Lead Scoring Case Study

By:

Ojas Haldankar

Nikhil Kishore

Niharika Trivedi

Problem Statement

- Help is needed for an X Education to identify the most promising leads—those that have the highest chance of becoming clients.
- The business wants us to create a model in which every lead must be given a lead score, meaning that consumers who have higher lead scores are more likely to convert than those who have lower lead scores.
- The intended lead conversion rate, as stated by the CEO in particular, is approximately 80%.

Objective

- * Build a logistic regression model that will allow the business to target potential leads by giving each lead a score between 0 and 100.
- * A higher score would indicate that the lead is hot and likely to convert.
- * A lower number would indicate that the lead is cold and unlikely to convert.

Approach

- The analysis has been performed by developing different models with following considerations:
  1. Feature selection and refinement from business standpoint
  2. Considering variability in features
  3. Making sure actionable insights can be generated

- The various steps performed are:
  1. Reading and understanding the data
  2. Data Cleaning & EDA
  3. Data Preparation
  4. Splitting the data into training and testing sets and rescaling
  5. Model Selection
  6. ROC Curve and Optimization of Sensitivity-Specificity trade off
  7. Validating training data with optimum parameter values
  8. Predictions on testing data
  9. Implementing Precision & Recall approach

Analysis Methodology

Assigning importance to variable buckets, doing sense checks on the data type of each variable

Feature Selection, Feature Engineering Rationale - 1

Variables	Description	Class	Categorical ?	Levels	Action Required	Rationale
Prospect ID	A unique ID with which the customer is identified.	Object	NA	NA	Delete	Record Variable
Lead Number	A lead number assigned to each lead procured.	int 64	NA	NA	Delete	Record Variable
Lead Origin	The origin identifier with which the customer was identified to be a lead. Includes API, Landing Page Submission, etc.	Object	Yes	5	Delete	Purely technical metric. Lead Source is a better metric
Lead Source	The source of the lead. Includes Google, Organic Search, Olark Chat, etc.	Object	Yes	21	Reduce to Top 5 and Other (Only 4% contribution for levels 6-21)	Crucial to understand which marketing medium has highest conversion
Do Not Email	An indicator variable selected by the customer wherein they select whether of not they want to be emailed about the course or not.	Object	Yes	2	Convert to dummy, change "Yes/No" to 1/0	Do not disturb might be less likely to convert ?
Do Not Call	An indicator variable selected by the customer wherein they select whether of not they want to be called about the course or not.	Object	Yes	2	Convert to dummy, change "Yes/No" to 1/0	Do not disturb might be less likely to convert ?
Converted	The target variable. Indicates whether a lead has been successfully converted or not.	int 64	No	1	Nothing	Output Metric
TotalVisits	The total number of visits made by the customer on the website.	float 64	No	NA	Check for Outliers	Higher engagement metric
Total Time Spent on Website	The total time spent by the customer on the website.	int 64	No	NA	Check for Outliers	Higher engagement metric
Page Views Per Visit	Average number of pages on the website viewed during the visits.	float 64	No	NA	Check for Outliers	Higher engagement metric
Last Activity	Last activity performed by the customer. Includes Email Opened, Olark Chat Conversation, etc.	Object	Yes	17	Delete	Last Mile Attribution is outdated + Already covered in Last Notable Activity
Country	The country of the customer.	Object	Yes	38	Delete (26% missing values)	Missing Values + Insuffient info in the brief on it's importance

Feature Selection, Feature Engineering Rationale - 2

Variables	Description	Class	Categorical ?	Levels	Action Required	Rationale
Specialization	The industry domain in which the customer worked before. Includes the level 'Select Specialization' which means the customer had not selected this option while filling the form.	Object	Yes	19	Delete (15% missing values)	Insufficient information on specific courses offered by X Education, hence which specializations might have a higher chance to convert. **No Selection is top category**
How did you hear about X Education	The source from which the customer heard about X Education.	Object	Yes	10	Delete (24% missing values)	Missing Values + Verbal anwers could be inaccurate ?
What is your current occupation	Indicates whether the customer is a student, umemployed or employed.	Object	Yes	6	Delete (29% missing values)	Could have been useful metric, but large % of missing values
What matters most to you in choosing this course	An option selected by the customer indicating what is their main motto behind doing this course.	Object	Yes	3	Delete (29% missing values)	Large % of missing values + question is a bit qualitative / generic
Search	Indicating whether the customer had seen the ad in any of the listed items.	Object	Yes	2	Convert to dummy	Crucial for marketing ROI
Magazine	Indicating whether the customer had seen the ad in any of the listed items.	Object	Yes	1	Delete (Only No responses)	
Newspaper Article	Indicating whether the customer had seen the ad in any of the listed items.	Object	Yes	2	Convert to dummy	Crucial for marketing ROI
X Education Forums	Indicating whether the customer had seen the ad in any of the listed items.	Object	Yes	2	Convert to dummy	Crucial for marketing ROI
Newspaper	Indicating whether the customer had seen the ad in any of the listed items.	Object	Yes	2	Convert to dummy	Crucial for marketing ROI
Digital Advertisement	Indicating whether the customer had seen the ad in any of the listed items.	Object	Yes	2	Convert to dummy	Crucial for marketing ROI
Through Recommendations	Indicates whether the customer came in through recommendations.	Object	Yes	2	Convert to dummy	Crucial for marketing ROI
Receive More Updates About Our Courses	Indicates whether the customer chose to receive more updates about the courses.	Object	Yes	1	Delete (Only No responses)	Only No as response

Feature Selection, Feature Engineering Rationale - 3

Variables	Description	Class	Categorical ?	Levels	Action Required	Rationale
Tags	Tags assigned to customers indicating the current status of the lead.	Object	Yes	26	Delete (36% missing values)	High % of missing values + requires NLP / Sentiment Analysis
Lead Quality	Indicates the quality of lead based on the data and intuition the the employee who has been assigned to the lead.	Object	Yes	5	Delete (51% missing values)	Could have been useful check on sales team's intuition, but high % of missing values
Update me on Supply Chain Content	Indicates whether the customer wants updates on the Supply Chain Content.	Object	Yes	1	Delete (Only No responses)	Only No as response
Get updates on DM Content	Indicates whether the customer wants updates on the DM Content.	Object	Yes	1	Delete (Only No responses)	Only No as response
Lead Profile	A lead level assigned to each customer based on their profile.	Object	Yes		Delete (29% missing values)	Insufficient Information
City	The city of the customer.	Object	Yes	7	Delete (15% missing values)	High % of missing values + insufficient info in the brief
Asymmetrique Activity Index	An index and score assigned to each customer based on their activity and their profile	Object	No		Delete (51% missing values)	Insufficient Information + Derived Metric of existing data
Asymmetrique Profile Index		Object	No			Insufficient Information + Derived Metric of existing data
Asymmetrique Activity Score		int 64	No			Insufficient Information + Derived Metric of existing data
Asymmetrique Profile Score		int 64	No			Insufficient Information + Derived Metric of existing data
I agree to pay the amount through cheque	Indicates whether the customer has agreed to pay the amount through cheque or not.	Object	Yes	1	Delete (Only No responses)	Only No as response
a free copy of Mastering The Interview	Indicates whether the customer wants a free copy of 'Mastering the Interview' or not.	Object	Yes	2	Convert to dummy	2 Level variable with a pointed / specific question
Last Notable Activity	The last notable acitivity performed by the student.	Object	Yes	16	Reduce to Top 3 and Other	Could be crucial for marketing ROI

Feature Selection, Engineering : Key Actions Taken

Variable buckets

- **High Domain Importance + Data Sufficiency**
  - Lead Source, Do Not Email, Do Not Call, Website Visits, Time Spent, Page Views, Where Seen Ad (Search, Digital, Through Recommendation), Last Notable Activity
    - Action : Reduce # of levels, ex. Lead Source (club low contribution levels as “other”)
    - Action : Convert to 1/0 from “Yes/No”, ex. Email, Call
    - Action : Check for Outliers in the numeric variables

- **High Domain Importance + Data Insufficiency**
  - Country, City, Specialization, Occupation, Receive Updates, Tags, Lead Quality
    - Action: Delete (Large % of Missing Values (>20%) or Only “No” in response (Receive Updates)

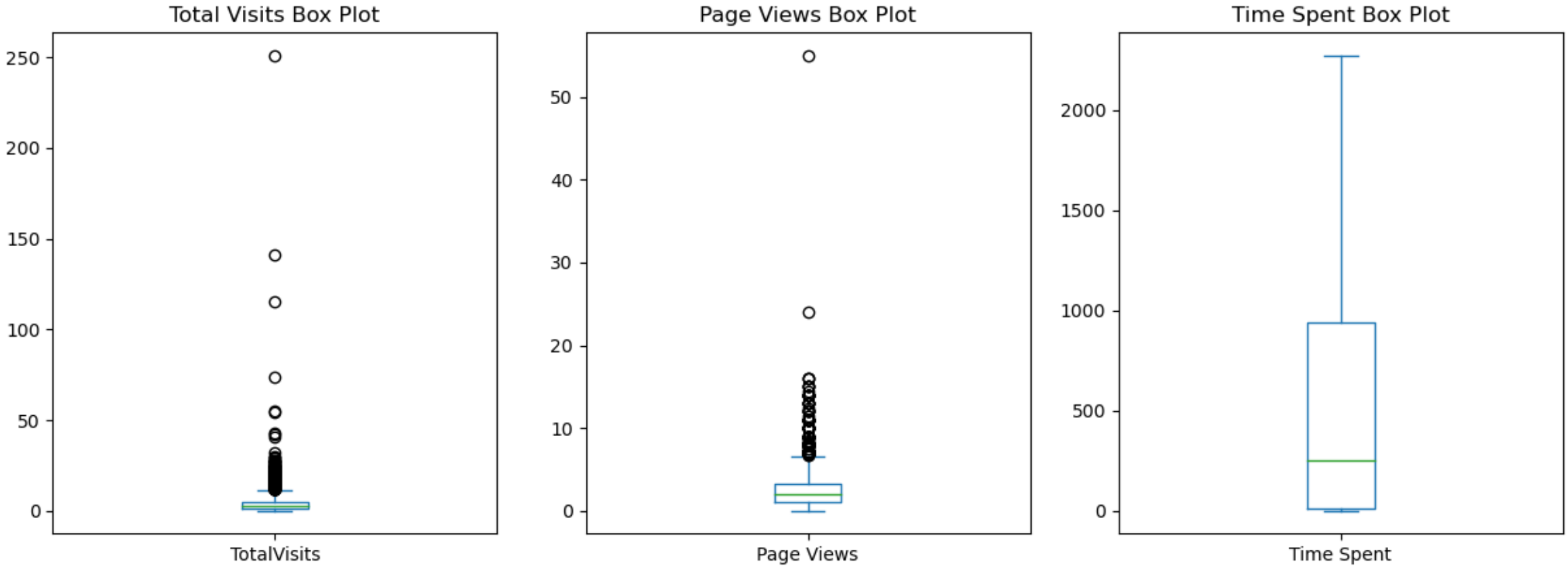
- **Low Domain Importance + Data Sufficiency**
  - Lead Origin (Technical Variable), Last Activity (Already covered in Last Notable Activity)
    - Action : Delete as they do not contribute to the modelling process / already covered in other variable

- **Insufficient Information on Importance in the brief + Data Insufficiency**
  - Assymetrique Index, Score Variables
    - Action : Delete (No information in the brief, large % of missing values)

Exploratory Data Analysis

Outliers, Missing Values, Reducing # of Levels in Cat Variable,
Correlation Grid

Outlier Analysis – Numeric Variables



Time Spent metric is stable with no outliers present : no action required

Total Visits, Page Views, domain knowledge suggests that isolated cases of large values may occur
- Do not have sufficient information to conclude if this is a technical error or not
- The high values are still in the hundreds and not in the thousands

Missing Value Treatment done as per requirement

> 15% missing values

Variable	% Missing
Update me on Supply Chain Content	51.6
Asymmetrique Profile Index	45.6
Asymmetrique Activity Score	45.6
Asymmetrique Profile Score	45.6
I agree to pay the amount through cheque	45.6
Lead Quality	36.3
Search	29.3
City	29.3
What matters most to you in choosing this course	29.1
Specialization	26.6
What is your current occupation	23.9
How did you hear about X Education	15.6
Asymmetrique Activity Index	15.4

Deletion of Columns

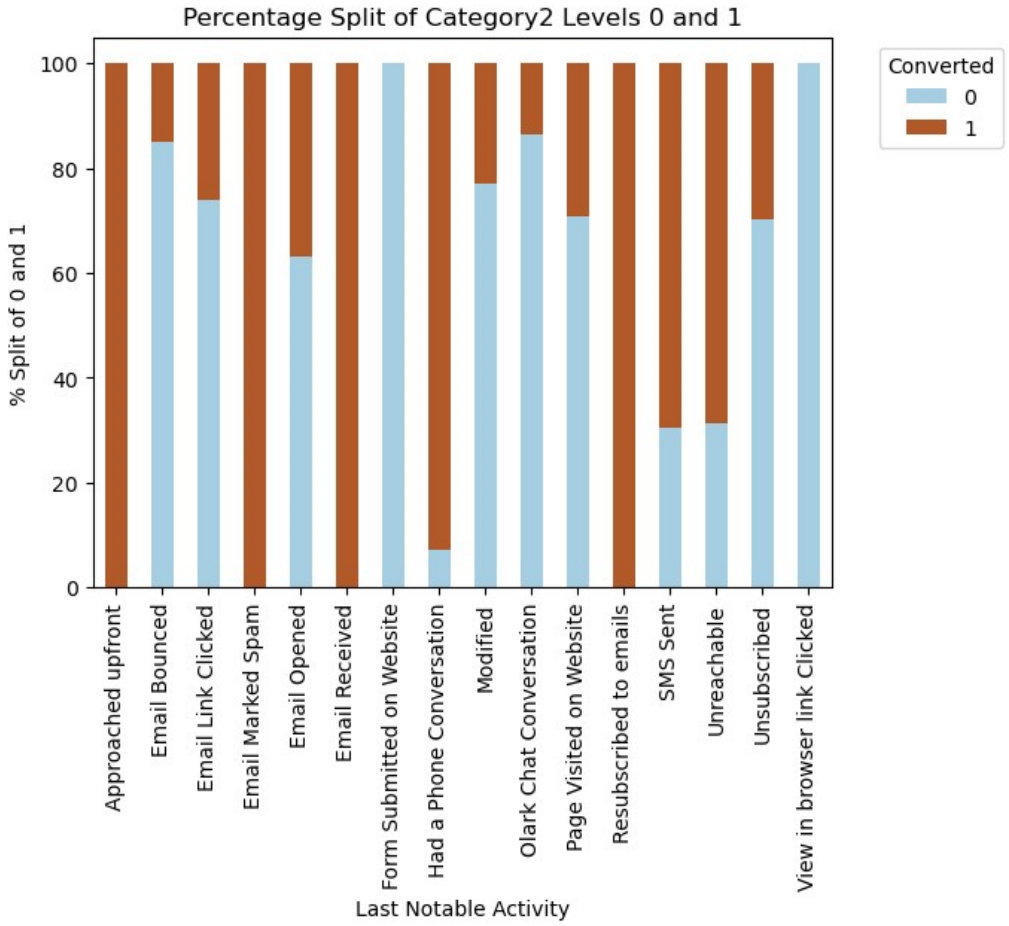
0.4 – 1.5% missing values

Variable	% Missing
Total Visits	1.5
Page Views	1.5

Imputation with Median
- High variance in both variables

Reducing # of levels in “Last Notable Activity”

Modified	36.907866
Email Opened	30.671447
SMS Sent	23.381139
Page Visited on Website	3.455020
Olark Chat Conversation	1.988266
Email Link Clicked	1.879618
Email Bounced	0.651890
Unsubscribed	0.488918
Unreachable	0.347675
Had a Phone Conversation	0.152108
Email Marked Spam	0.021730
Approached upfront	0.010865
Resubscribed to emails	0.010865
View in browser link Clicked	0.010865
Form Submitted on Website	0.010865
Email Received	0.010865

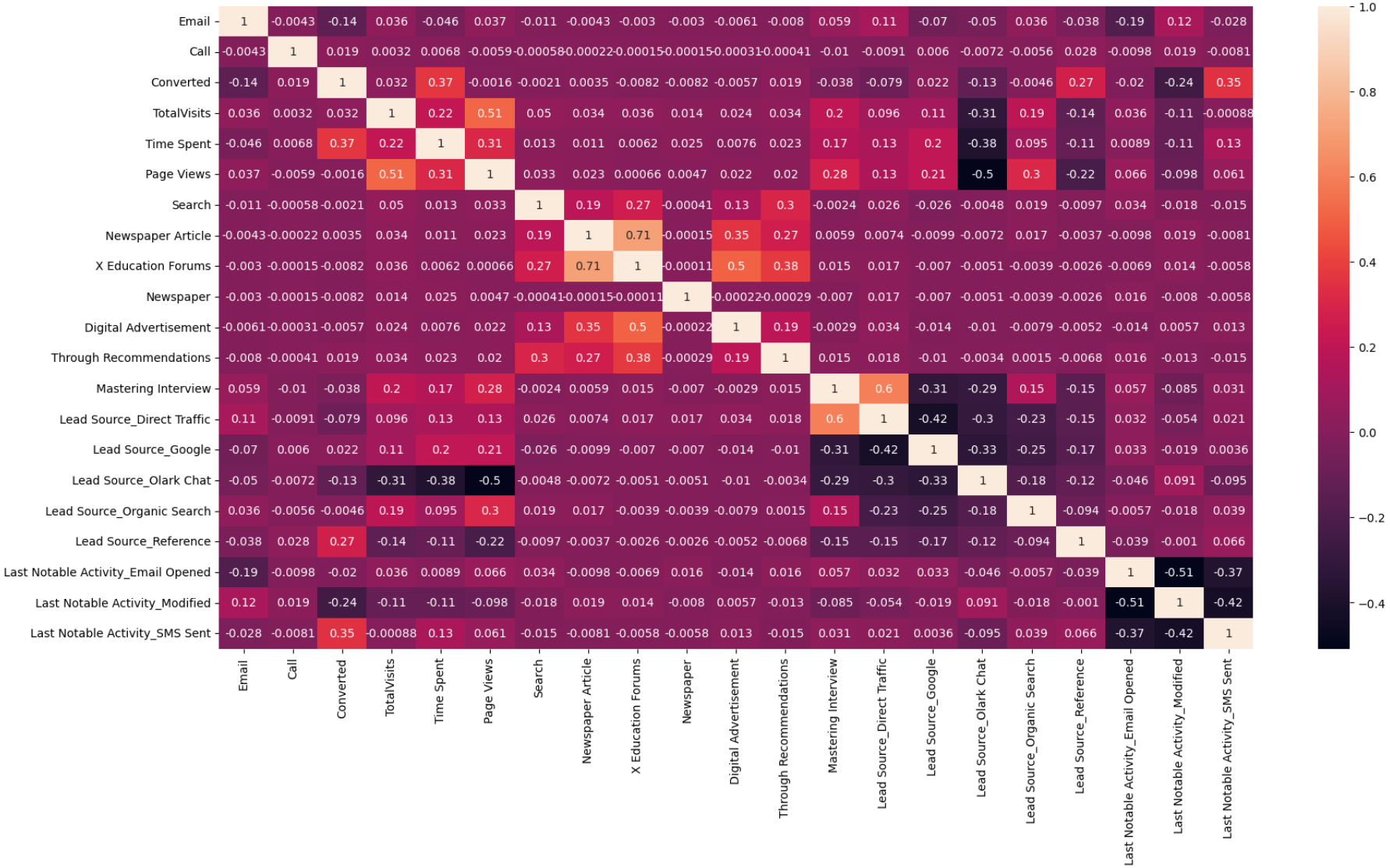


Levels beyond the Top 3 clubbed as “Other”

- Low contribution to total
- Conversion % is not substantially higher vs the Top 3 levels

High correlation between the numeric variables

- Correlation Grid between Feature Variables



Numeric Variables (Total Visits, Time Spent, Page Views) are expected to be well correlated, since they all are a surrogate to consumer engagement

Missing Value Treatment

Prospect ID	0.0
Lead Number	0.0
Lead Origin	0.0
Lead Source	0.0
Do Not Email	0.0
Do Not Call	0.0
Converted	0.0
TotalVisits	0.0
Total Time Spent on Website	0.0
Page Views Per Visit	0.0
Last Activity	0.0
Country	0.0
Specialization	0.0
What is your current occupation	0.0
What matters most to you in choosing a course	0.0
Search	0.0
Newspaper Article	0.0
X Education Forums	0.0
Newspaper	0.0
Digital Advertisement	0.0
Through Recommendations	0.0
Tags	0.0
A free copy of Mastering The Interview	0.0
Last Notable Activity	0.0

dtype: float64

- **40%** threshold was considered. All the columns having more than 40% of missing records were dropped.
- Categorical columns were either imputed with mode or some new category after performing feature analysis and domain understanding.

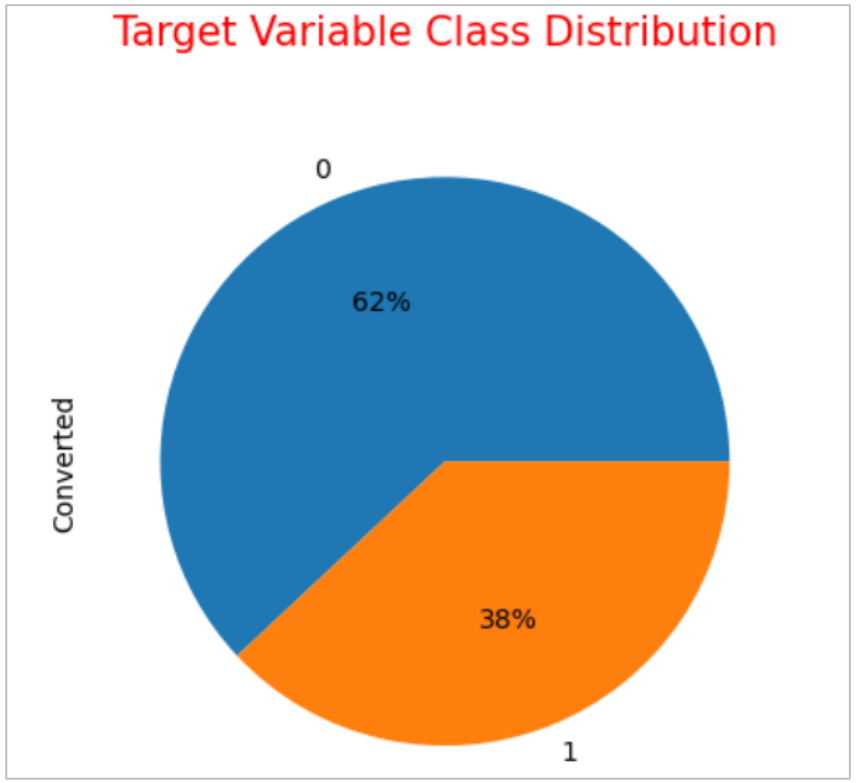
Univariate Analysis

- Around **93%** of customers are leads generated from either **Landing Page Submission** (54%) or **API** (39%).
- Approximately **80%** of the traffic is from **Google**, **Direct Traffic** or **Olark Chat**.
- Most of the customers are ok with receiving **emails** or **calls**.
- Around **68%** of customer have recently **opened emails** (38%) or **have sent sms** (30%).
- **71%** of the customers are from **India** and in **26%** of the cases, country name is missing.
- Around **60%** of the customers/students are **unemployed**.
- **70%** of the customers are looking for **Better Career Prospects**.
- Seems like customer have seen the ad or shown interest for program through some different sources as for listed sources the response is clearly less.
- **32%** of customers want a free copy of **Mastering The Interview**.

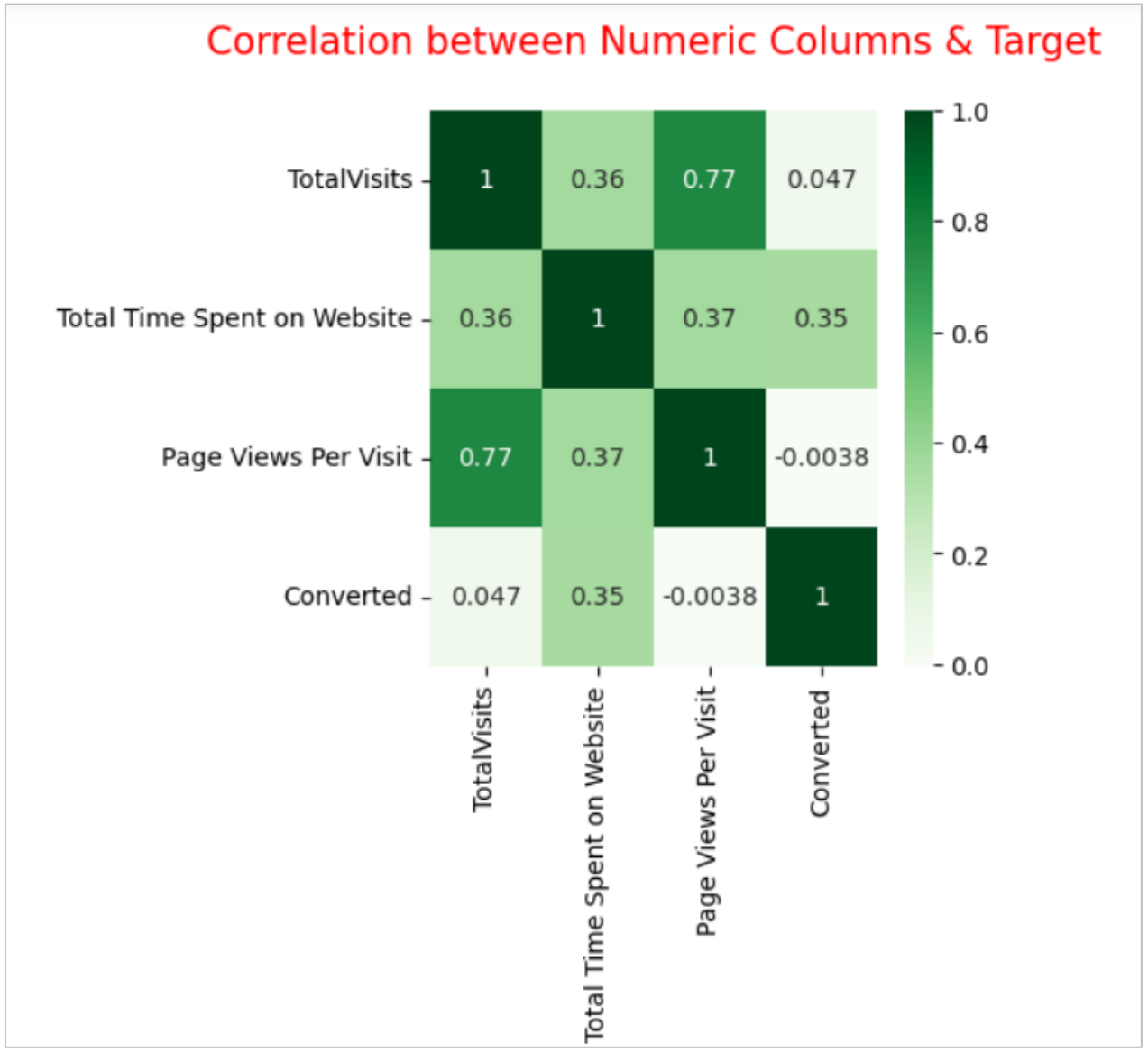
Bivariate Analysis

- **Lead Add Form** have better proportion of conversion, but the leads are low in number.
- **Reference** lead source has the highest chance of conversion. **Olark Chat** has more associated leads but comparatively less conversion probability.
- People who choose to get **emails** or **calls** have good conversion rate.
- **Sending Sms** to customers improves the chances of conversion.
- Customers from **India** are major in fraction and have shown positive response.
- **Working professionals** have very good chance of conversion followed by **unemployed** people.
- People opting the course for **Better Career Prospects** have high chance of conversion.
- People who did not demand for free copy of **Mastering The Interview** have shown better conversion.
- Customers who have **approached by sending sms** have higher chances of conversion.

Target Variable Analysis



The present/initial conversion rate was **38%**.



Model Building & Evaluation

Final List of Feature Variables in each model

Category	ML Model (1)	ML Model (2)	ML Model (3)
Email/ Call	**Do Not Email_Yes**	**Do Not Email_Yes**	**Do Not Email_Yes**
Last Activity		Last Activity_Olark Chat Conversation	Last Activity_Olark Chat Conversation
Last Activity		Last Activty_Converted to Lead	Last Activty_Converted to Lead
Last Activity		Last Activty_Email Bounced	
Last Notable Activity	**Last Notable Activity_SMS Sent**	**Last Notable Activity_SMS Sent**	**Last Notable Activity_SMS Sent**
Last Notable Activity			Last Notable Activity_Unsubscribed
Last Notable Activity	Last Notable Activity_Email Opened	Last Notable Activity_Email Opened	
Last Notable Activity	Last Notable Activity_Modified	Last Notable Activity_Modified	
Last Notable Activity		Last Notable Actiity_Email Link Clicked	
Lead Source		Lead Source_Welingak Website	Lead Source_Welingak Website
Lead Source	Lead Source_Direct Traffic		
Lead Source	Lead Source_Olark Chat		
Lead Source	Lead Source_Organic Search		
Lead Source	Lead Source_Reference		
Lead Source		Lead Source_Direct Traffic	
Lead Origin		Lead Origin_Lead Add Form	
Website Metrics			Page Views Per Visit
Website Metrics		Total Visits	
Website Metrics	**Total Time Spent on Website**	**Total Time Spent on Website**	**Total Time Spent on Website**
Tags			Tags_Closed by Horizzon
Tags			Tags_Lost to EINS
Tags			Tags_Ringing
Tags			Tags_Switched Off
Tags			Tags_will revert after reading the e-mail
What Matters Most			What Matters Most
Current Occupation		Working Professionals	
Through Recommendations	Through Recommendations		

Notable common features
in both Models

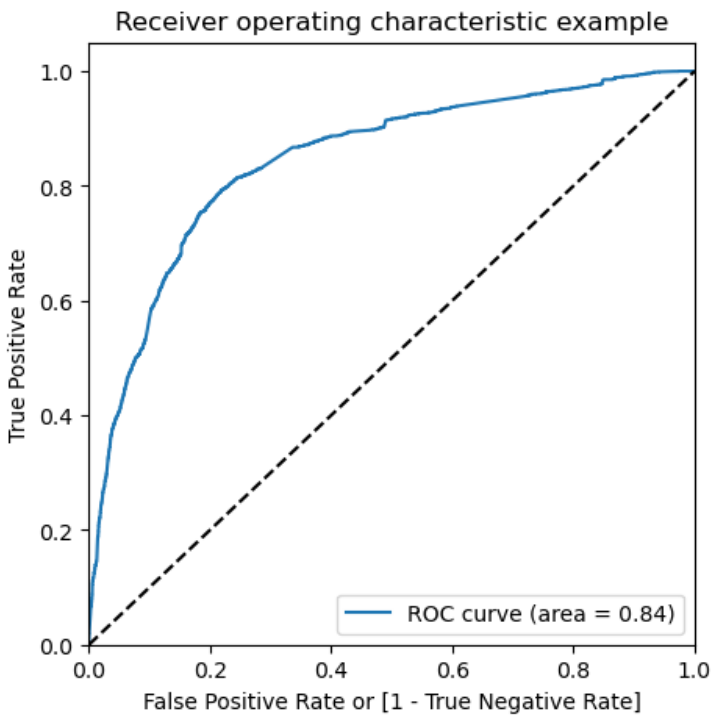
1. **Do not E-mail**
(Might not be interested in further details – lower chance of conversion)

2. **Last Notable Activity_SMS Sent**
(Taking action – higher chance of conversion)

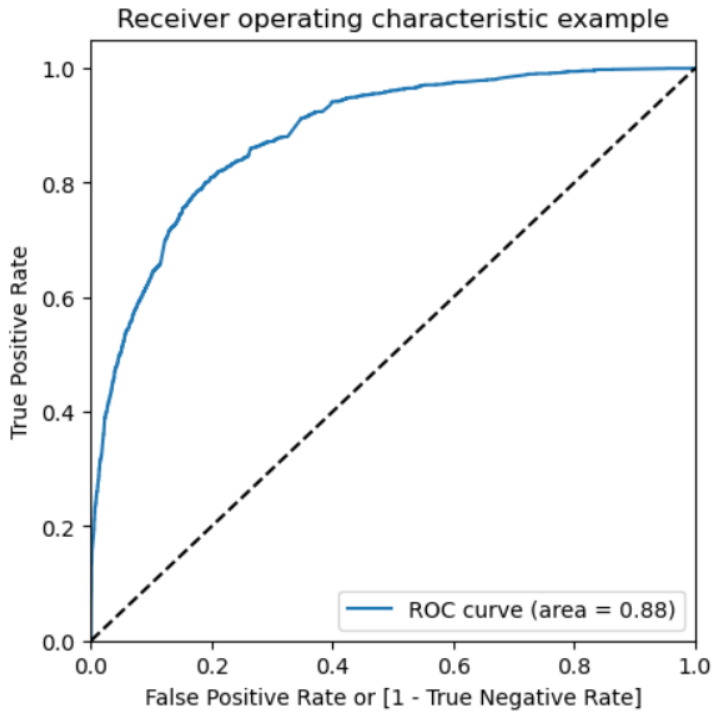
3. **Total Time Spent on Website**
(Well known engagement metric – higher chance of conversion)

ROC for the Models

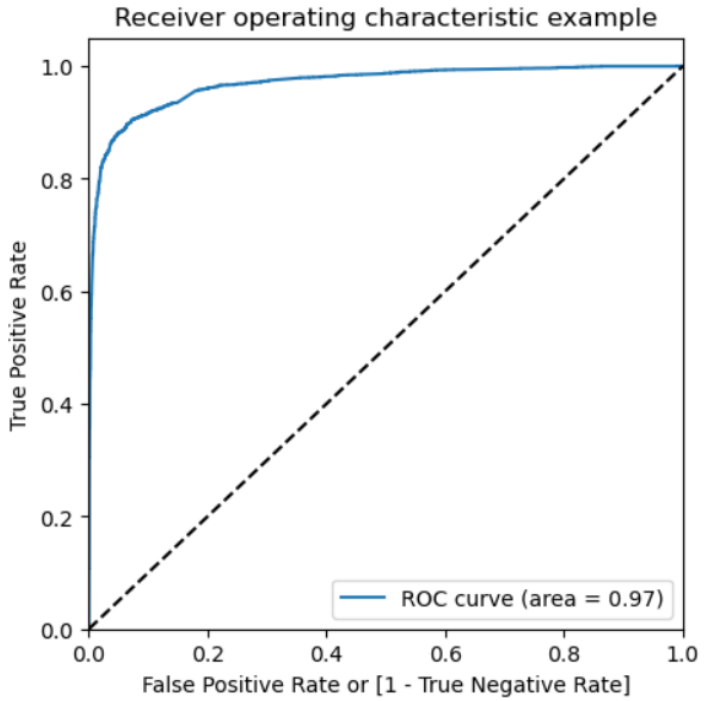
ML Model (1)



ML Model (2)

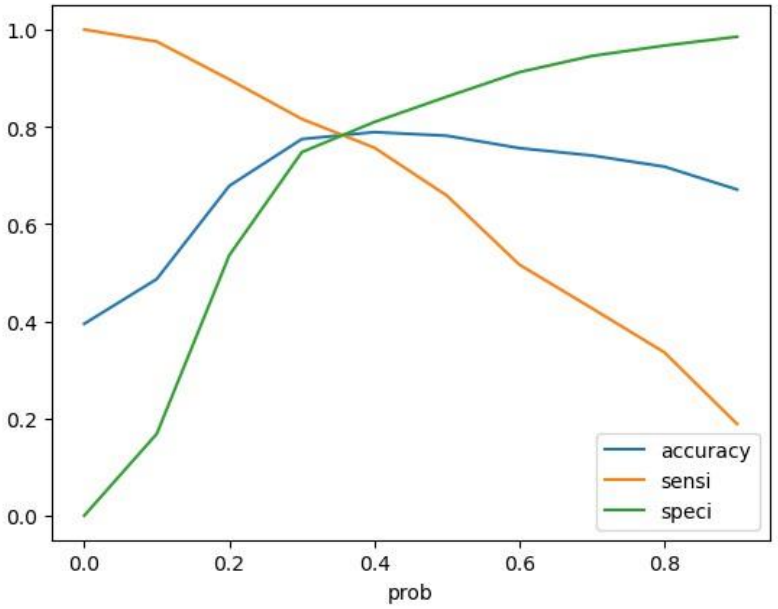


ML Model (3)



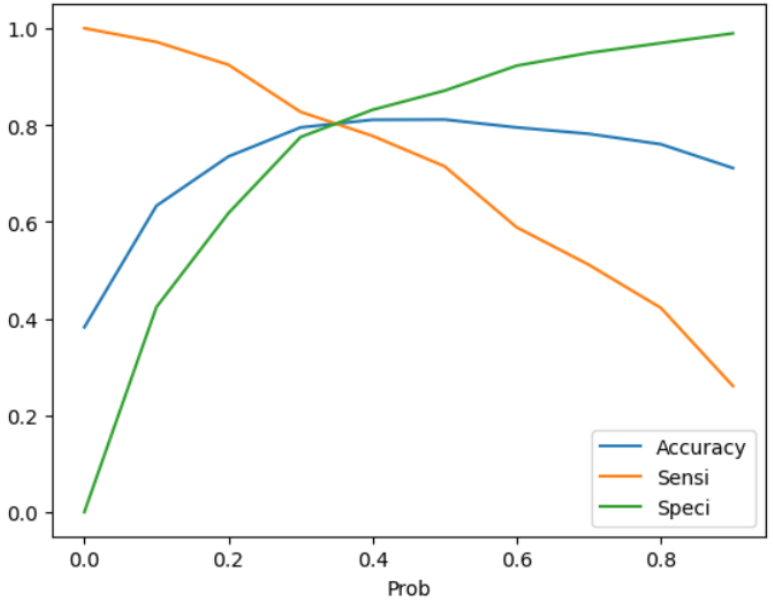
Sensitivity – Specificity Trade off on Training Data

ML Model (1)



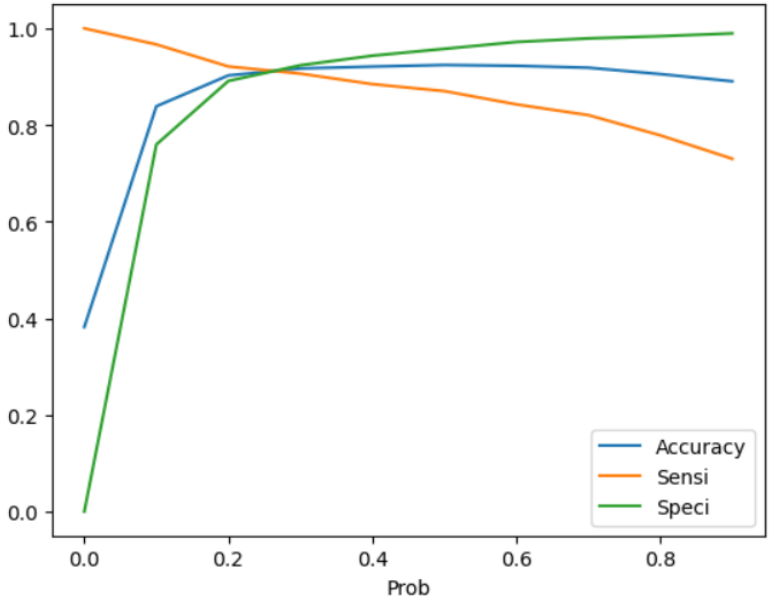
Cut off – 0.35

ML Model (2)



Cut off – 0.37

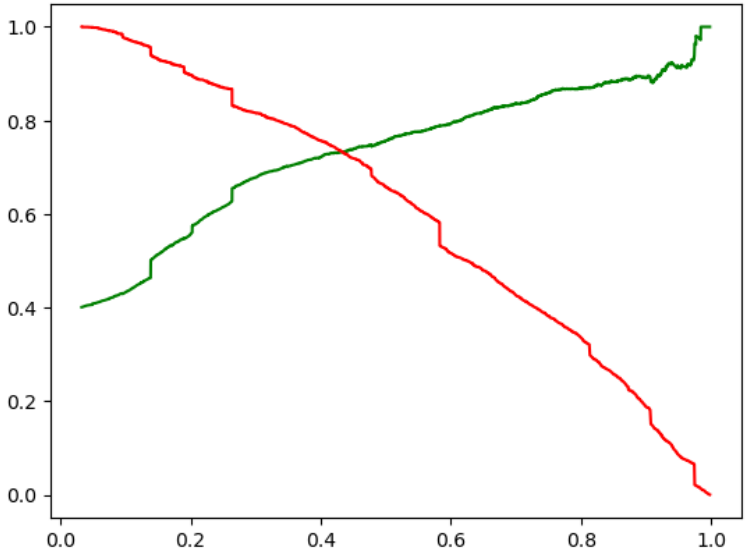
ML Model (3)



Cut off – 0.3

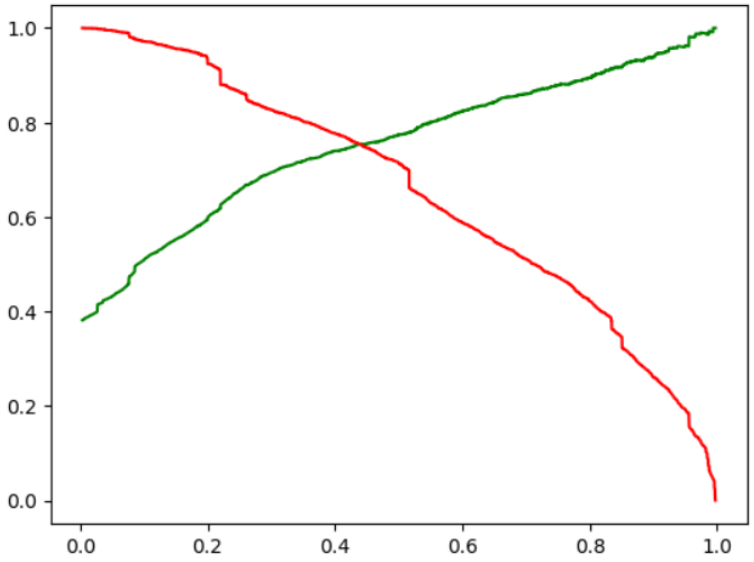
Precision Recall Trade off on Training Data

ML Model (1)



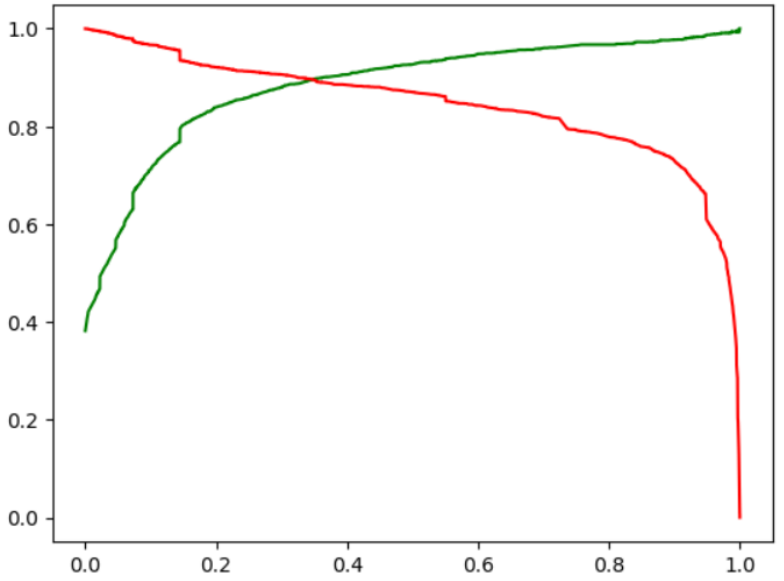
Cut off – 0.42

ML Model (2)



Cut off – 0.41

ML Model (3)



Cut off – 0.37

Accuracy Metrics for Train and Test

Metric	Data	ML Model 1	ML Model 2	ML Model 3
Accuracy	Train	78.5	80.6	91.7
Sensitivity	Train	79.1	79.3	90.6
Specificity	Train	78.1	81.4	92.4
FPR	Train	21.8	18.6	7.6
PPV	Train	70.2	75.5	87.9
NPV	Train	85.1	86.4	94.1
Accuracy	Test	79.2	81.8	91.3
Sensitivity	Test	80.1	81.1	89.9
Specificity	Test	78.7	82.3	92.1
FPR	Test	21.2	17.7	7.8
PPV	Test	67.7	73.4	87.3
NPV	Test	87.6	87.8	93.8

1. All the models are stable.

- Accuracy, Sensitivity, Specificity consistent on Train and Test for both the models

2. All the above metrics > 75% on both train and test in all the models

Recommendations

The factors which are important in deciding the Conversion Rate are:

- **Total Time Spent on Website** - more the time spent, better is the conversion. Thus, team should focus on optimization of web page.
- When customer have not opted for **email follow up**, conversion rate is usually lower.
- When the customer is a **Working Professional**, chances of conversion are higher.
- Leads coming from **Reference** have better chances of conversion.
- When the last notable activity was **SMS sent**, conversion seems to rise.
- When the Lead Source is **Welingak Website**, it shows better conversion.
- **Lead Add Form** have better proportion of conversion, but the leads are low in number. Business should focus on this source for getting more leads.