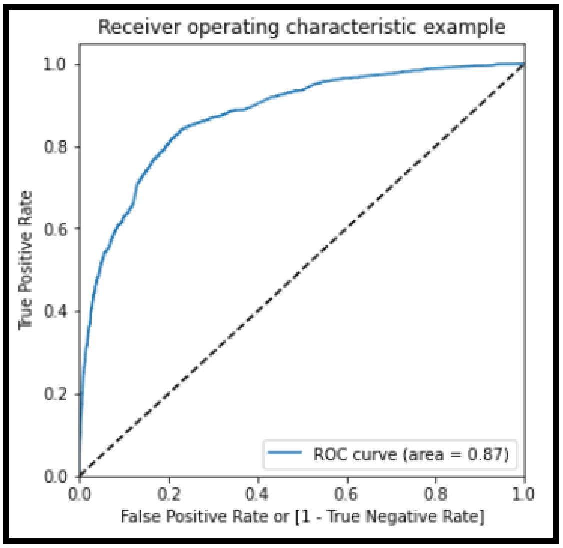
All the variables had VIF values of less than 7. Hence, there were no major multicollinearity issues. We did not remove 'Page Views Per Visit' 'Total Visits' 'What is your current occupation_Unemployed' because,

- from business perspective these variables might be important
- we might run the risk of overfitting the model by removing them
- their VIF values are close to 5

Hence, we finalised this model and used this for further analysis and predictions.

Further ROC curve was plotted to check the stability of the model and optimal cutoff point was calculated keeping accuracy, sensitivity and specificity in mind.

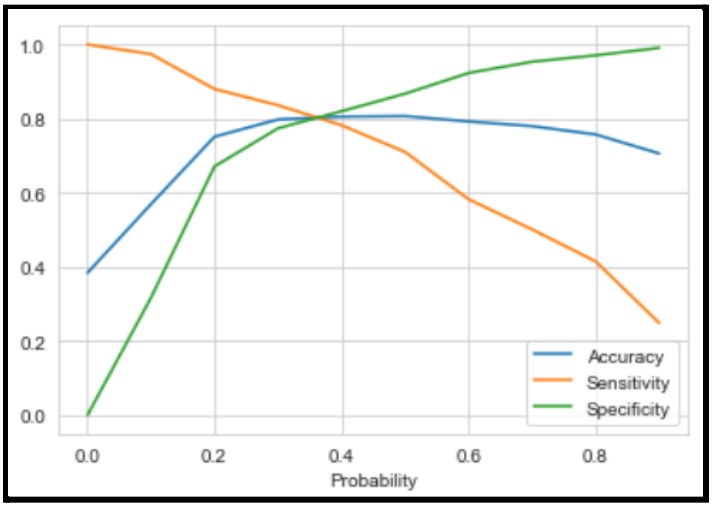
ROC Curve



The curve is closer to the left side of the border than to the right side hence our model is having great accuracy.

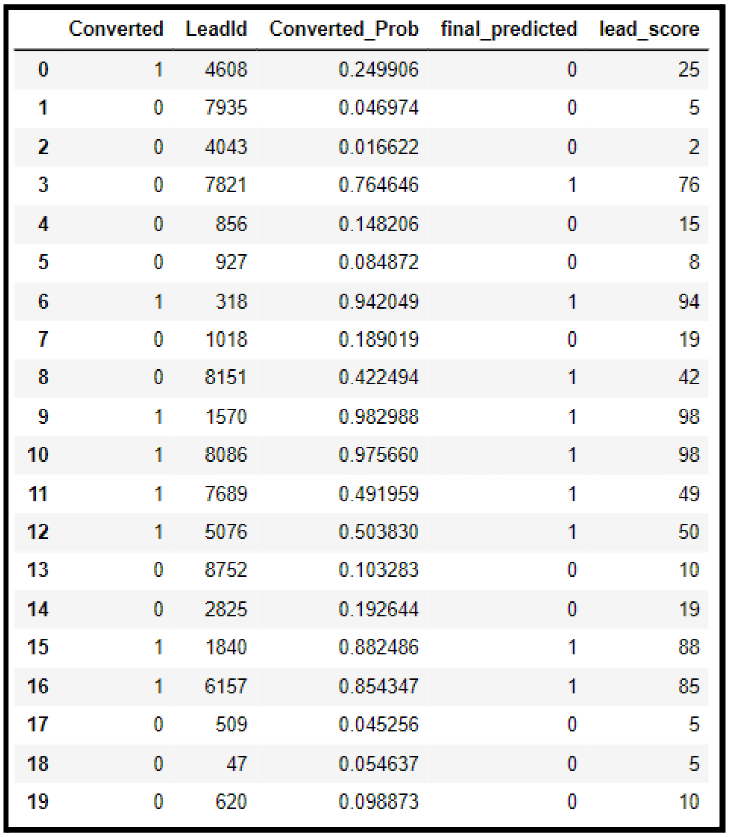
The area under the curve is 87% of the total area.

Probability Cutoff Point



Probability cutoff point was at around 0.37 as this is where the sensitivity, accuracy and specificity converged.

Based on the cutoff point Lead score was calculated and assigned to the train and test data.



	Converted	LeadId	Converted_Prob	final_predicted	lead_score
0	1	4608	0.249906	0	25
1	0	7935	0.046974	0	5
2	0	4043	0.016622	0	2
3	0	7821	0.764646	1	76
4	0	856	0.148206	0	15
5	0	927	0.084872	0	8
6	1	318	0.942049	1	94
7	0	1018	0.189019	0	19
8	0	8151	0.422494	1	42
9	1	1570	0.982988	1	98
10	1	8086	0.975660	1	98
11	1	7689	0.491959	1	49
12	1	5076	0.503830	1	50
13	0	8752	0.103283	0	10
14	0	2825	0.192644	0	19
15	1	1840	0.882486	1	88
16	1	6157	0.854347	1	85
17	0	509	0.045256	0	5
18	0	47	0.054637	0	5
19	0	620	0.098873	0	10

Predictions were then made using the model on train and test data. These predictions were then evaluated. The evaluations have been summarised in the below table.

Evaluation Parameter	**Train data**	**Test data**
Accuracy	80.7%	81.63%
Sensitivity	80.07%	83.56%
Specificity	80.47%	80.41%
Precision	76.96%	73.05%
Recall	71.01%	83.56%

As we can see the evaluations were coherent between the train and test data. The model was found to be stable and sufficiently accurate.
Following conclusions were derived from the analysis.

Conclusions

- We have successfully run the model on test and train data and evaluated the results for both of them. The results are coherent and the models seem to be accurate up to 80% on train data and 82% on test data.
- Precision and Recall tradeoff has been evaluated and shown by a plot with meeting point at 0.45.
- We have considered a probability cutoff point of 0.37 as this is where the values of accuracy, sensitivity and specificity converged.
- Accuracy, Sensitivity and Specificity values of the test set are around 81%, 79% and 82% which are approximately closer to the respective values calculated using the trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80% as desired.
- In business terms, this model has an ability to adjust with the company's requirements in the coming future.

This concludes that the model is in a stable state and we can successfully draw business related conclusions from it.

Important features responsible for good conversion rate or the ones' which contribute the most towards the probability of a lead getting converted are:

1. Total Time Spent on Website (coeff.= 4.0574)
2. Lead Origin_Lead Add Form (coeff.= 3.2794) and
3. What is your current occupation_Unemployed (coeff.= - 2.5238)