Case Study:

Lead Scoring (Logistic Regression)

AVISHEK BANERJEE
SMRUTI RANJAN PARIDA

Problem Statement

- **An education company X** markets its courses on several websites and search engines like Google. Once people land on the website, they might browse the courses or fill up a form 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**.

- 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%**.

- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. To make this process more efficient, the company wishes to identify the most potential leads, also known as **'Hot Leads'**.

- **The objective** is 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.

- Available data:
  Leads dataset from the past with around 9000 data points. This dataset consists of various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

Summary of Dataset

count	Value Counts	% Missing data	Unique values	Top Value	Frequency
Prospect ID	9240	0.0	9240	fc751588-d915-439a-adb6-ab527374f4c4	1
Lead Origin	9240	0.0	5	Landing Page Submission	4886
Lead Source	9204	0.4	21	Google	2868
Do Not Email	9240	0.0	2	No	8506
Do Not Call	9240	0.0	2	No	9238
Last Activity	9137	1.1	17	Email Opened	3437
Country	6779	26.6	38	India	6492
Specialization	5860	36.6	18	Finance Management	976
How did you hear about X Education	1990	78.5	9	Online Search	808
What is your current occupation	6550	29.1	6	Unemployed	5600
What matters most to you in choosing a course	6531	29.3	3	Better Career Prospects	6528
Search	9240	0.0	2	No	9226
Magazine	9240	0.0	1	No	9240
Newspaper Article	9240	0.0	2	No	9238
X Education Forums	9240	0.0	2	No	9239
Newspaper	9240	0.0	2	No	9239
Digital Advertisement	9240	0.0	2	No	9236
Through Recommendations	9240	0.0	2	No	9233
Receive More Updates About Our Courses	9240	0.0	1	No	9240
Tags	5887	36.3	26	Will revert after reading the email	2072
Lead Quality	4473	51.6	5	Might be	1560
Update me on Supply Chain Content	9240	0.0	1	No	9240
Get updates on DM Content	9240	0.0	1	No	9240
Lead Profile	2385	74.2	5	Potential Lead	1613
City	5571	39.7	6	Mumbai	3222
Asymmetrique Activity Index	5022	45.6	3	02.Medium	3839
Asymmetrique Profile Index	5022	45.6	3	02.Medium	2788
I agree to pay the amount through cheque	9240	0.0	1	No	9240
A free copy of Mastering The Interview	9240	0.0	2	No	6352
Last Notable Activity	9240	0.0	16	Modified	3407

Column	No of Values	% Missing data	mean	std	min	0.25	0.5	0.75	max
Lead Number	9240	0.0	617188	23406	579533	596484	615479	637387	660737
Converted	9240	0.0	0.38	0.49	0	0.0	0	1	1.0
TotalVisits	9103	1.5	3.44	4.85	0	1.0	3	5	251.0
Total Time Spent on Website	9240	0.0	487.70	548.02	0	12.0	248	936	2272.0
Page Views Per Visit	9103	1.5	2.36	2.16	0	1.0	2	3	55.0
Asymmetrique Activity Score	5022	45.6	14.30	1.38	7	14.0	14	15	18.0
Asymmetrique Profile Score	5022	45.6	16.34	1.8114	11	15.0	16	18	20.0

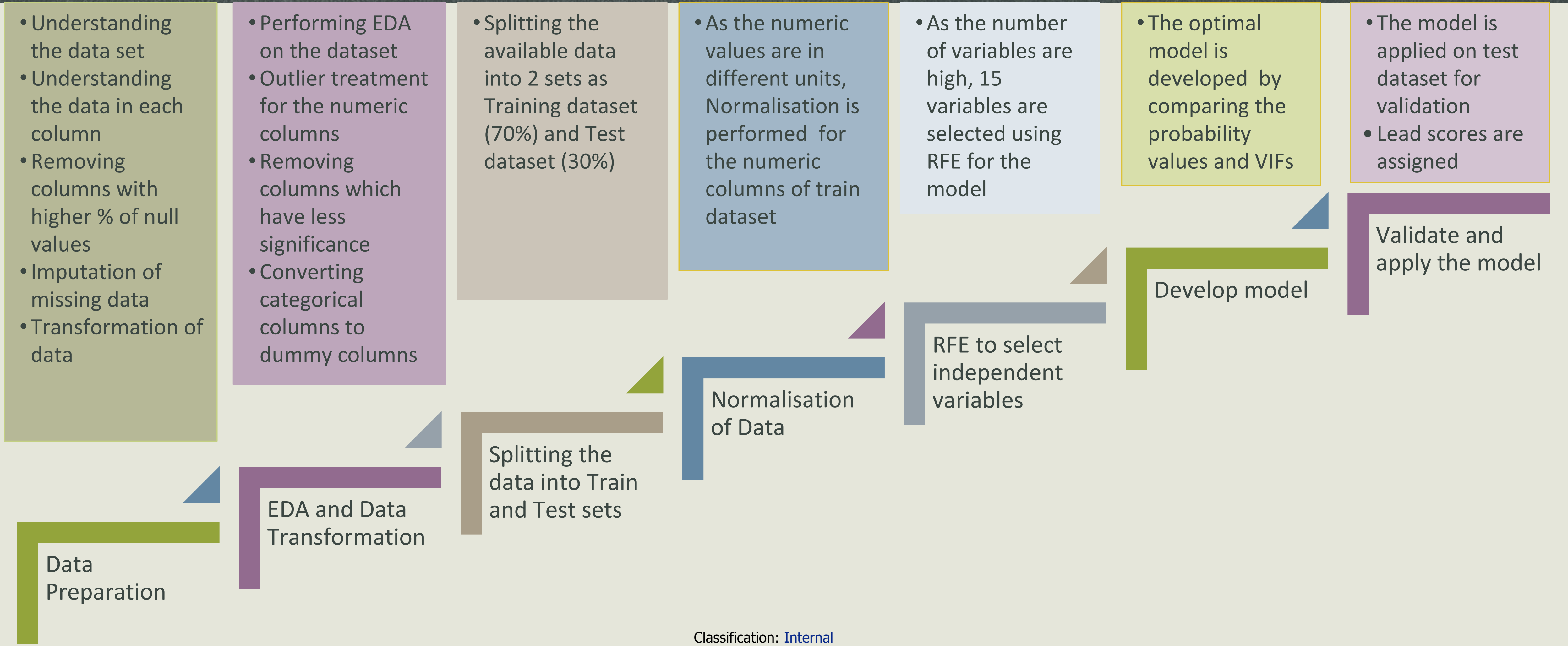
Columns with high % of missing data:

- How did you hear about X Education
- Lead Profile
- Lead Quality
- Asymmetrique Activity Index & Asymmetrique Activity Score
- Asymmetrique Profile Index & Asymmetrique Profile Score

Columns with same value in all the rows:

- 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

Solution Approach



Data Preparation

- Understanding the data set
- Understanding the data in each column
- Removing columns with higher % of null values
- Imputation of missing data
- Transformation of data

EDA and Data Transformation

- Performing EDA on the dataset
- Outlier treatment for the numeric columns
- Removing columns which have less significance
- Converting categorical columns to dummy columns

Splitting the data into Train and Test sets

- Splitting the available data into 2 sets as Training dataset (70%) and Test dataset (30%)

Normalisation of Data

- As the numeric values are in different units, Normalisation is performed for the numeric columns of train dataset

RFE to select independent variables

- As the number of variables are high, 15 variables are selected using RFE for the model

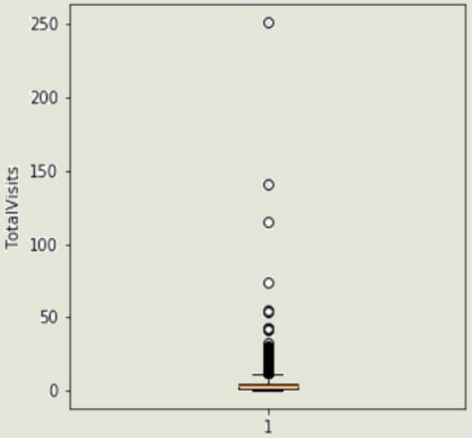
Develop model

- The optimal model is developed by comparing the probability values and VIFs

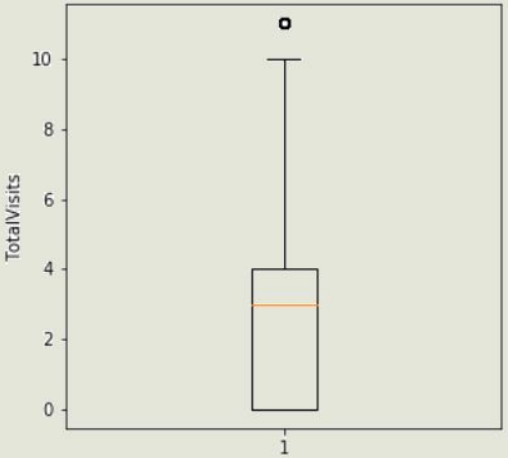
Validate and apply the model

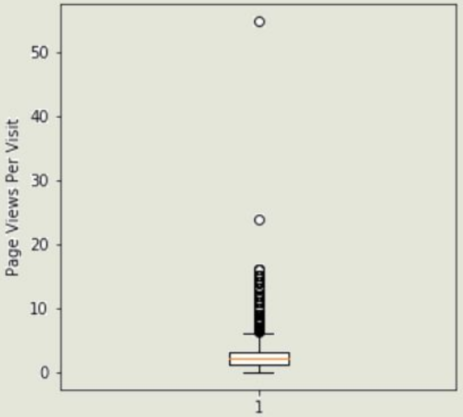
- The model is applied on test dataset for validation
- Lead scores are assigned

EDA: Outliers in Data

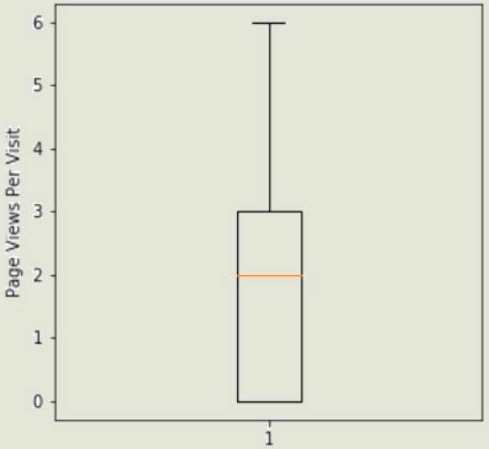


The parameters Total Visits and Page Views seem to have Outliers. The same were removed using IQR.

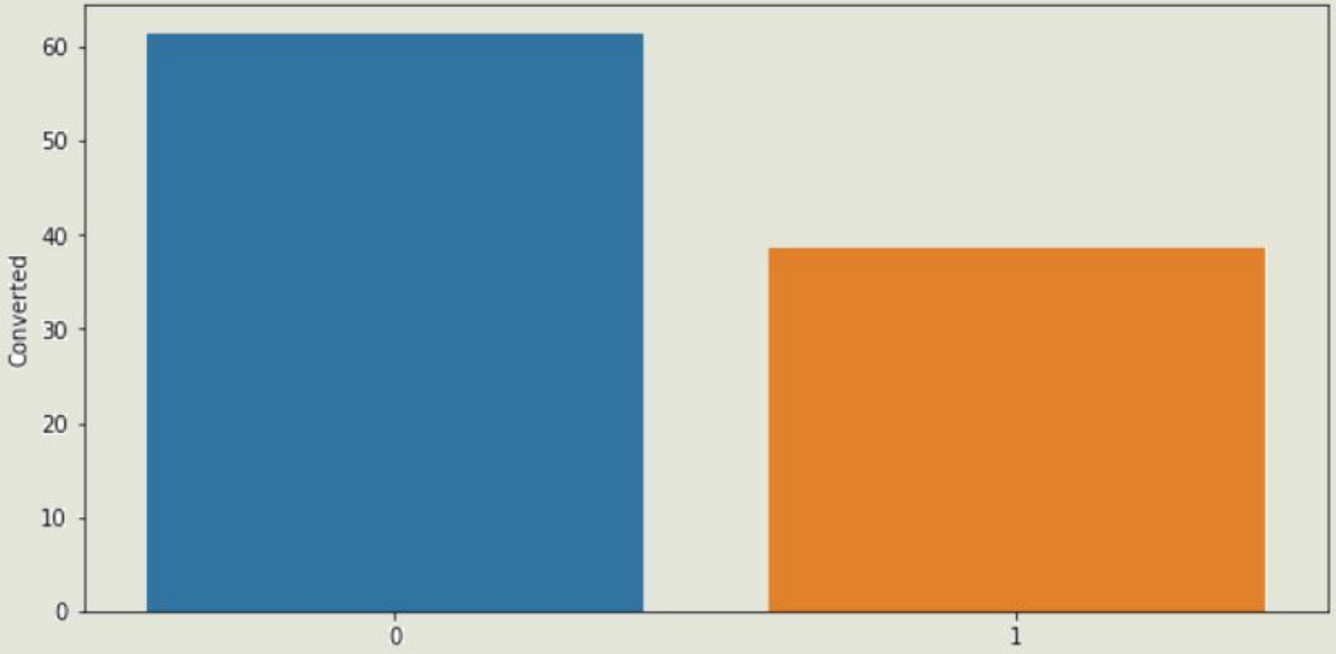




After Outlier Treatment

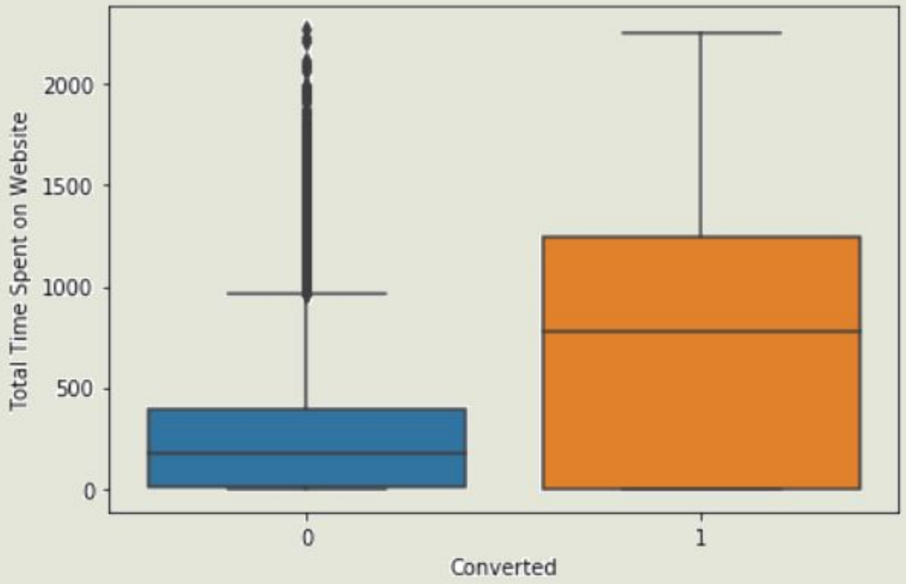


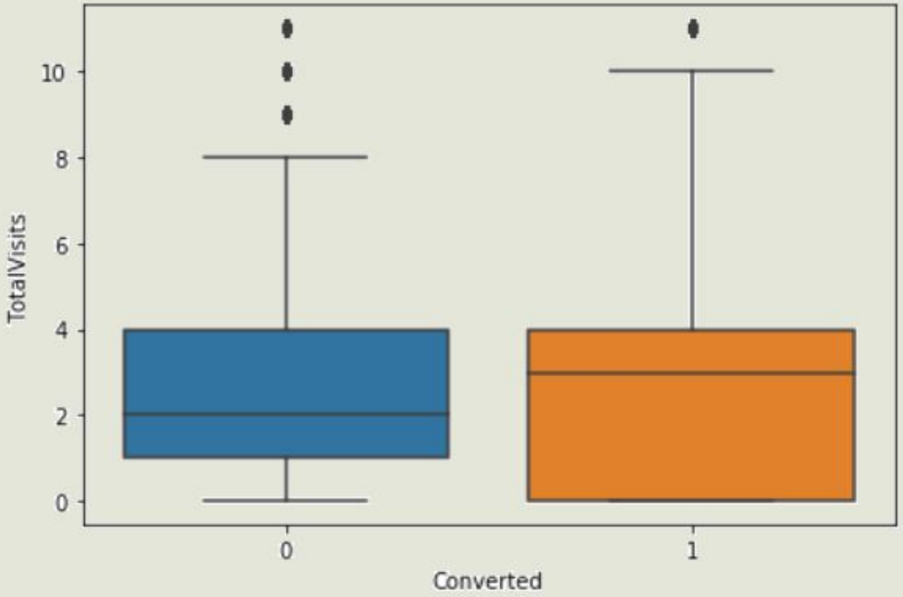
EDA: Univariate Analysis



Current conversion rate:
~ 37%

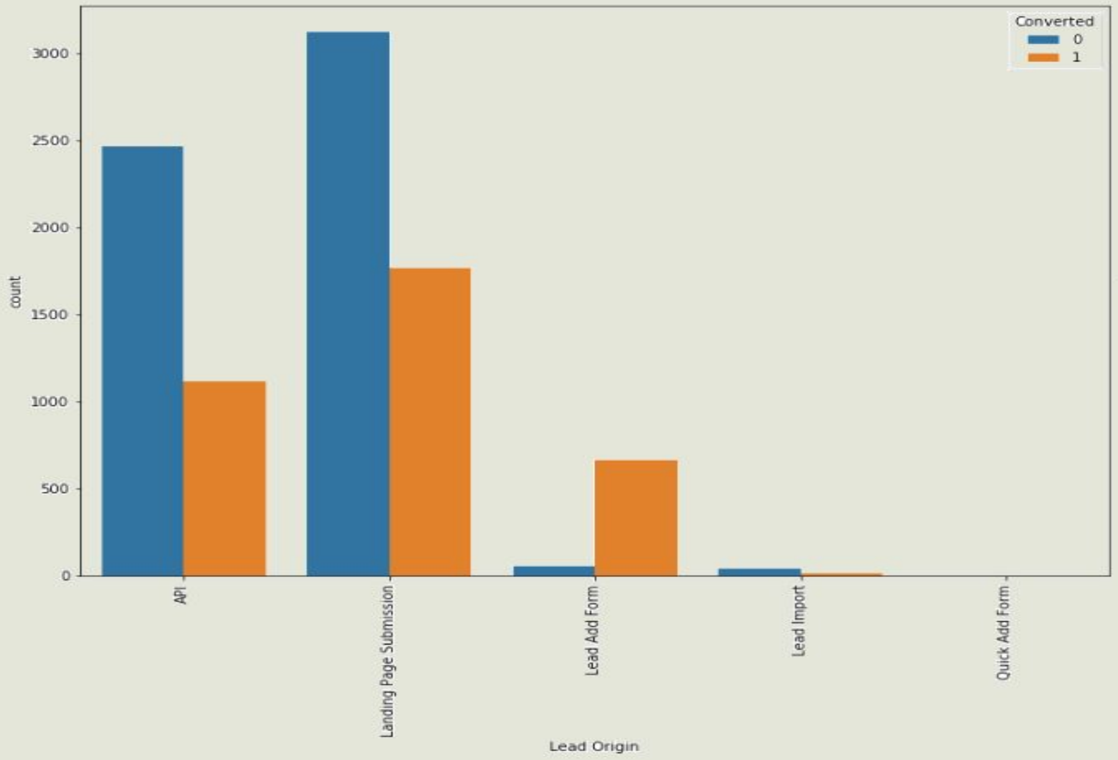
EDA: Univariate Analysis

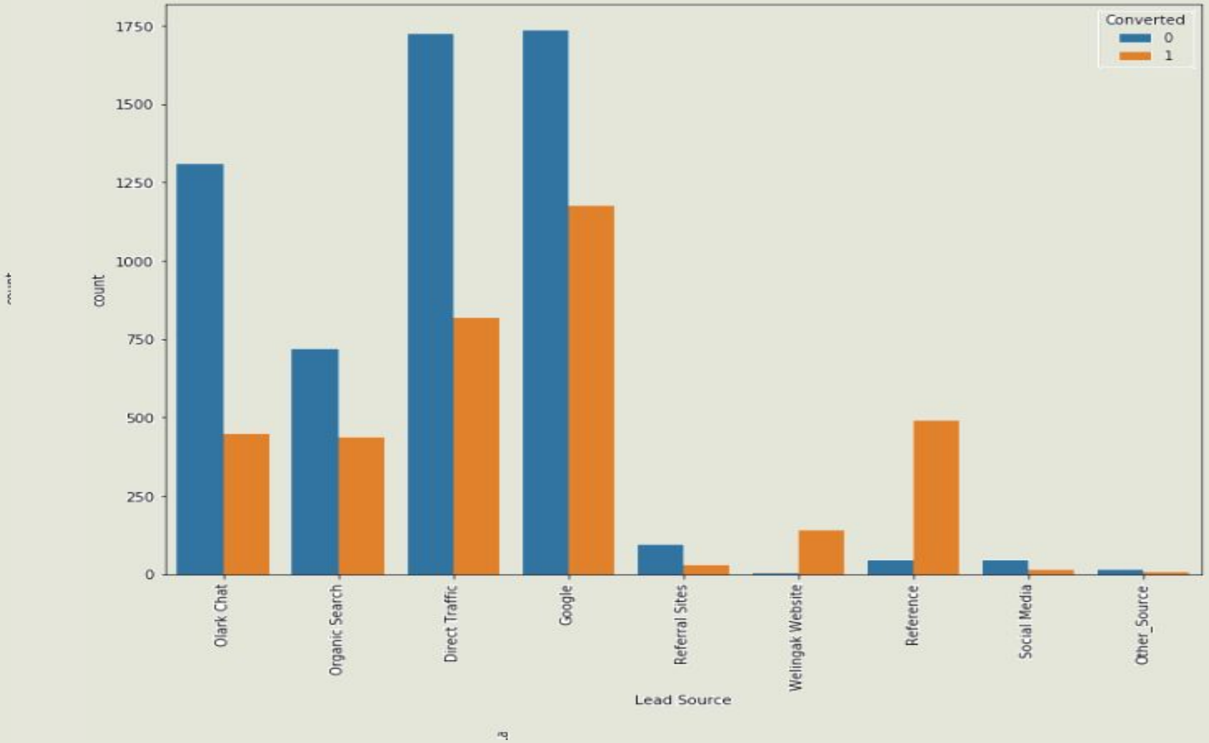




Time Spent on Website: The median of Time Spent for converted leads is much higher than the non-converted ones.
Total Visits: The median of Total Visits for the converted leads is higher than the non-converted ones.

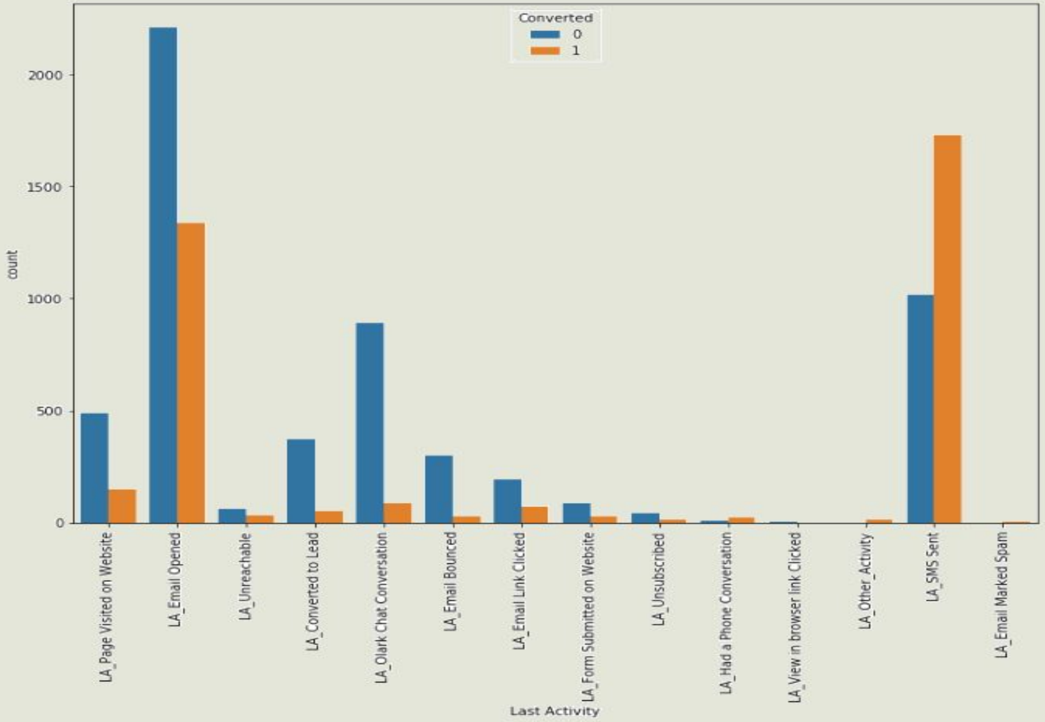
EDA: Univariate Analysis

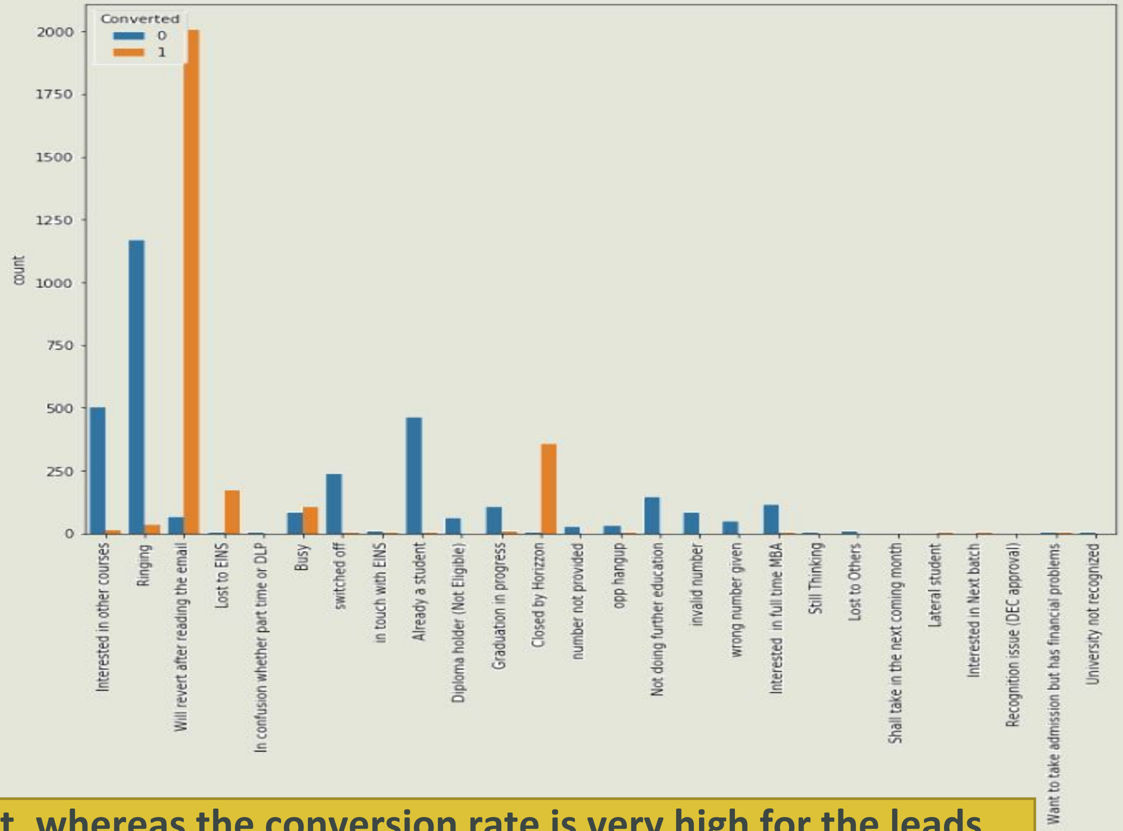




Lead Origin: Most of the leads are coming through “API” or “Landing Page Submission”. The conversion rate for “Lead Import” is high.
Lead Source: The conversion rates are very high for leads sourced from “Reference” and “Welingak Website” (though data points are less).

EDA: Univariate Analysis

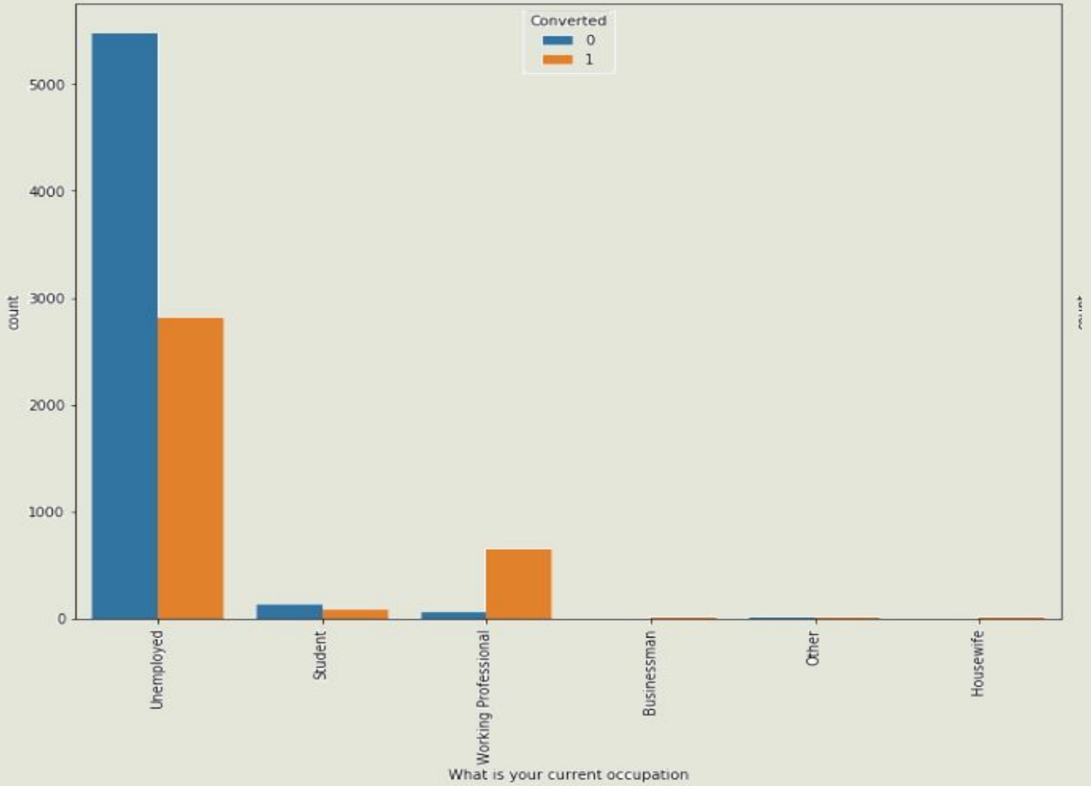


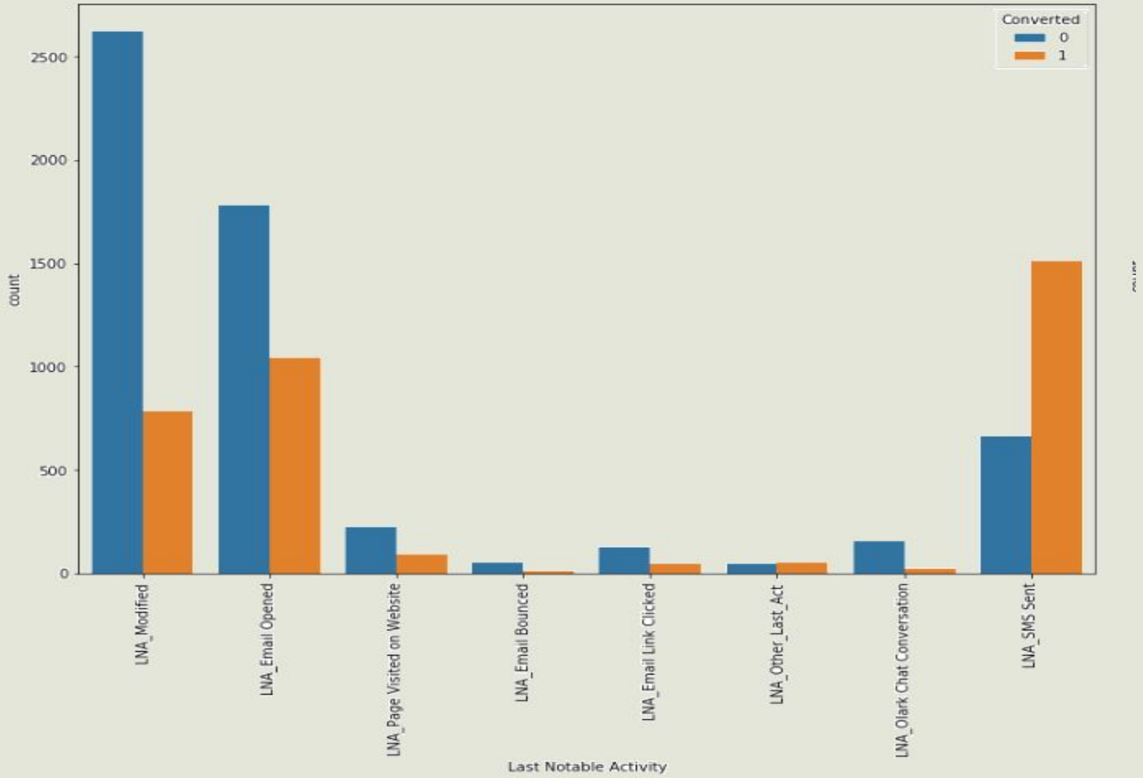


Last Activity: The leads with last activity as "Email Opened" are the highest, whereas the conversion rate is very high for the leads with last activity as "SMS Sent".

Tags: The leads with Tag as "Will revert back after reading the email" have significantly higher conversion rate. The leads with Tag as "Closed by Horizon" seem to have already converted.

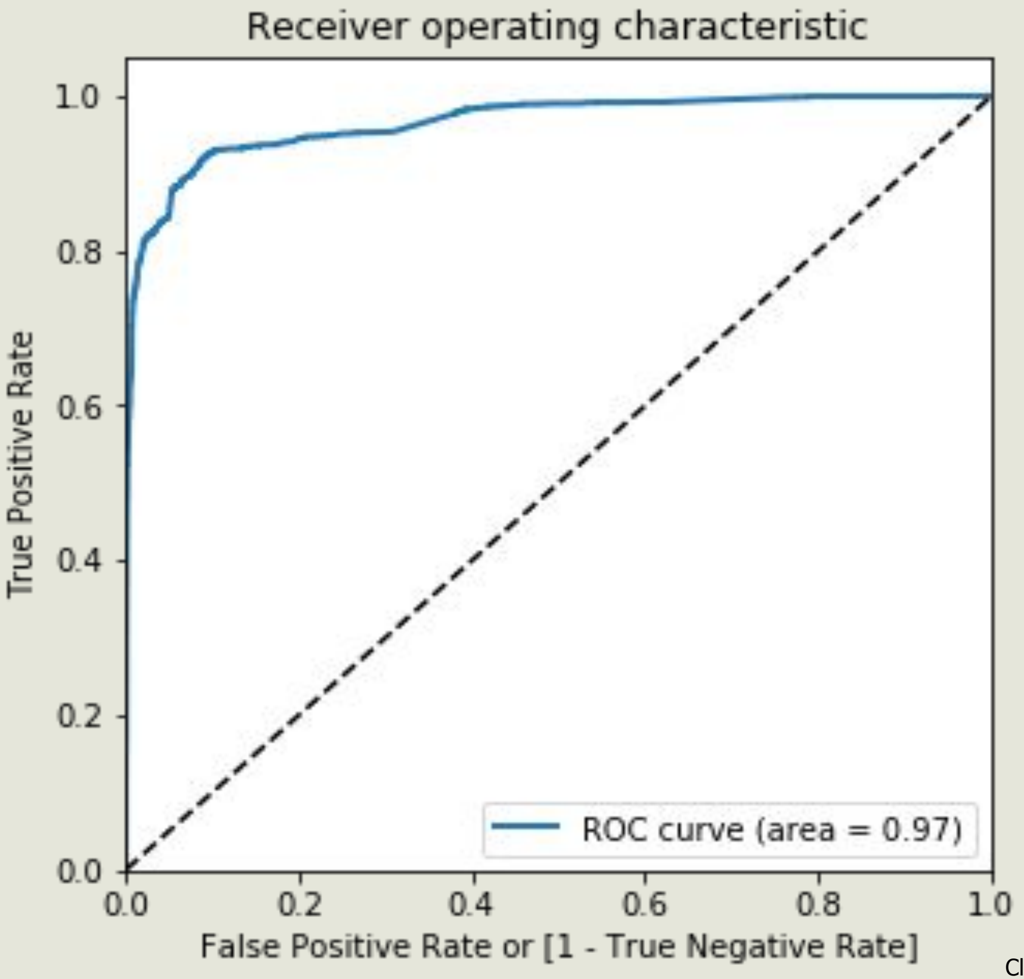
EDA: Univariate Analysis





Occupation: The conversion rate is very high for “Working Professional”.
Last Notable Activity: The leads with last notable activity as “SMS Sent" have significantly higher conversion rate.

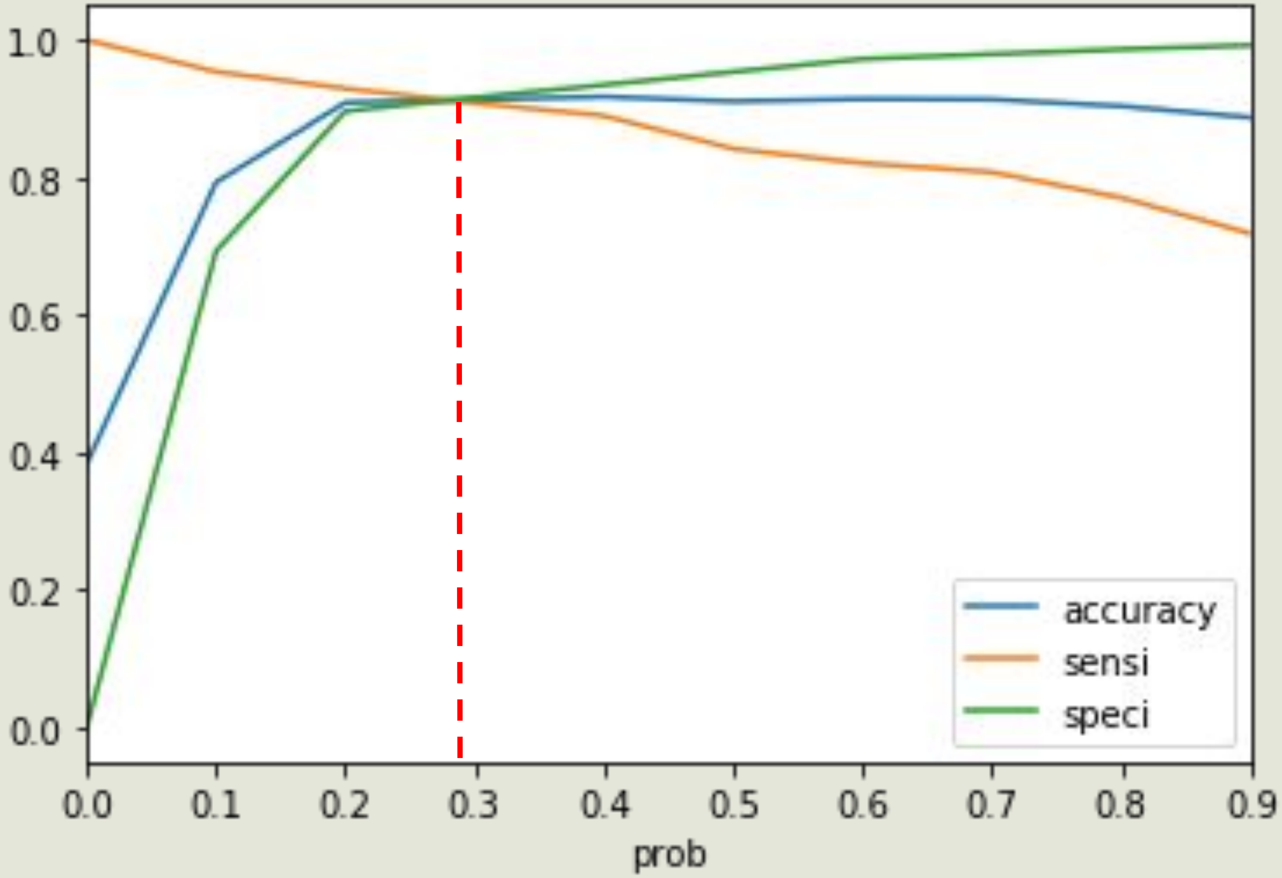
Logistic Regression Model: ROC Curve



The area under the curve is 97%.

Classification: Internal

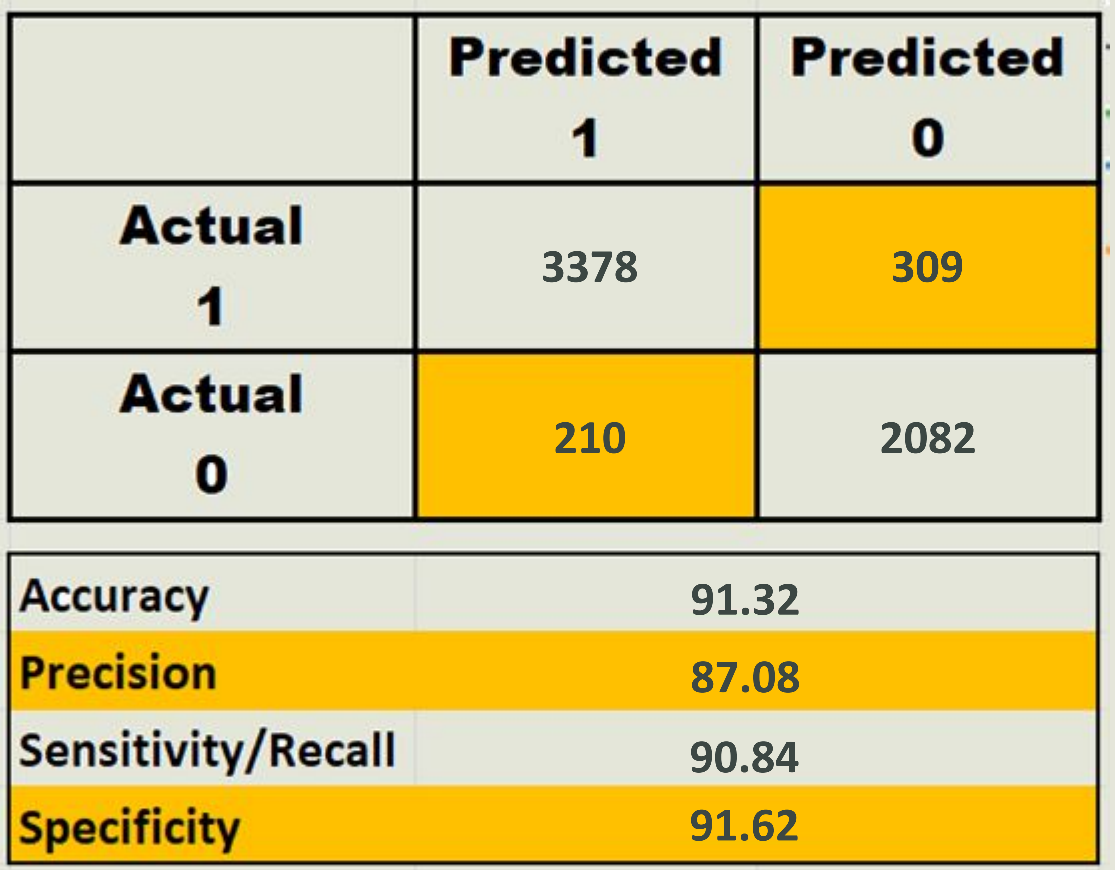
Logistic Regression Model:
Accuracy, Sensitivity and Specificity for various probabilities



At cutoff probability of 0.3, the model seems to have Good Accuracy, Sensitivity and Specificity.

Logistic Regression Model: Confusion Matrix

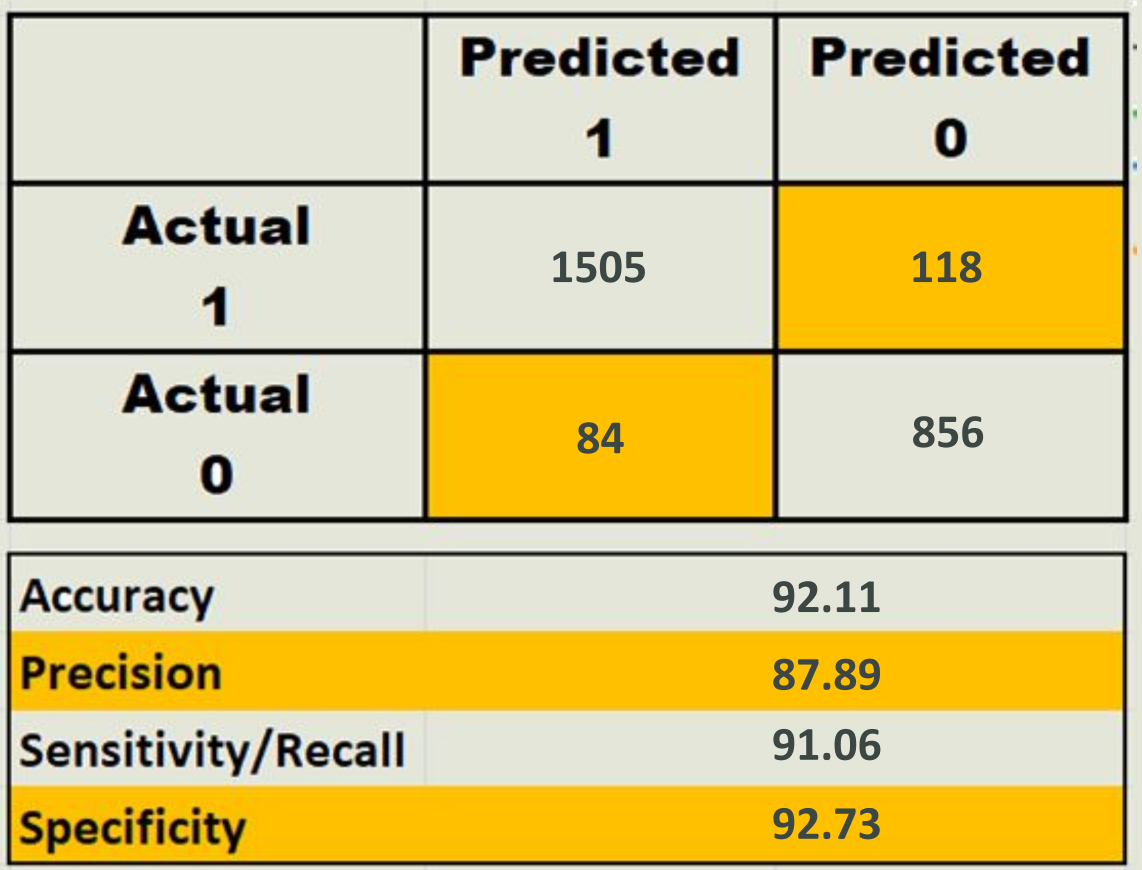
Train Dataset



	Predicted 1	Predicted 0
Actual 1	3378	309
Actual 0	210	2082

Accuracy	91.32
Precision	87.08
Sensitivity/Recall	90.84
Specificity	91.62

Test Dataset



	Predicted 1	Predicted 0
Actual 1	1505	118
Actual 0	84	856

Accuracy	92.11
Precision	87.89
Sensitivity/Recall	91.06
Specificity	92.73

Parameters in the Final Model

The following parameters are considered to calculate the lead scores:

- Total Time Spent on Website
- Welingak Website (Lead Source)
- Student of Some School (Lead Profile)
- SMS Sent (Last Activity)
- Modified (Last Notable Activity)
- Closed by Horizon (Tags)
- Lost to ENIS (Tags)
- Will revert after reading the email (Tags)
- Invalid number (Tags)
- Switched off (Tags)
- Ringing (Tags)
- Interested in other courses (Tags)

Lead Score > 30 can be considered as cutoff for good conversion rate (Hot Leads).

Top Variables for Lead Conversion

- Top 3 variables to contribute towards probability of lead conversion:

  1) Tags
  2) Lead Source and Profile
  3) Total Time Spent on Website

- Top 3 dummy variables to increase the probability of lead conversion:

  1) Closed by Horizon (Tags)
  2) Lost to EINS (Tags)
  3) Welingak Website (Lead Source)

Strategy: Convert all of the potential leads

- The sales team has enough bandwidth with interns.

- Strategy should be to prioritise and increase sensitivity, which will result in a lower value of cutoff probability.

- This will eventually increase the number of true positives and decrease the number of false negatives, thus capturing almost all of the potential leads.

Strategy: Make only necessary calls and minimise useless calls

- The sales team has new work after reaching its target, and needs to focus on only necessary calls.
- Strategy should be to decrease the number of false positives and increase the number of true positives.
- This can be achieved by prioritising and increasing precision, which will result in a higher cutoff probability, thus capturing the most potential leads.