Lead Scoring Case Study

Building Logistic Regression Model

Aditya Anand (adityaanand2701@gmail.com)
&
Hrithik Chand (hrithiknovember@gmail.com)

Agenda

- Problem Statement
- Brief description of data cleaning and imputing missing value
- Brief description of data analysis and data preparation
- Model Building and Evaluation
- Final Conclusion

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course 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. Moreover, the company also gets leads through past referrals. 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. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’. If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Brief description of Data Cleaning and Imputing missing Value

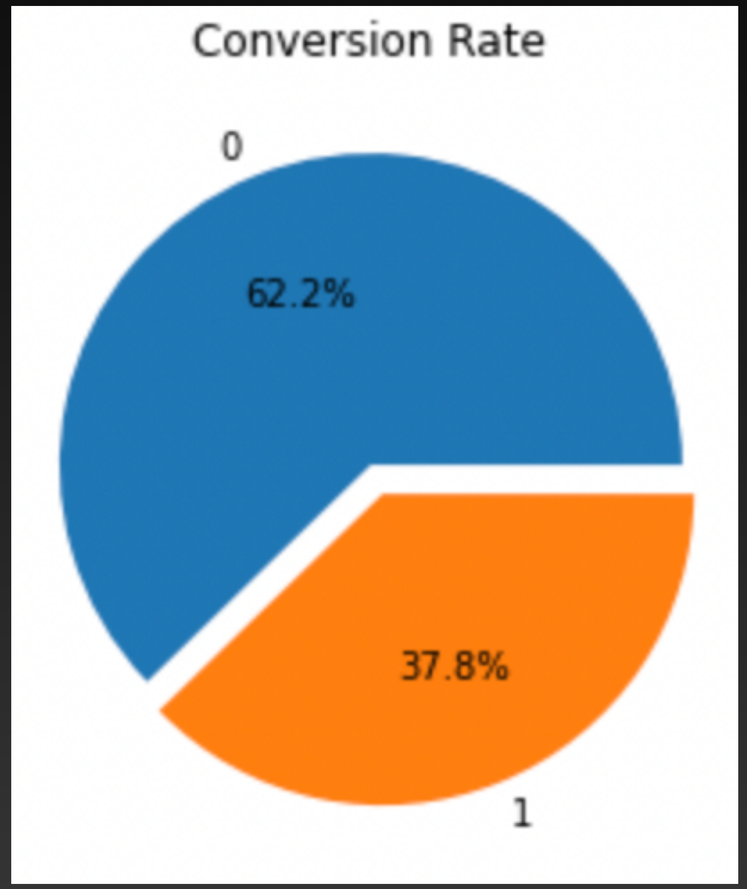
- Dropped unnecessary columns that has unique value for each rows
- Replaced the value ‘Select’ in the dataset with null values
- Dropped all columns which have more than 40% missing values
- For all categorical feature replaced the missing value with mode
  - Except for ‘Specialization’, here we replaced null value with ‘NA’ as we also have student data in our dataset
- Dropped complete rows for those column which have less than 2% missing values
- For numerical feature which has outlier, dropped compete rows which has value greater than 99 percentile

Brief description of Data analysis and Data Preparation

- Data Preparation:
  - Converted the binary variable (Yes/No) to 1/0
  - Binning the categorical features as some values has very less value count
  - Created the dummy variable for the categorical features
  - Dropped those columns which are highly skewed

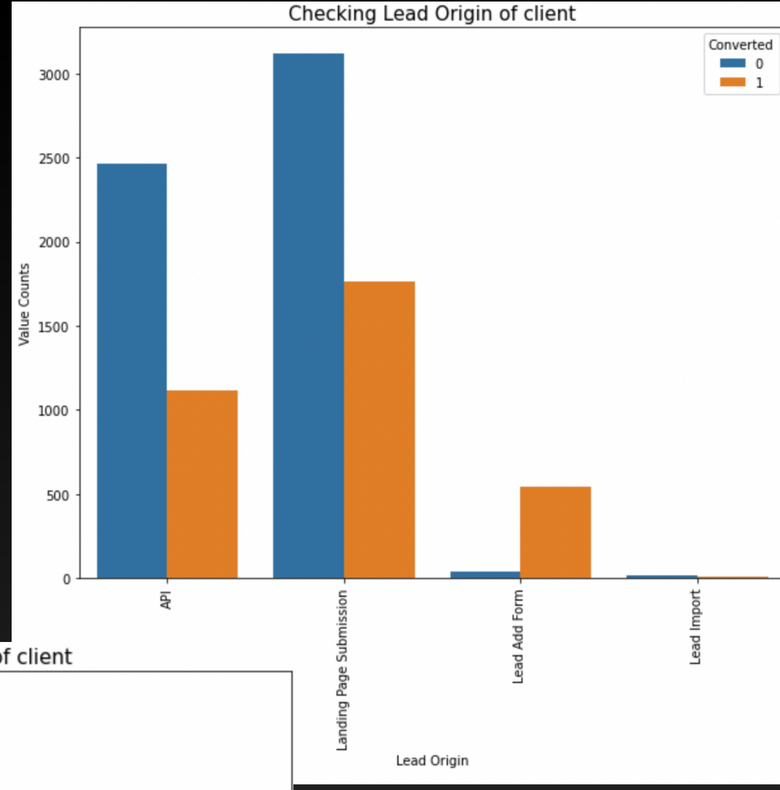
Brief description of Data analysis and Data Preparation

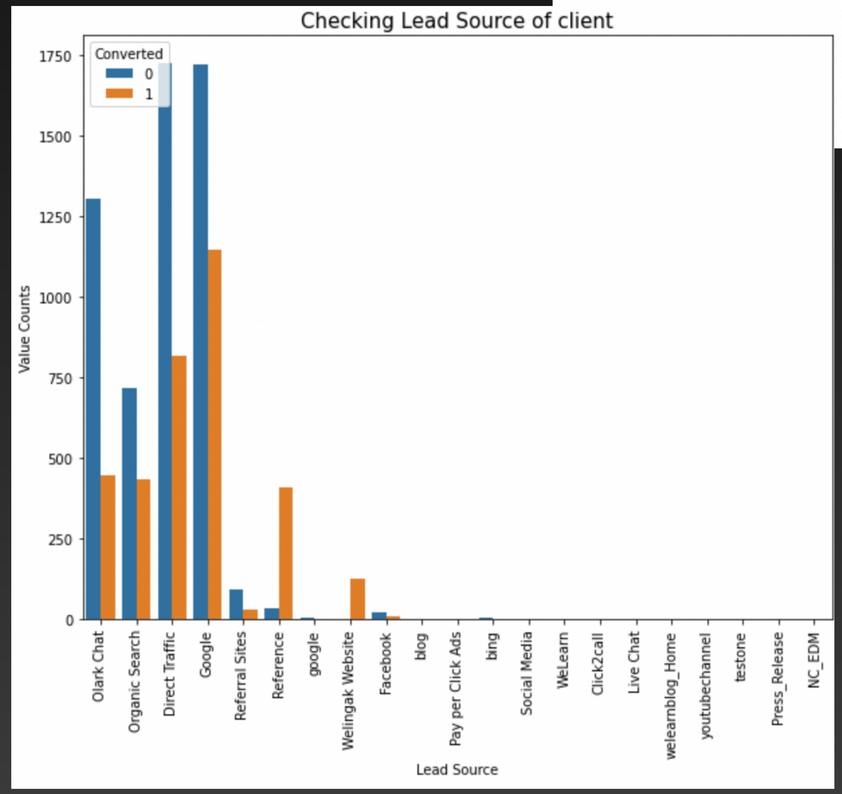
- Data Analysis:
  - Checking Conversion Rate
    - Cons rate is 37.8% -> data seems to be balanced and good for model building



Brief description of Data analysis and Data Preparation

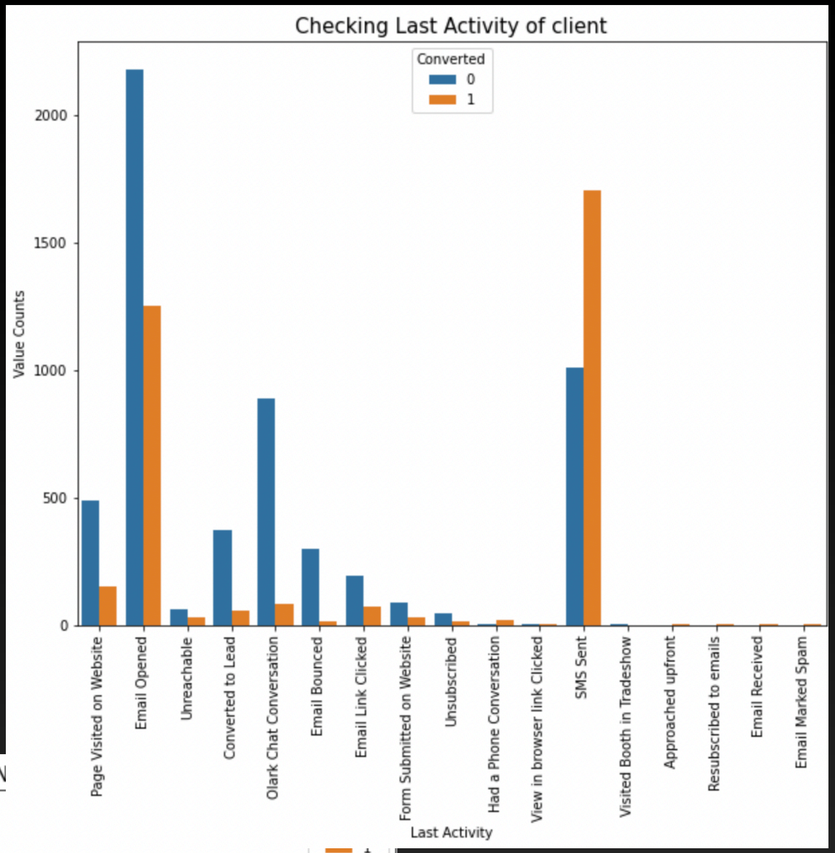
- Data Analysis:
  - Checking Lead Origin and Lead Source
    - Lead Add Form ha higher conversion rate and Landing page submission has maximum conversion
    - Client whose source are from google has higher conversion

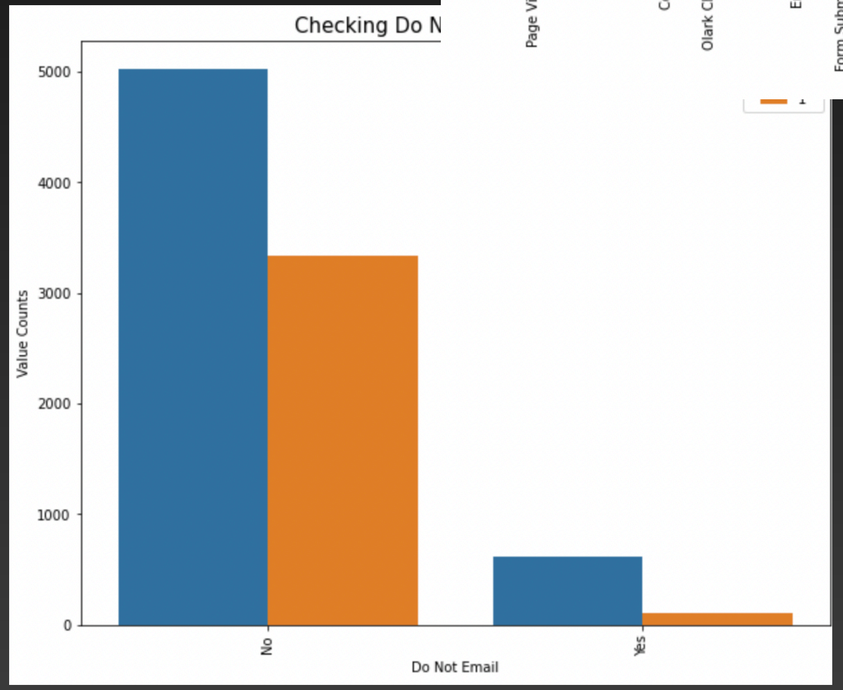




Brief description of Data analysis and Data Preparation

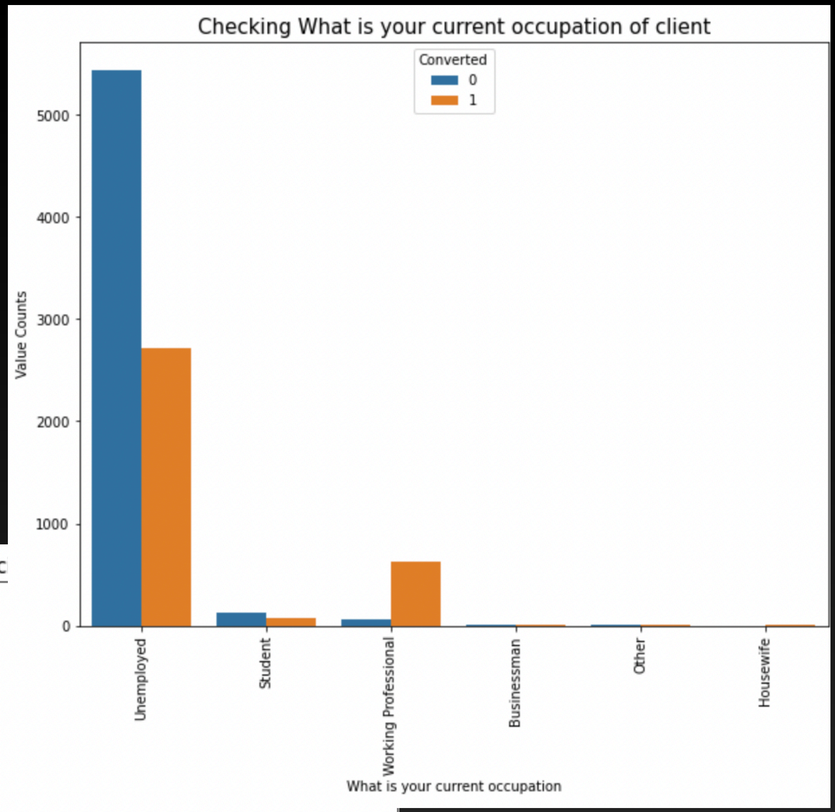
- Data Analysis:
  - Checking Do not email & Last Activity
    - Client who didn’t opted for email has higher conversion chances
    - Client whose last activity was sms sent and email opened has higher conversion chances

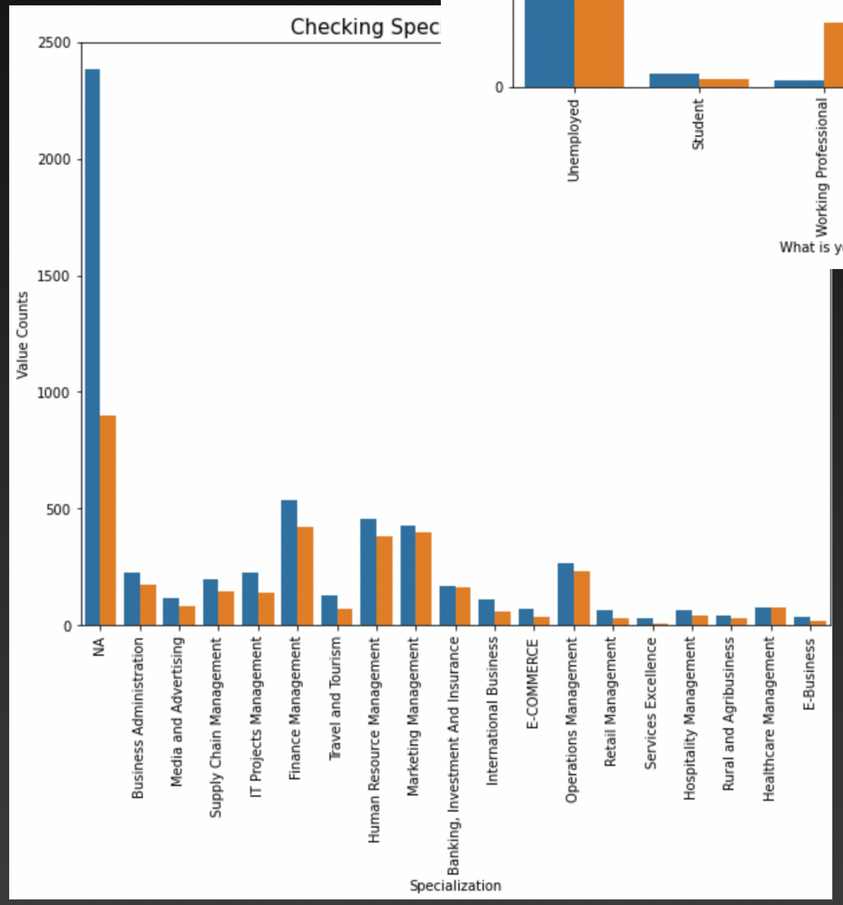




Brief description of Data analysis and Data Preparation

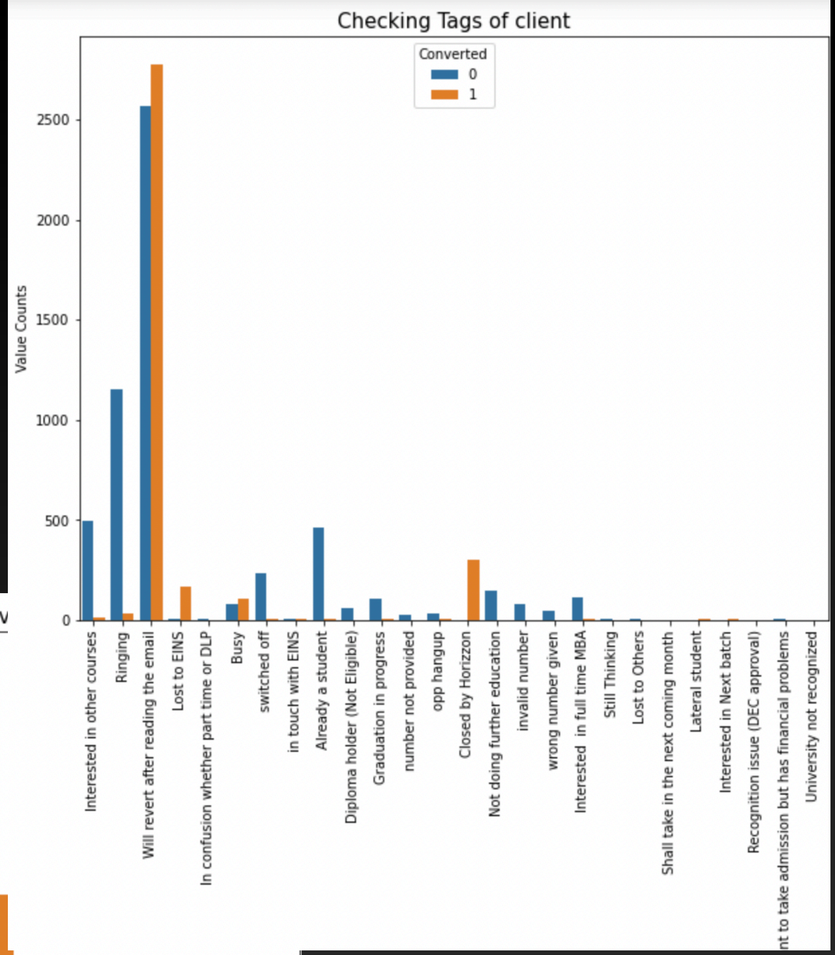
- Data Analysis:
  - Checking Specialisation & current occupation
    - Client who are unemployed & working professional has higher conversion chances
    - Client whose specialization is Management has higher conversion chances

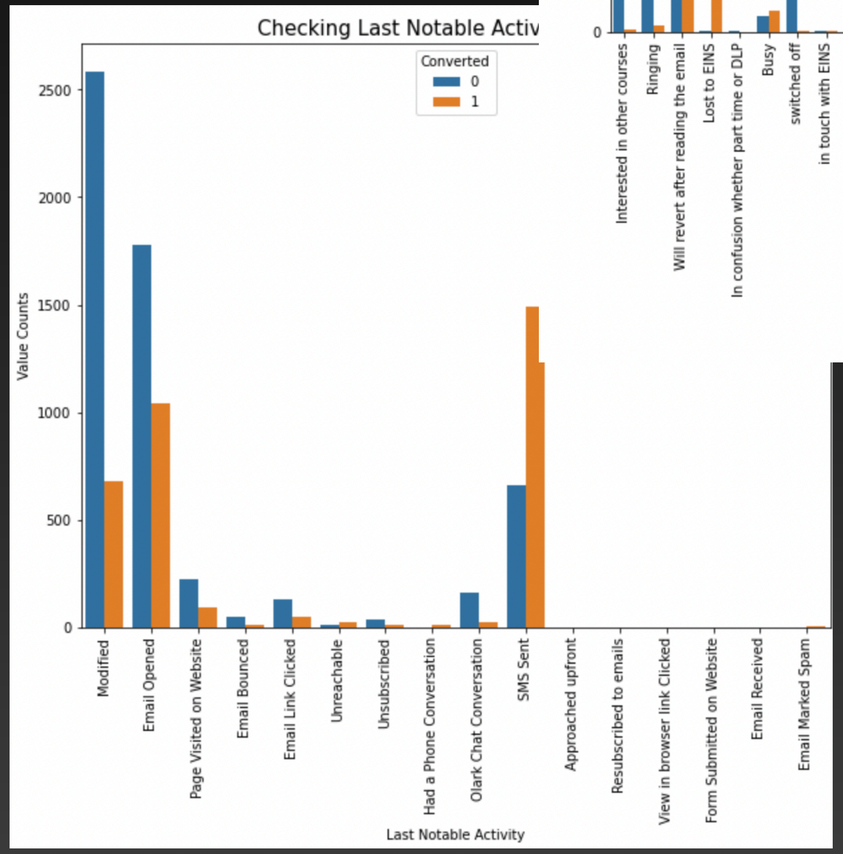




Brief description of Data analysis and Data Preparation

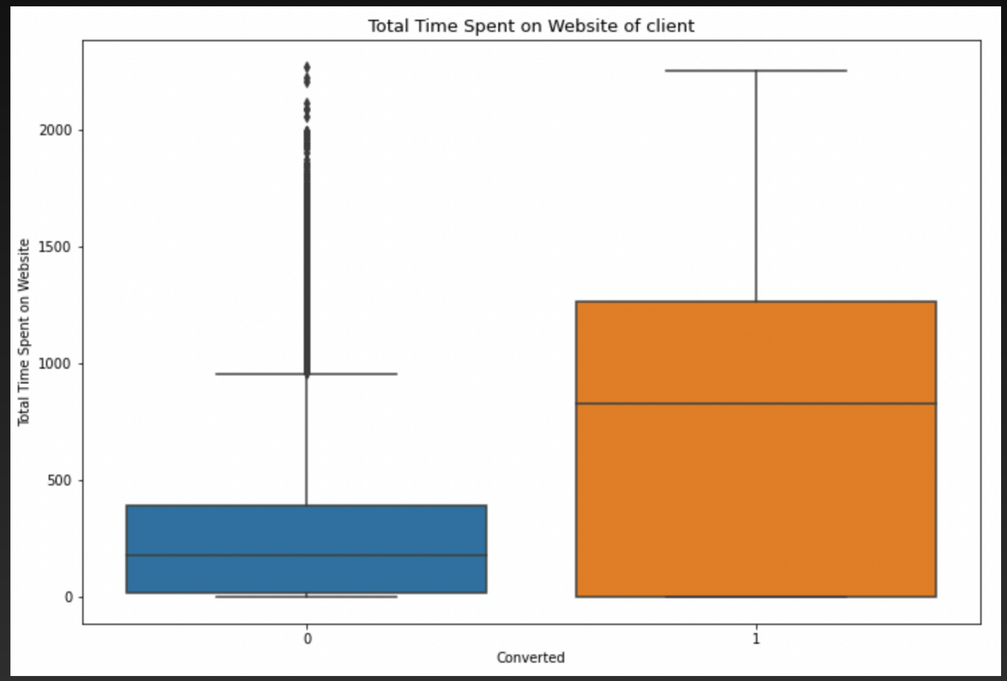
- Data Analysis:
  - Checking Tags & Last Notable activity
    - Client who will revert after reading email has higher conversion chances
    - Client whose last note;e activity is sms sent & email opened has higher conversion chances





Brief description of Data analysis and Data Preparation

- Data Analysis:
  - Checking Total time spent on website
    - Clients who is spending more time on website has higher conversion rate



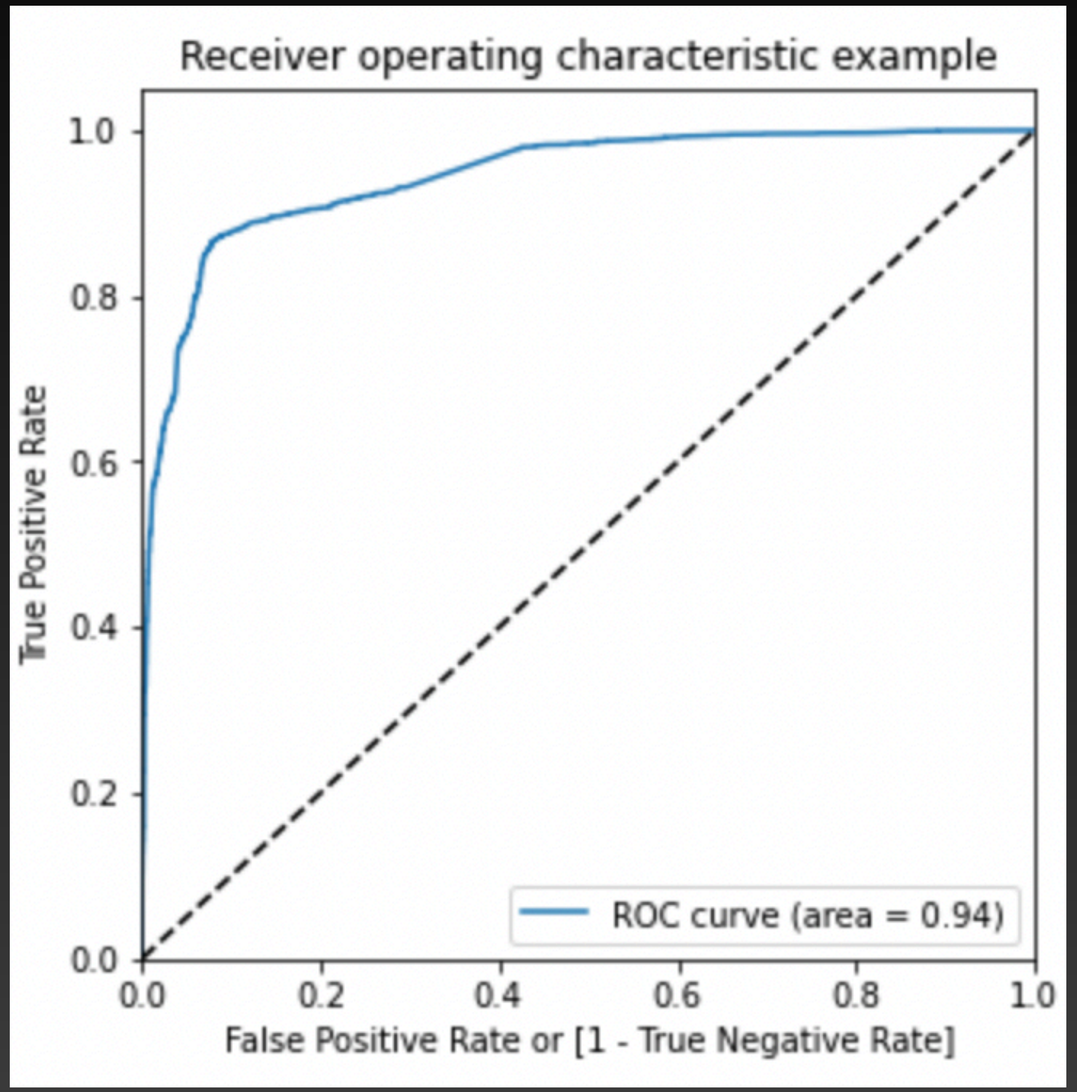
Model Building and Evaluation

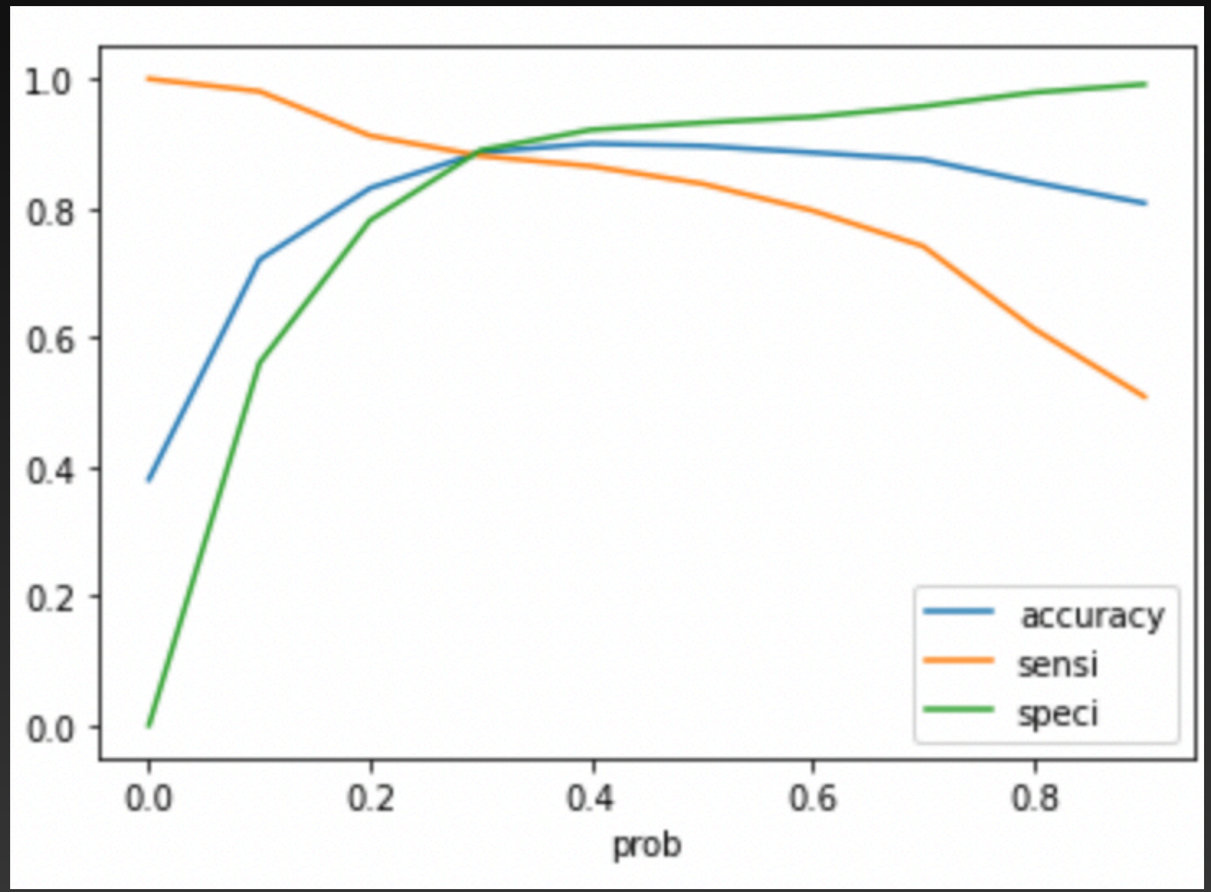
- Model Building:
  - Created dummy variable for all the categorical feature
  - Split the dataset into Test & Train in 30% - 70% respectively
  - Scaled all the numerical feature in the Train data set
  - Started building the Logistic Regression model
    - First select the top 15 feature using RFE (coarse tuning)
    - After selecting top 15 feature, started checking significance of each feature manually by seeing the p values and VIF and dropped feature which are either insignificant or has high VIF

Model Building and Evaluation

- Model Evaluation:
  - Selected the threshold probability as 0.5 and make the prediction on train set and check the accuracy of model
  - To select the optimal threshold draw the ROC curve and cut off prob curve
  - Optimal threshold calculated as 0.3
  - Finally make the prediction on Test set and compare the accuracy with train set
    - Train Set:
    - Accuracy : 88.63% Sensitivity : 88.08% Specificity : 88.97%
    - Test Set:
    - Accuracy : 88.46% Sensitivity : 88.43% Specificity : 88.47%

Model Building and Evaluation





Model Building and Evaluation - Final Model

Dep. Variable:	Converted	**No. Observations:**	6246
Model:	GLM	**Df Residuals:**	6233
Model Family:	Binomial	**Df Model:**	12
Link Function:	logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1757.1
Date:	Sun, 07 Nov 2021	**Deviance:**	3514.3
Time:	01:01:32	**Pearson chi2:**	9.94e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.3891	0.361	-9.395	0.000	-4.096	-2.682
Do Not Email	-1.6911	0.204	-8.286	0.000	-2.091	-1.291
Total Time Spent on Website	1.0562	0.045	23.299	0.000	0.967	1.145
Lead Origin_Lead Add Form	3.6188	0.266	13.614	0.000	3.098	4.140
What is your current occupation_Unemployed	-1.2752	0.334	-3.823	0.000	-1.929	-0.621
What is your current occupation_Working Professional	1.5402	0.412	3.742	0.000	0.733	2.347
Last Notable Activity_Olark Chat Conversation	-1.2150	0.352	-3.453	0.001	-1.905	-0.525
Last Notable Activity_SMS Sent	2.6184	0.116	22.656	0.000	2.392	2.845
Tags_Busy	3.5856	0.305	11.763	0.000	2.988	4.183
Tags_Closed by Horizzon	8.1714	0.751	10.882	0.000	6.700	9.643
Tags_Lost to EINS	7.8044	0.563	13.866	0.000	6.701	8.908
Tags_Ringing	-1.0903	0.319	-3.416	0.001	-1.716	-0.465
Tags_Will revert after reading the email	3.9190	0.206	19.070	0.000	3.516	4.322

	Features	VIF
3	What is your current occupation_Unemployed	4.76
11	Tags_Will revert after reading the email	4.09
10	Tags_Ringing	1.69
6	Last Notable Activity_SMS Sent	1.49
4	What is your current occupation_Working Profes...	1.48
8	Tags_Closed by Horizzon	1.31
2	Lead Origin_Lead Add Form	1.30
1	Total Time Spent on Website	1.13
7	Tags_Busy	1.12
9	Tags_Lost to EINS	1.12
0	Do Not Email	1.09
5	Last Notable Activity_Olark Chat Conversation	1.04

Final Conclusion

- Most important feature for model :
  - The total time spend on the Website.
  - When the lead origin is Lead add format.
  - When the current occupation was:
    - Unemployed
    - Working Professional
  - When the last notable activity was:
    - SMS
    - Olark chat conversation

- By focusing on these variable X Education Forms can grow and convert most lead.

Thank you