Lead Score Case Study

Group Members:
1. Ankur Bansal
2. Deepti Dixit
3. Prateek Verma

Problem Statement

- X Education is marketing its courses to industry professionals.
- 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.
- The company gets leads for its courses via websites, search engines like Google, past referrals etc., however its lead conversion rate is very poor, approx. 30%
- The company wants to increase its lead coversion rate to about 80% by identifying most pomising leads, i.e. the leads that are most likely to convert into paying customers.

Objective of the Project

To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Overall Approach

Data Preparation and Cleaning

- Processing data from Leads.csv file
- Identifying and treating missing and null values in data.
- Dealing with outliers.
- Analysing columns in dataframe.
- Binning and Grouping of Values.

EDA and Data Visualization

- Univariate and Bivariate Analysis
- Visualising categorical and numerical variables.
- Dummy Variable creation

Model Building

- Splitting data into test-train sets.
- Target Variable is Converted.
- Building Logistic Regression model using Stats Model and RFE

Model Training and Evaluation

- Training Logistic Regression Model on training data.
- Evaluating based on p-value and VIF score.
- Plotting ROC curve
- Finding Optimal Cut-off Point
- Checking performance of the model using metrics like accuracy, sensitivity, specificity, precision and recall.

Model Testing

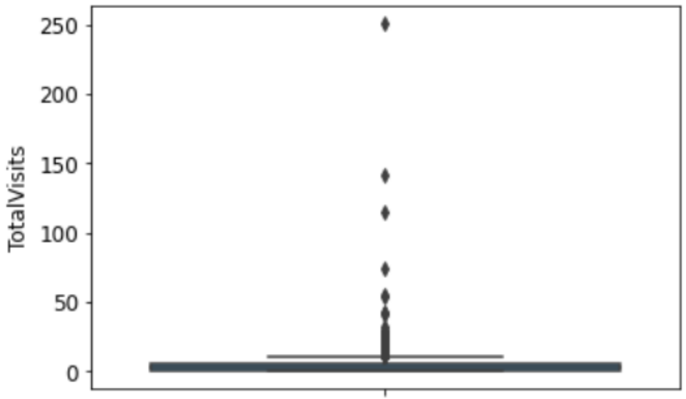
- Testing the model on test data.
- Extracting conclusion and recommendation.

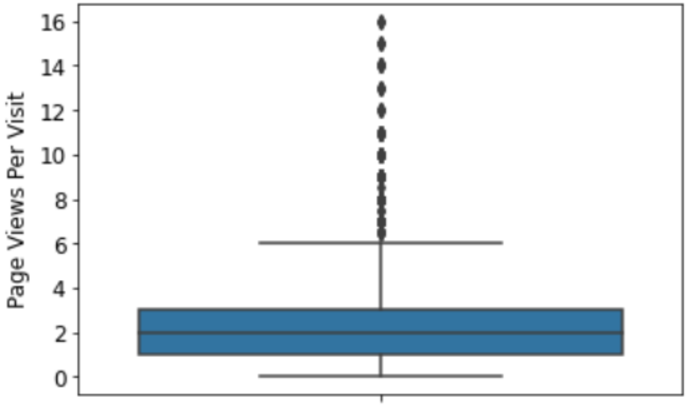
Data Preparation and EDA

Data Cleaning and Preparation

- Dataset: The dataset from X Education had 9000 data points set which consists of various attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. and target variable- ‘Converted’.
- The raw data had 37 columns (attributes) with 9240 rows of data.
- Columns with 'Select' as a value were removed and marked as NaN as part of cleaning.
- Attributes with only one unique value were removed as it will be of no use for us, such as Magazine, Receive More Updates About Our course etc.
- Columns with more than 45% of missing values were also dropped.
- Imputing Null Values- For instance, Specialization attribute is something which is not necessary that every candidate will have, so we have imputed null values with "Not Specified".
- Variables which does not provide diverse data were dropped, such as country column which had 95% of single value.

- To reduce variable in further analysis, values of columns were grouped/binned. For instance, we have grouped values of lead source, Specialization etc. which are less then or equal to 1 percent as others.
- Outliers were removed for numerical variables- Page Views Per Visit and TotalVisits.





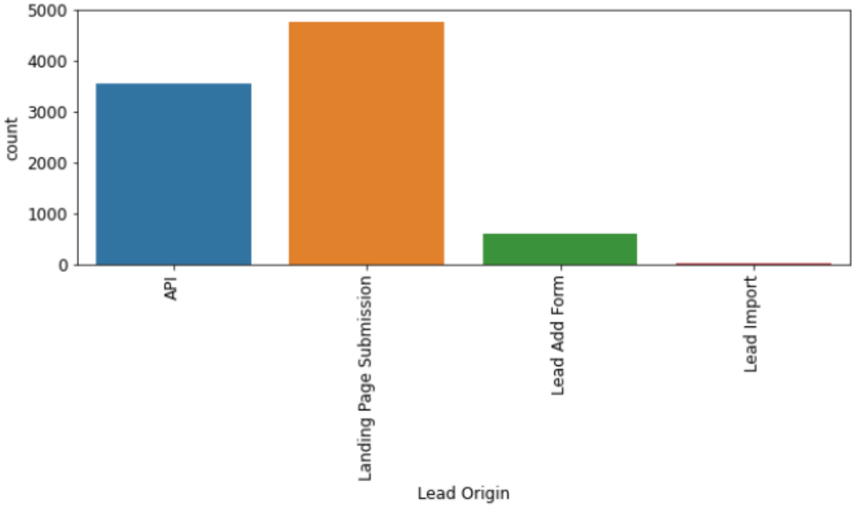
- After data cleaning process 97% of rows are retained and 38% of columns are retained.

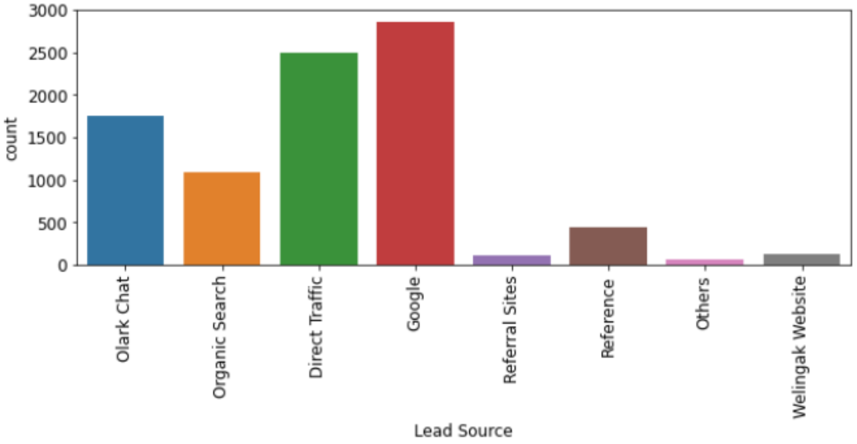
EDA and Data Visualization

- Observations from visualizing categorical variables:

i. Maximum leads are collected by landing page form submission.

ii. Maximum customers lands on website by finding it though google search engine or by directly typing web address of the website.

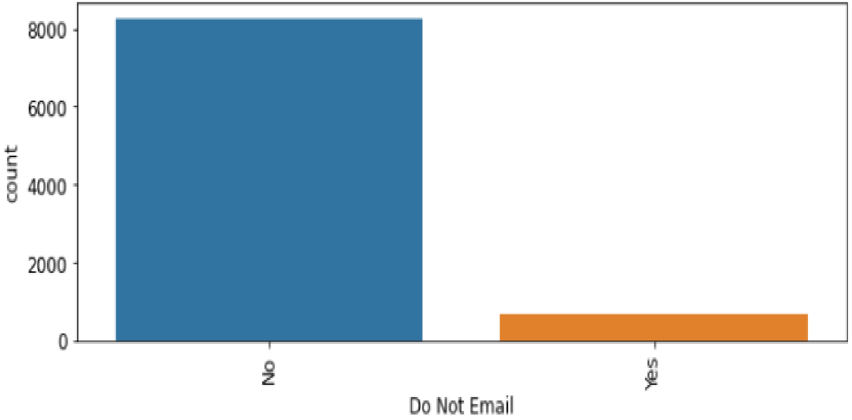


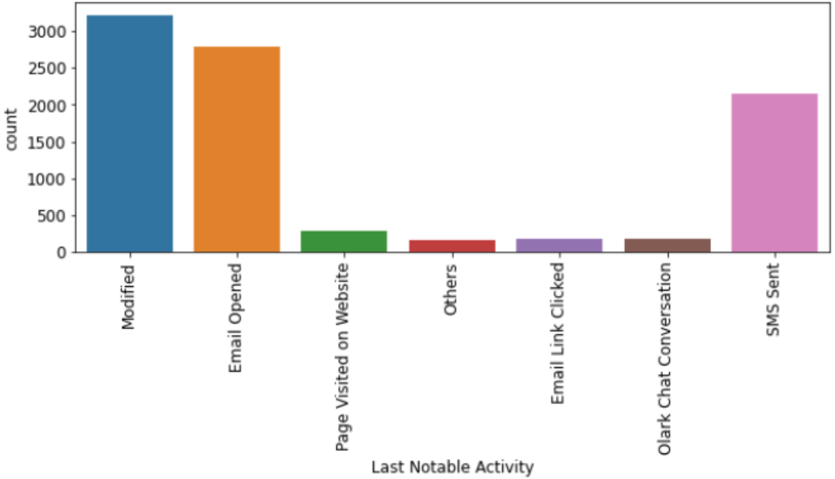


iii. Maximum customers prefer that we don't email them.

iv. It can be noticed that users open emails sent by the company but don’t often open the link in the email.

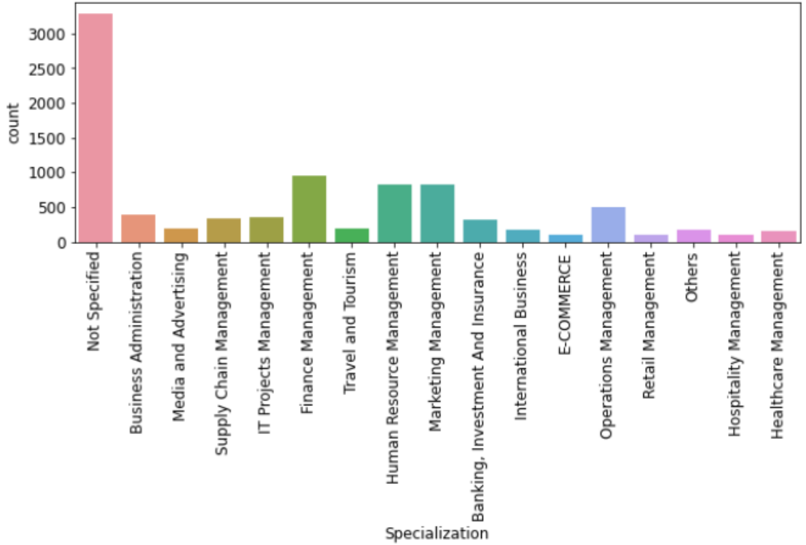
.

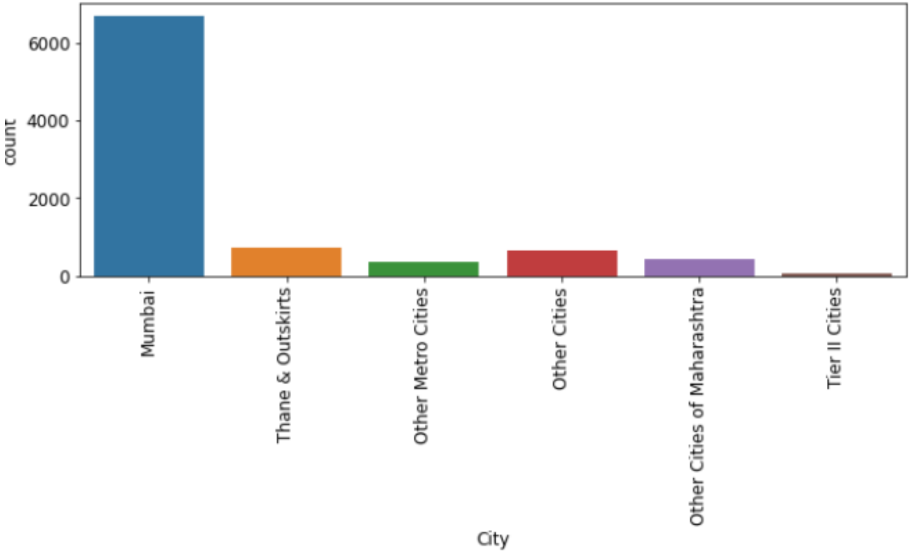




v. For Specialization, customers are either does not have any specialization (it can also mean that they are student) or they are from management background.

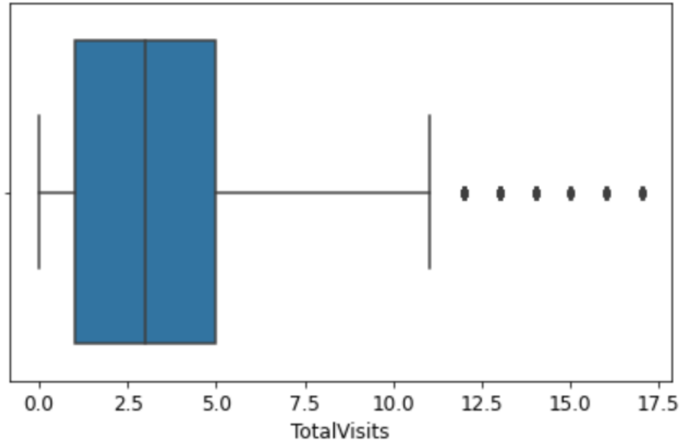
vi. Maximum customers are either from Mumbai or near Mumbai.

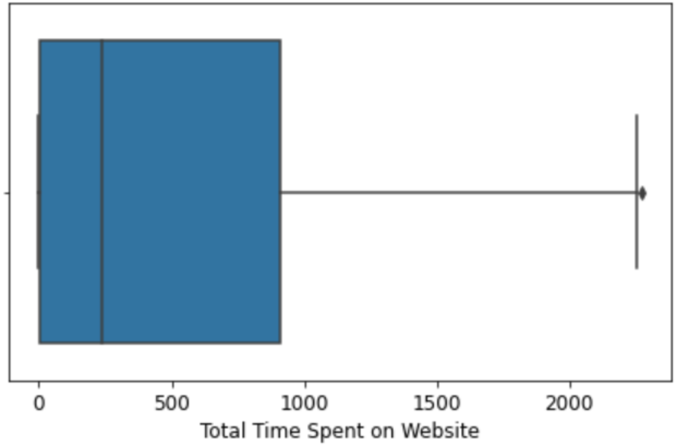


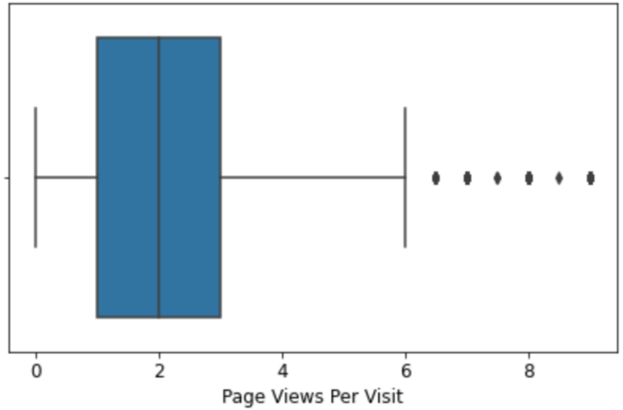


- Dummy variables were created for categorical data and original variables were dropped later on.

Analysis for Numerical Variables

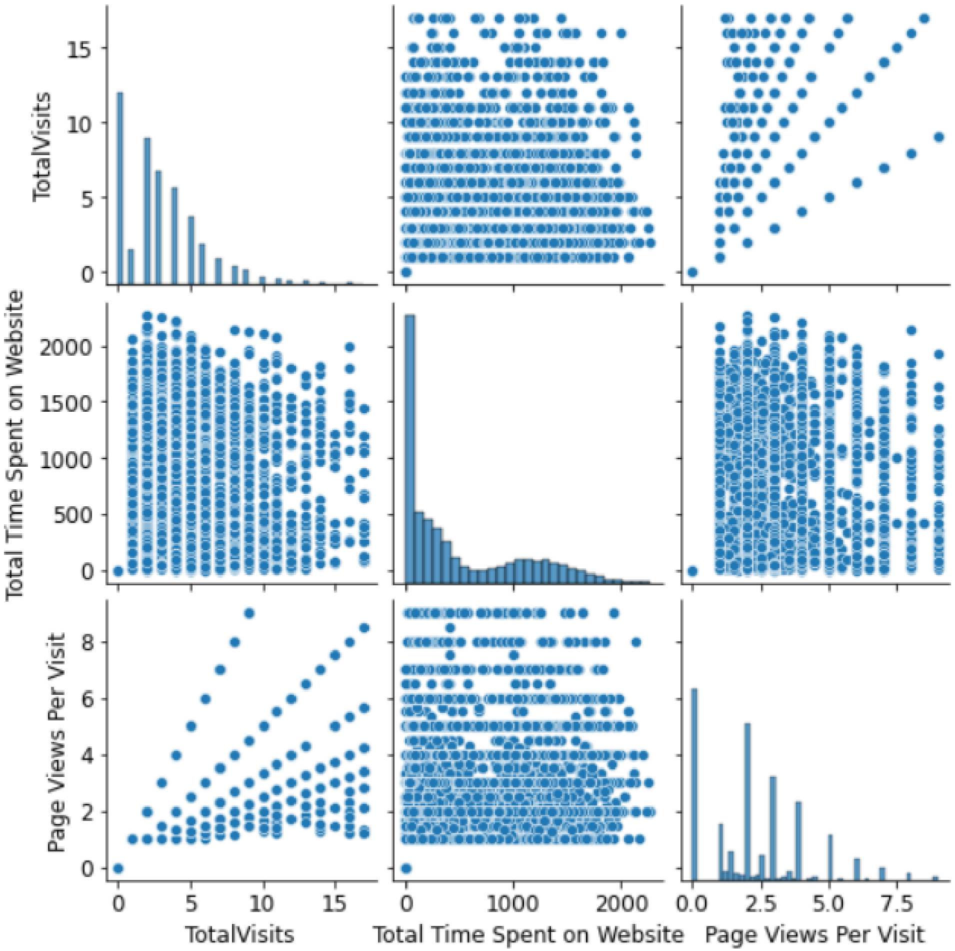






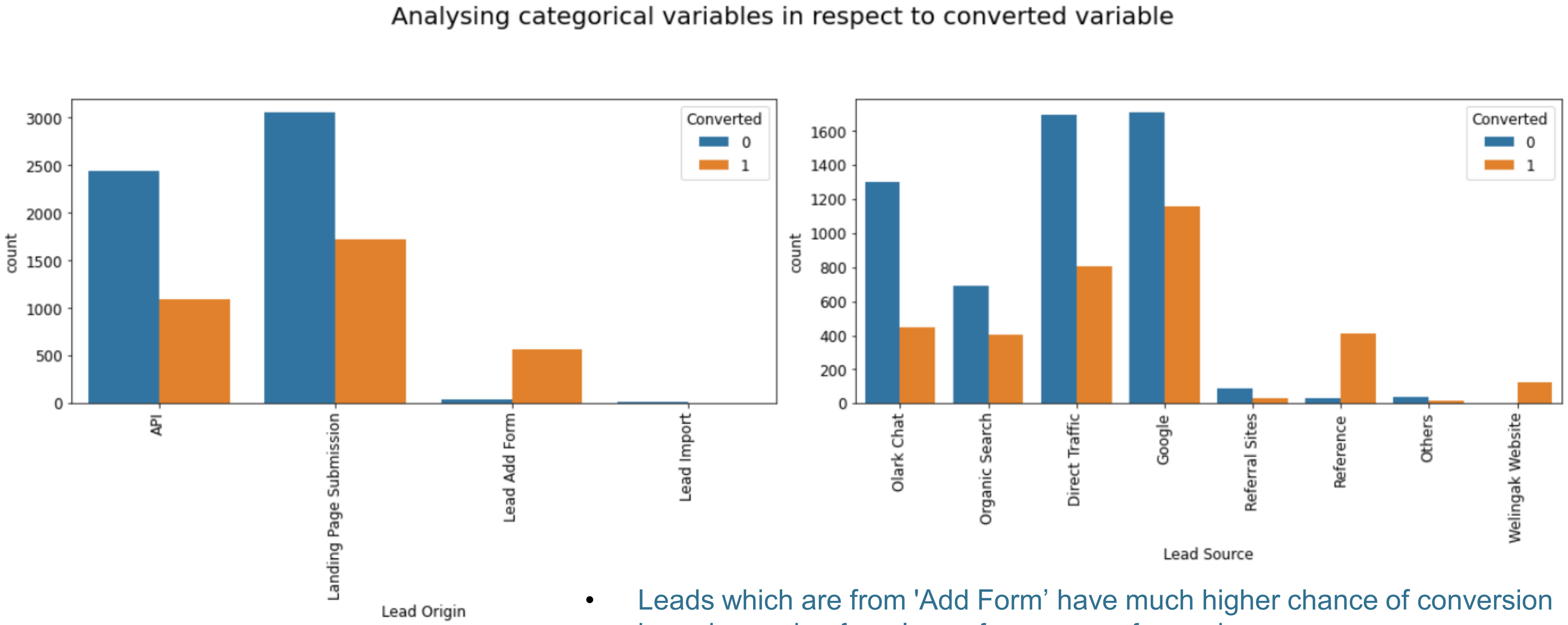
- Majority of people who apply for course visit 1-5 times on website.
- Majority of people who apply for course spent around 1-15 min on the website.
- Majority of people who apply for course browse 1-3 pages on the website

```
<Figure size 5760x4320 with 0 Axes>
```



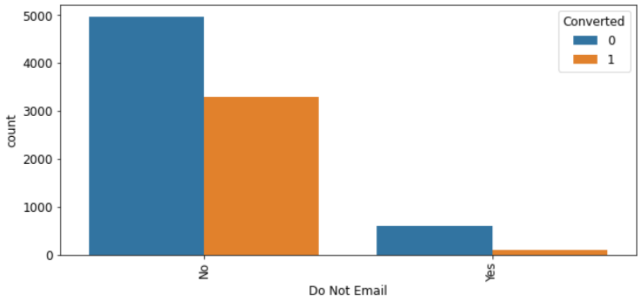
Linear relationship is identified between, total visits and page view per visit

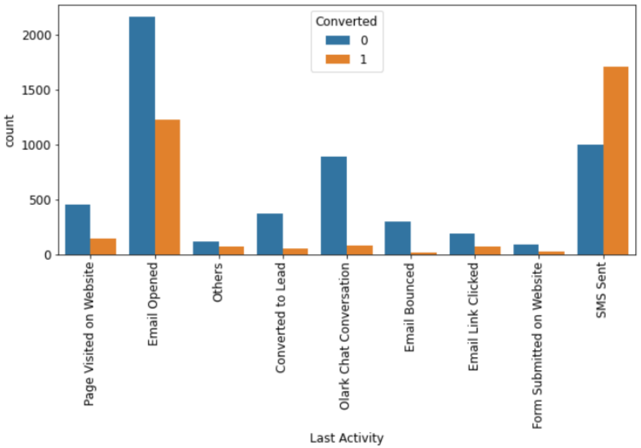
Analysis for Categorical Variables w.r.t Target Variable

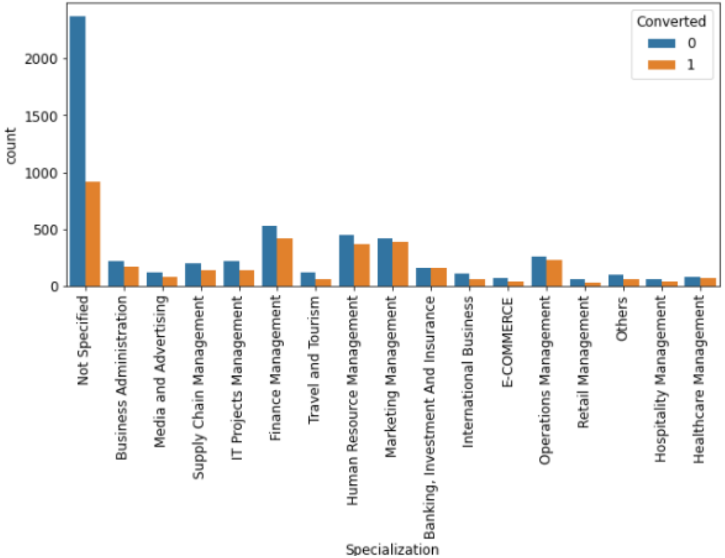


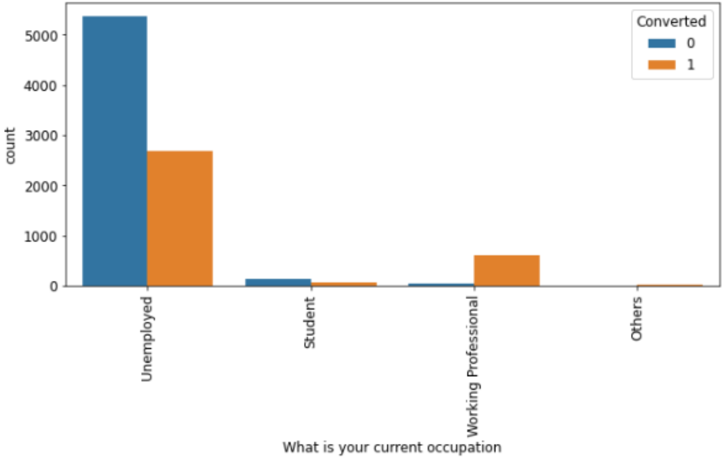
Analysing categorical variables in respect to converted variable

- Leads which are from 'Add Form’ have much higher chance of conversion based on ratio of total no. of responses for each category.
- Leads from reference has much higher chance of conversion in compare to other source based on ratio.

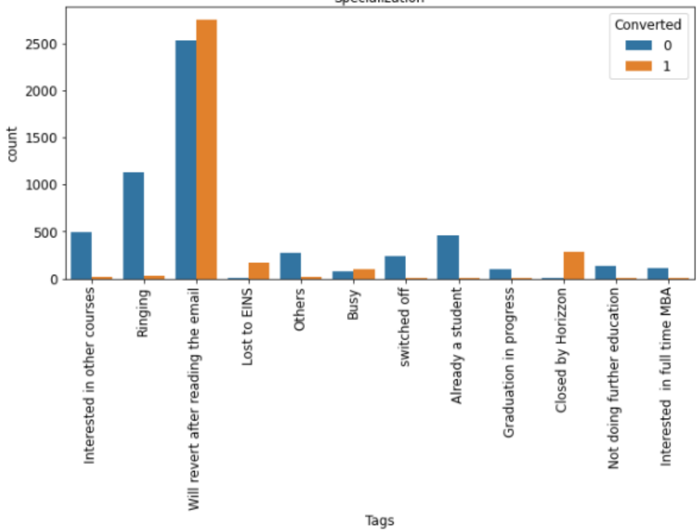


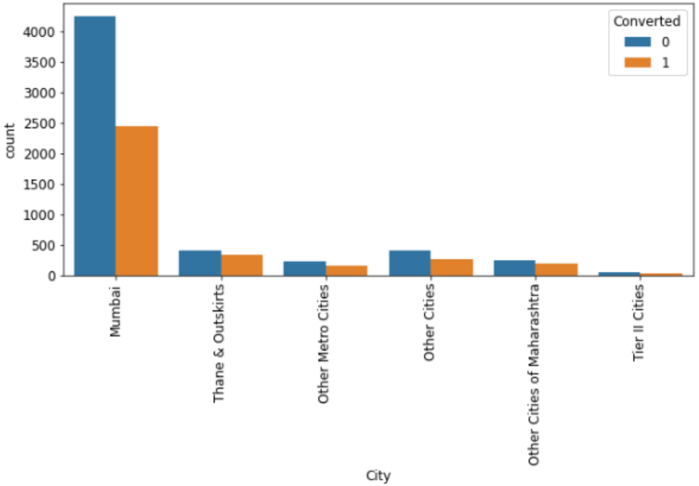


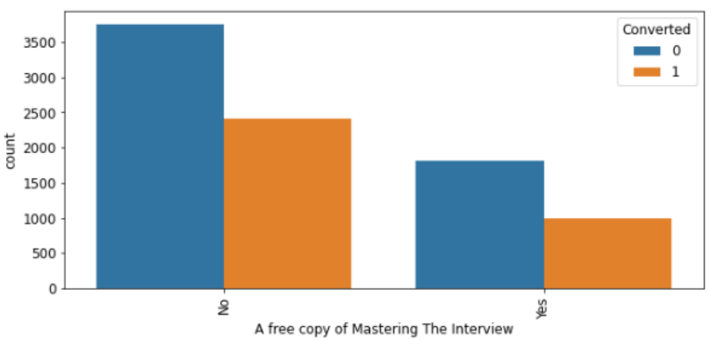


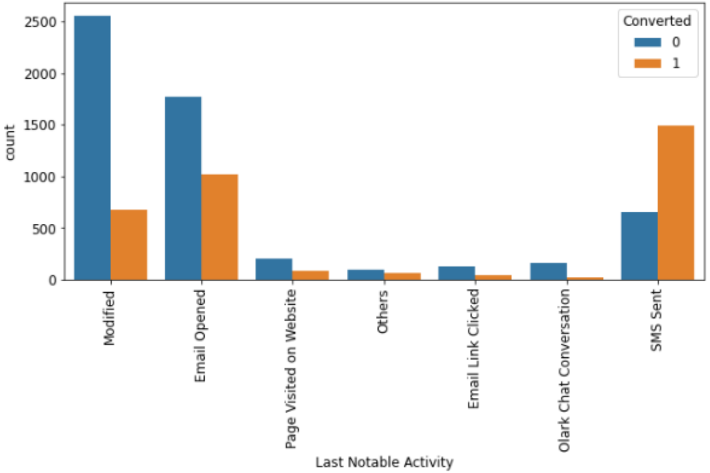


Working professionals has higher chance of conversion.

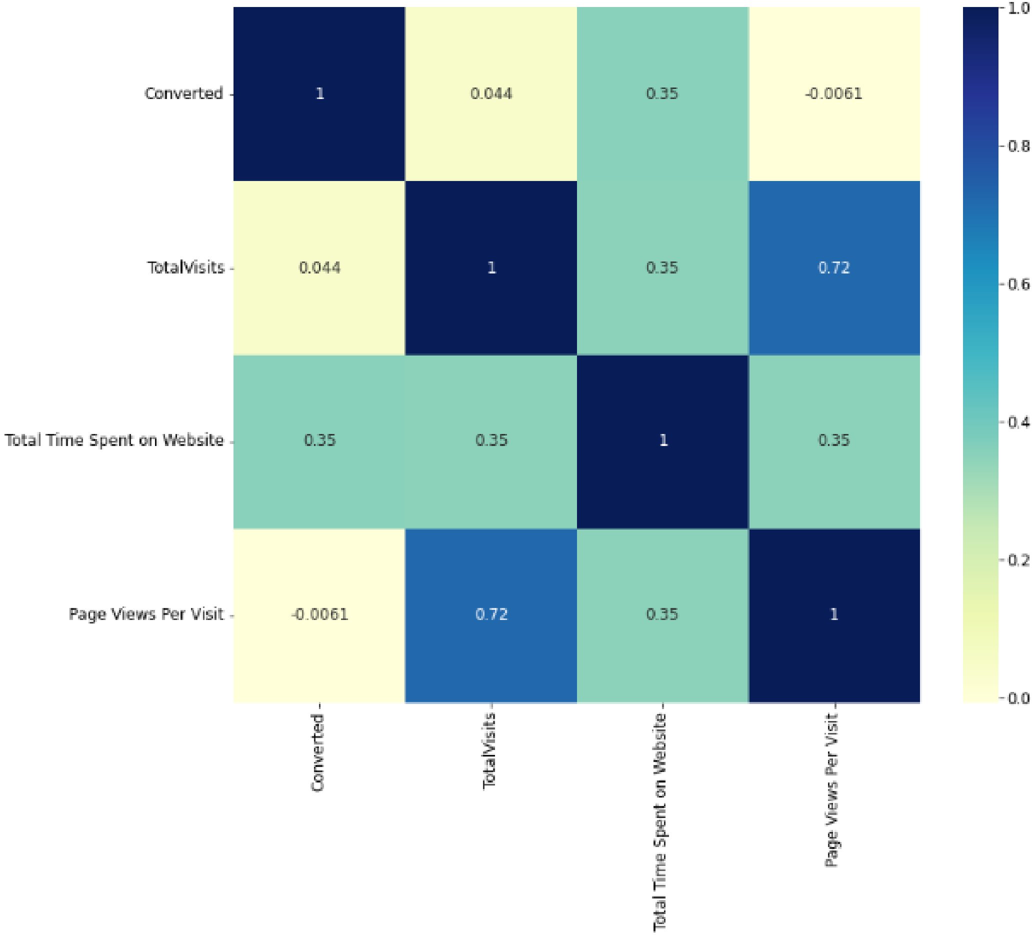








Conversion of customers is much higher using SMS marketing compared to email marketing based on ratio of total no. of responses for each category.



Correlation Matrix for Numerical Variables

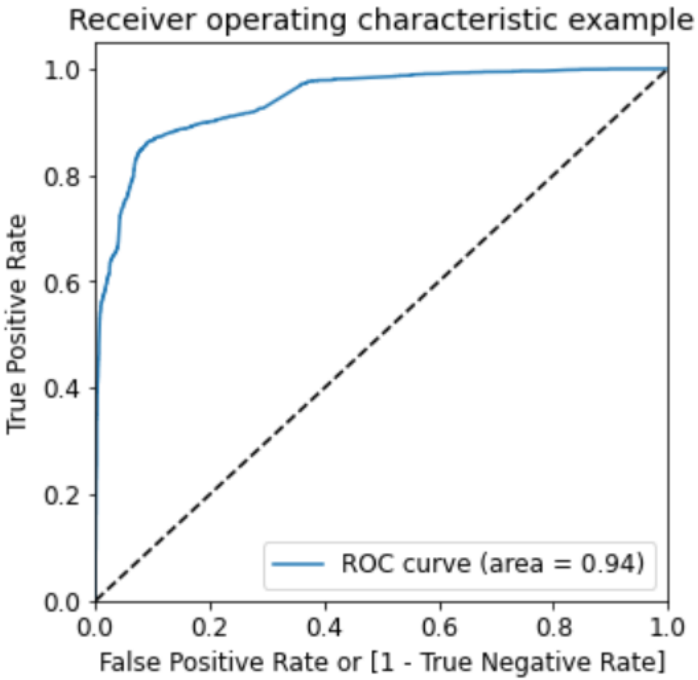
- Total Visits has positive high corelation with pages views per visit.
- Total time spend on website has positive high corelation with converted.
- Total visits has positive high corelation with Total time spent on website

Model Building and Evaluation

Model Building and Evaluation

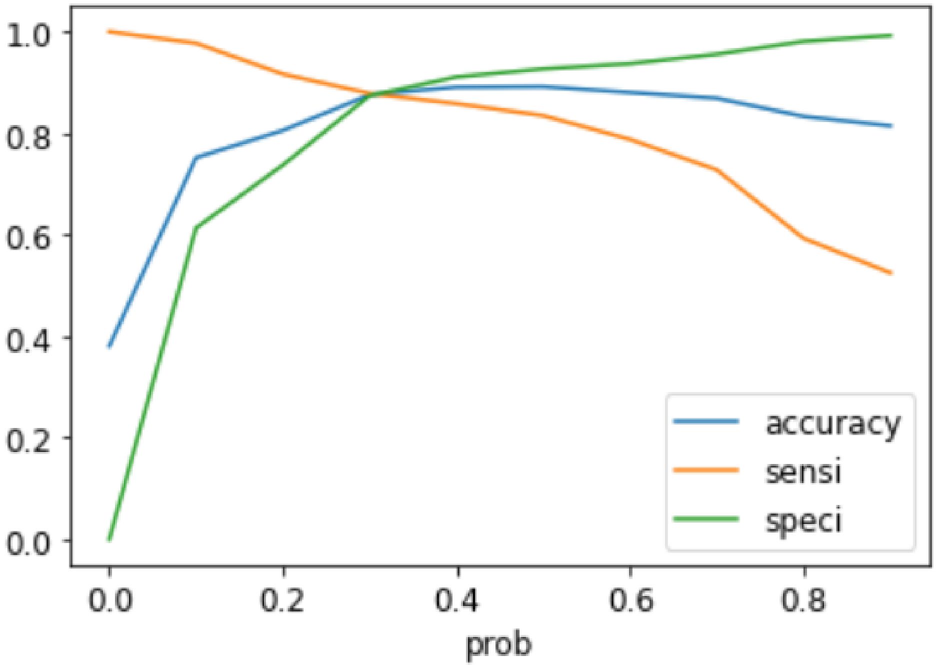
- The data was split into test and train sets with 70:30 ratio,
- Min-Max scaling was used to re-scale the numerical variables
- Logistic Regression model was built and RFE was used with 15 features for feature elimination.
- Model was iteratively built checking p-value and VIFs at each level and dropping columns with high VIF values (>5) and p-values (>0.05)
- Accuracy and Precision on training data at 0.3 Optimal Cut-Off was 89.18% and 81.07% respectively.
- Accuracy and Precision on test data was 89% and 80.9% respectively

ROC Curve



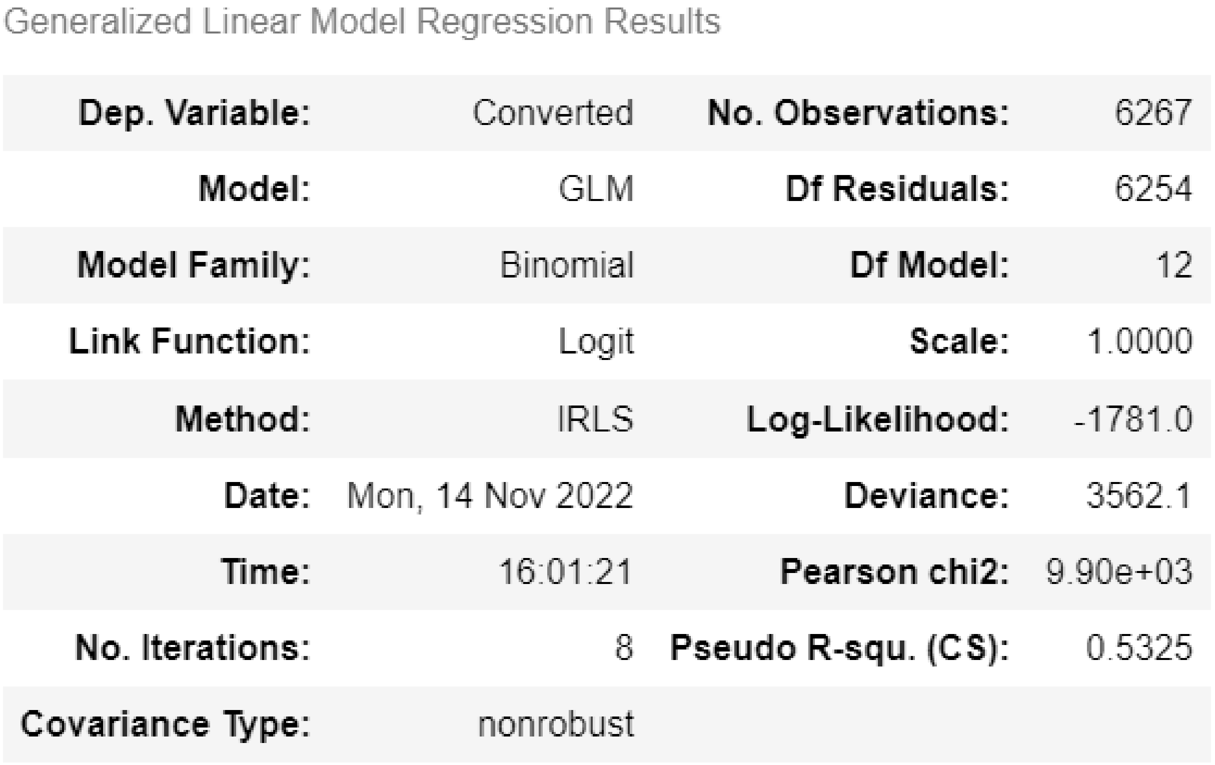
The ROC Curve should be a value close to 1. We are getting a good value of 0.94 indicating a good predictive model.

Finding Optimal Cut-off Point



As per the graph, 0.3 is the optimal point to take as a cut-off.

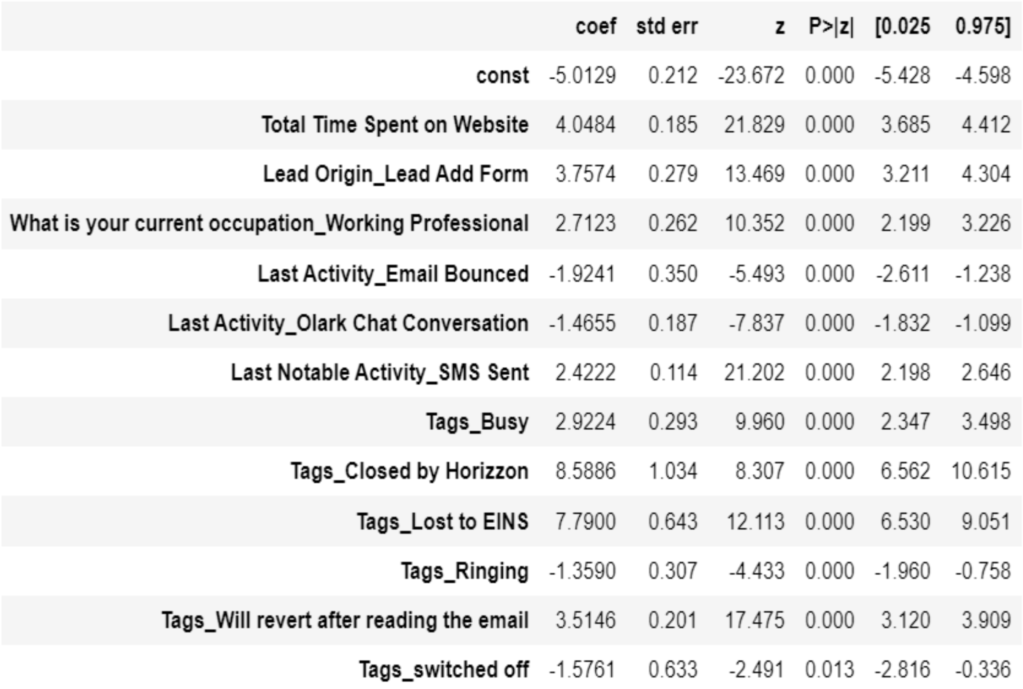
Final Logistic Regression Model



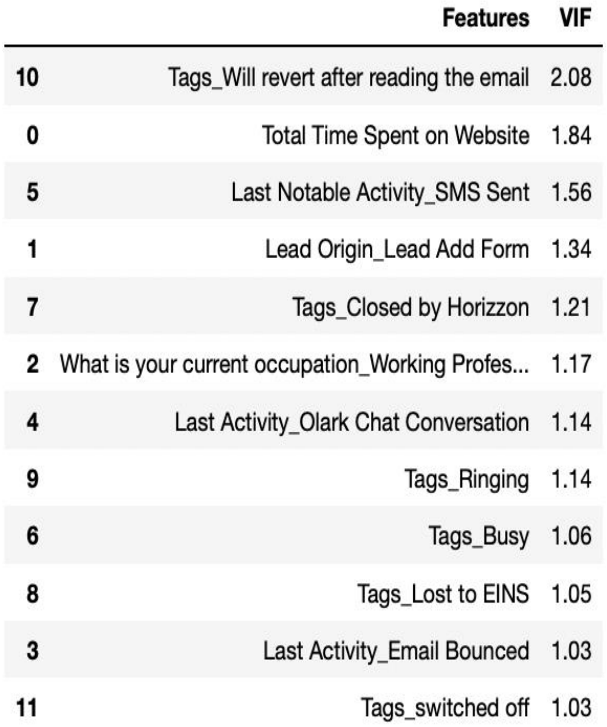
Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6267
Model:	GLM	**Df Residuals:**	6254
Model Family:	Binomial	**Df Model:**	12
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1781.0
Date:	Mon, 14 Nov 2022	**Deviance:**	3562.1
Time:	16:01:21	**Pearson chi2:**	9.90e+03
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.5325
Covariance Type:	nonrobust		

- All features p values are less then 0.05 at this stage.
- All features VIF are less then 5

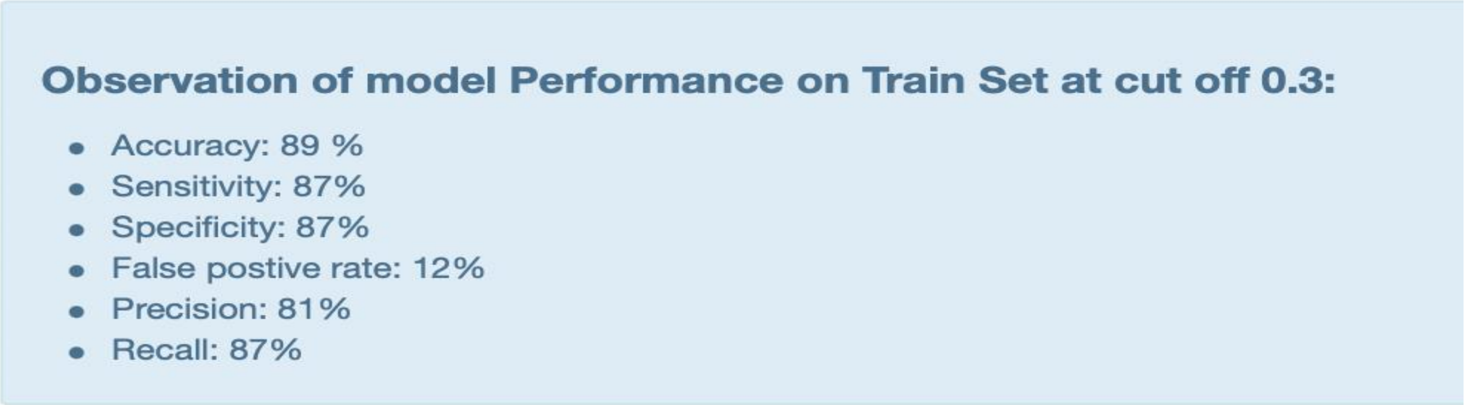


	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.0129	0.212	-23.672	0.000	-5.428	-4.598
Total Time Spent on Website	4.0484	0.185	21.829	0.000	3.685	4.412
Lead Origin_Lead Add Form	3.7574	0.279	13.469	0.000	3.211	4.304
What is your current occupation_Working Professional	2.7123	0.262	10.352	0.000	2.199	3.226
Last Activity_Email Bounced	-1.9241	0.350	-5.493	0.000	-2.611	-1.238
Last Activity_Olark Chat Conversation	-1.4655	0.187	-7.837	0.000	-1.832	-1.099
Last Notable Activity_SMS Sent	2.4222	0.114	21.202	0.000	2.198	2.646
Tags_Busy	2.9224	0.293	9.960	0.000	2.347	3.498
Tags_Closed by Horizzon	8.5886	1.034	8.307	0.000	6.562	10.615
Tags_Lost to EINS	7.7900	0.643	12.113	0.000	6.530	9.051
Tags_Ringing	-1.3590	0.307	-4.433	0.000	-1.960	-0.758
Tags_Will revert after reading the email	3.5146	0.201	17.475	0.000	3.120	3.909
Tags_switched off	-1.5761	0.633	-2.491	0.013	-2.816	-0.336



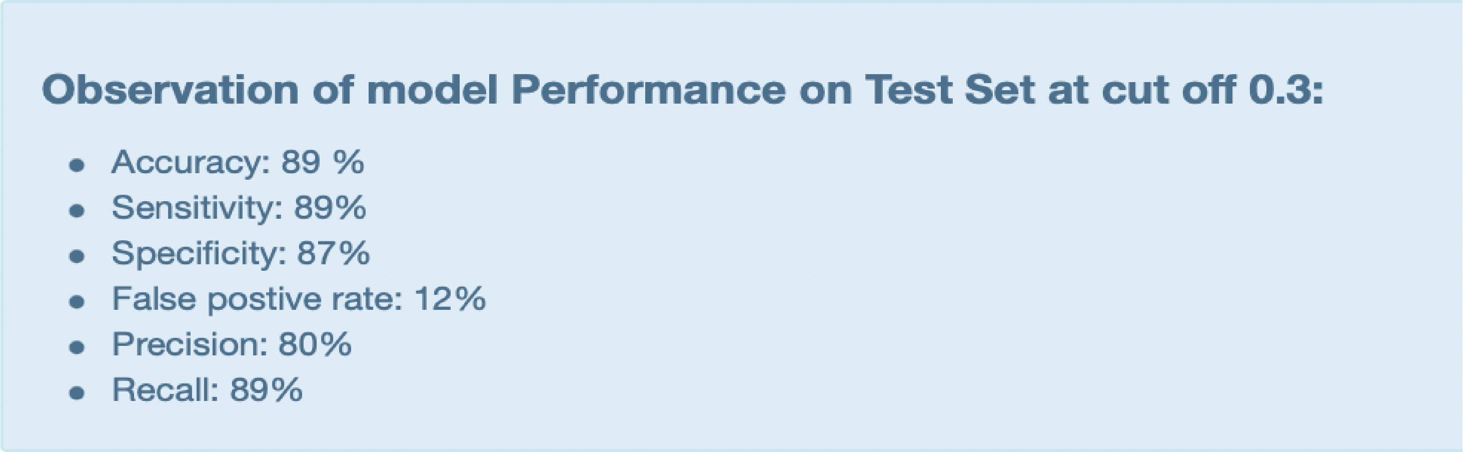
	Features	VIF
10	Tags_Will revert after reading the email	2.08
0	Total Time Spent on Website	1.84
5	Last Notable Activity_SMS Sent	1.56
1	Lead Origin_Lead Add Form	1.34
7	Tags_Closed by Horizzon	1.21
2	What is your current occupation_Working Profes...	1.17
4	Last Activity_Olark Chat Conversation	1.14
9	Tags_Ringing	1.14
6	Tags_Busy	1.06
8	Tags_Lost to EINS	1.05
3	Last Activity_Email Bounced	1.03
11	Tags_switched off	1.03

Model Evaluation



Observation of model Performance on Train Set at cut off 0.3:

- Accuracy: 89 %
- Sensitivity: 87%
- Specificity: 87%
- False postive rate: 12%
- Precision: 81%
- Recall: 87%



Observation of model Performance on Test Set at cut off 0.3:

- Accuracy: 89 %
- Sensitivity: 89%
- Specificity: 87%
- False postive rate: 12%
- Precision: 80%
- Recall: 89%

Conclusion and Recommendations

Top variables in model which contribute positively most towards the probability of a lead getting converted

- Tags_Closed by Horizon: With very high coefficient of 8.5886 indicates customers with this tag has very high probability to convert.
- Tags_Lost to EINS: With very high coefficient of 7.7900 indicates customers with this tag has very high probability to convert.
- "Total Time Spent on Website": With coefficient 4.0484 indicates customers who spent more time on website has good probability to convert.
- "Lead Origin_Lead Add Form" : With coefficient 3.7574 indicates customers who apply using "Add Form" has high probability to convert more.

Top variables in model which contribute negatively most towards the probability of a lead getting converted

- Last Activity_Email Bounced: With coefficient -1.9241 indicates customers who has provided wrong email id, has good probability to drop.
- Tags_switched off: With coefficient -1.5761 indicates customers who has phone switched off or provided wrong phone number, has good probability to drop.

Performance Evaluation.

- Overall Accuracy and Precision of the model is 89% and 80% respectively.
- The model has sensitivity of 89%, specificity of 87% and recall of 89%.

Recommendations:

- X Education must focus on working professionals who spent more time on website or have shown more interest going through courses, applied using Form on website and tagged with ‘Closed by Horizon’ and ‘Lost to EINS’ as they have higher probability of getting converted.
- Lead score on model run will help to identify hot leads, and hence sales team must spend more time on those users with higher leads.
- As per the analysis, users respond or prefer SMS as compared to emails, and hence sales team must focus on that strategy.

Thank You!