X EDUCATION

LEAD SCORING CASE STUDY

● ● ● ●

Padmakara Srinivas

Deepa Duraisamy

PROBLEM STATEMENT

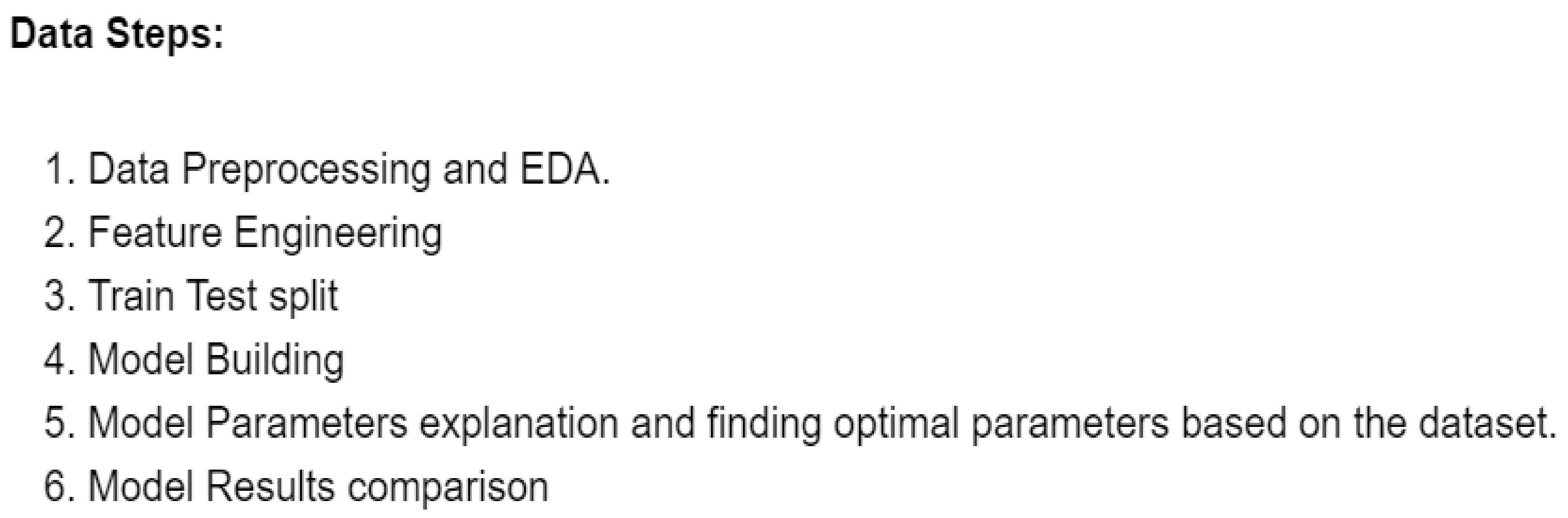
An organization, X Education needs help to select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires us to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

There are some more problems presented by the company which the model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well.

FINAL OBJECTIVE

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.

APPROACH



Data Steps:

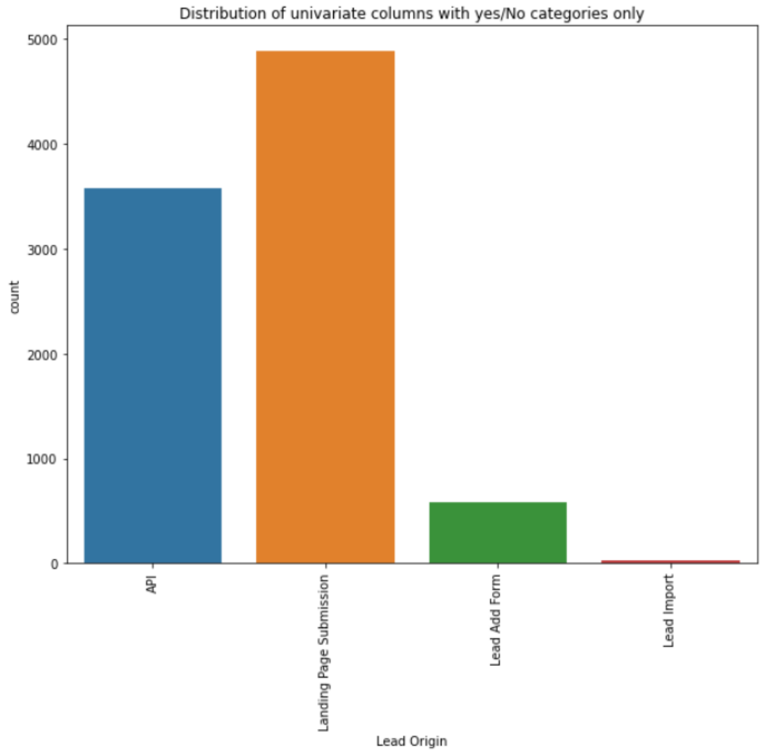
1. Data Preprocessing and EDA.
2. Feature Engineering
3. Train Test split
4. Model Building
5. Model Parameters explanation and finding optimal parameters based on the dataset.
6. Model Results comparison

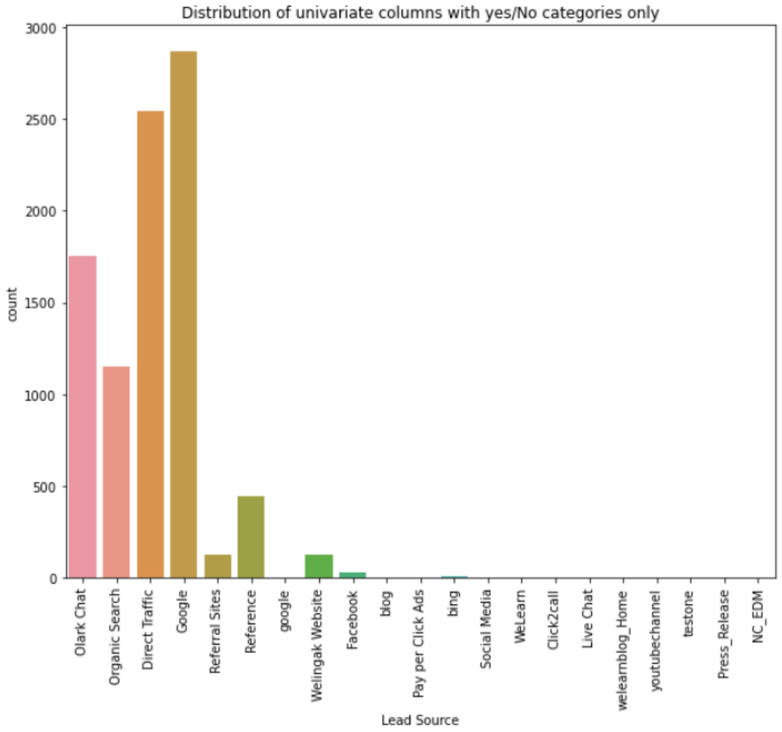
DATA PREPROCESSING

- Imported all required libraries, packages
- Read the csv file
- Analyze the data set using info(), shape and check statistical info of the dataset using describe()
- Head() to check data of first 5 records to understand the data
- Handle missing values – we used a 40% threshold – and dropped columns having more than 40% missing data. Out of 37 columns, 7 columns had data missing for more than 40% records. These columns were dropped.
- Converted Yes/No columns to 1s and 0s.
- Check columns where there is only one single value for the column – in this case, the column do not have a differentiating impact on the target variable and can be dropped – 5 columns have the same values for the entire column and hence can be eliminated.
- 2% of the rows have less than 5% data across. Hence, these rows can be dropped. Since this is only 2% of the dataset, it will not impact the overall statistics or the model.

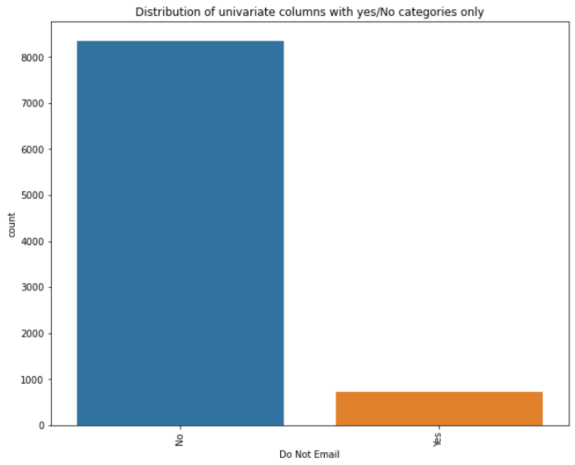
EDA INSIGHTS

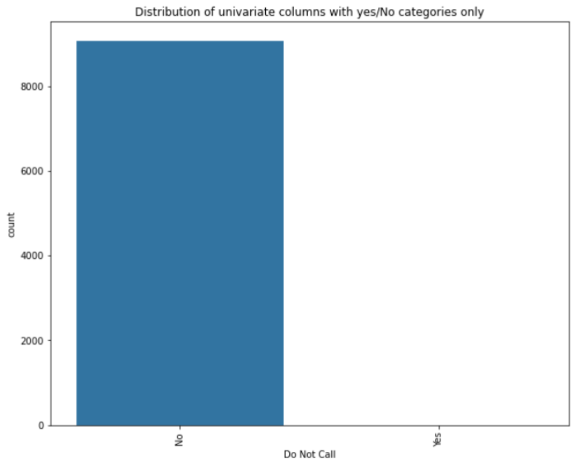
- We performed the univariate analysis of columns which have only Yes/No values.
- For other columns, check distribution across dataset. Categorical values with low representation are clubbed as Others.
- Distributions also with respect to the target variable Converted Yes/No
- Outlier treatment is not done formally since total visits, or page views could be genuine values.

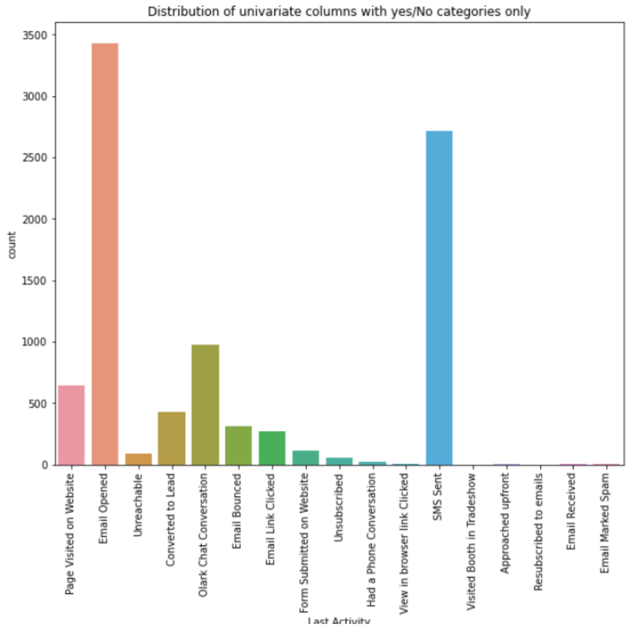


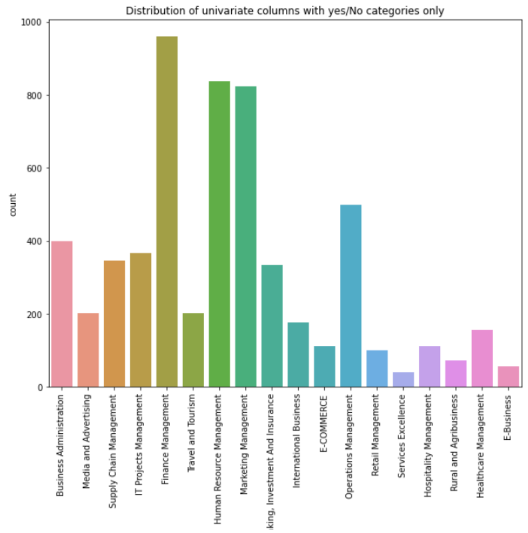


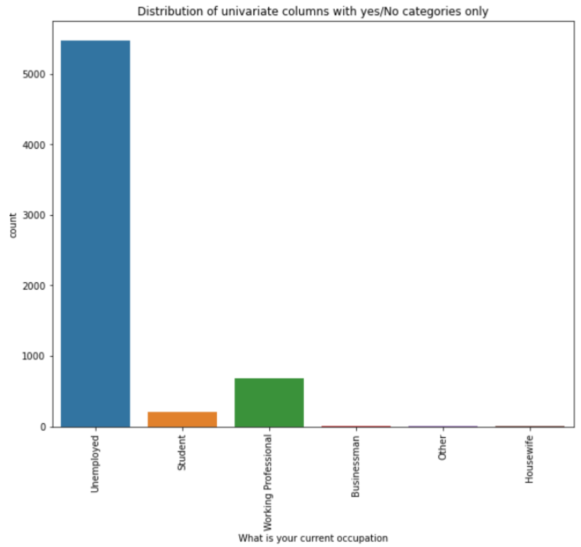
EDA INSIGHTS

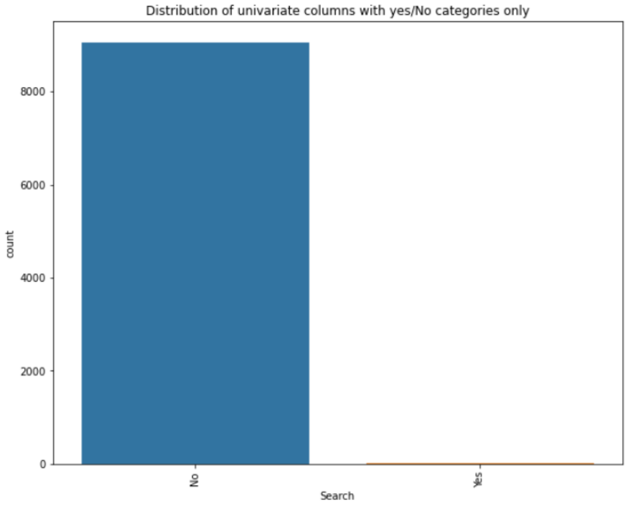




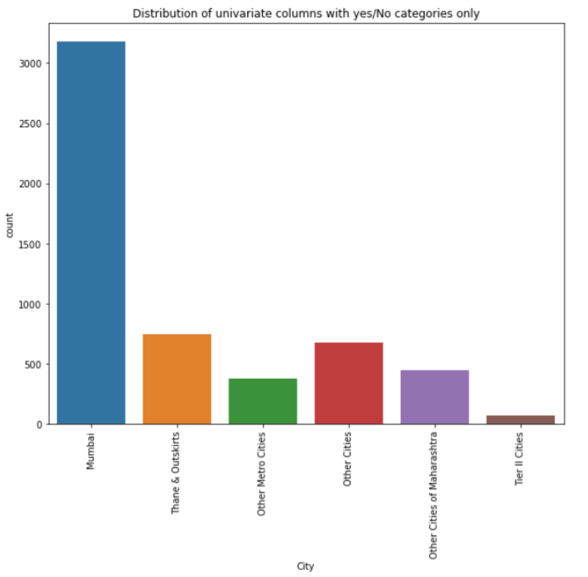


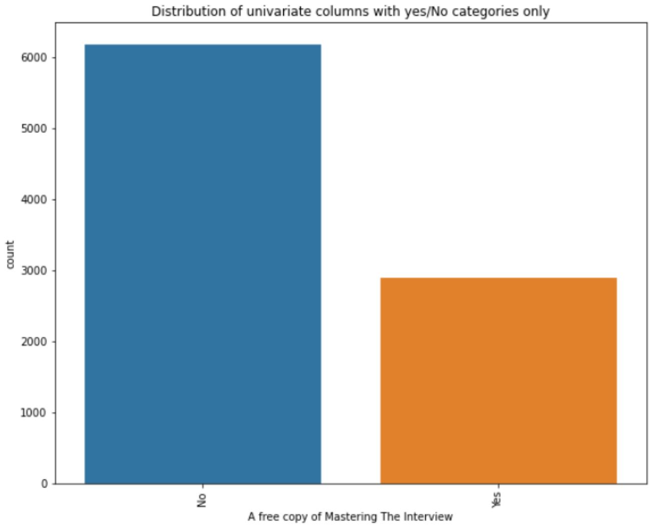


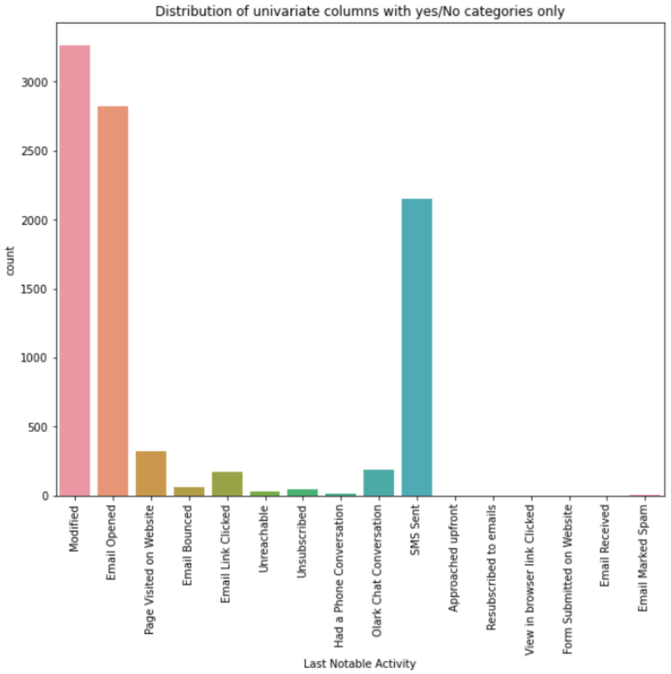




EDA INSIGHTS



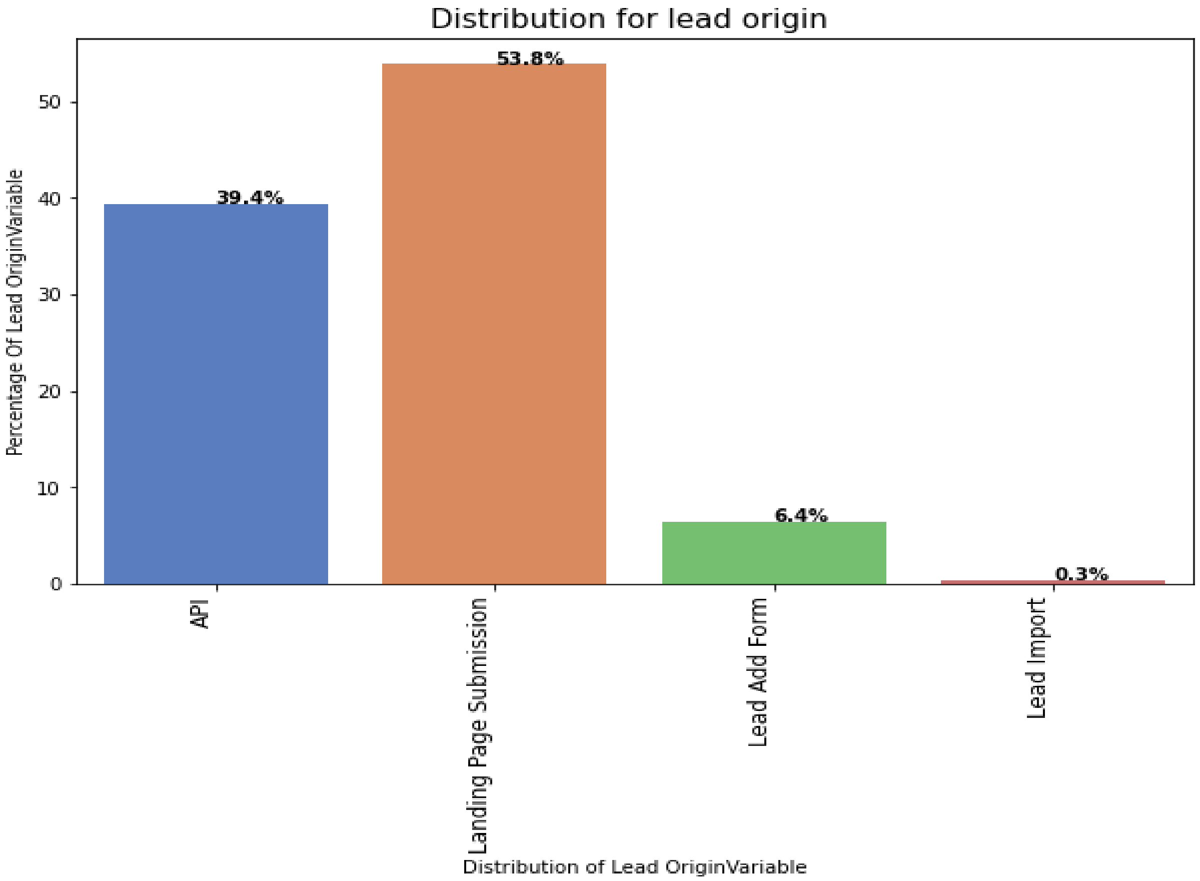




Insights:

1. Since Most of the yes/No columns have only One category(>90%), the column value does not significantly affect the target column Converted.
2. So Eliminating these columns("Do Not Call","Country","Search","What matters most to you in choosing a course","Newspaper Article","X Education Forums","Newspaper","Digital Advertisement","Through Recommendations") from the original dataset.

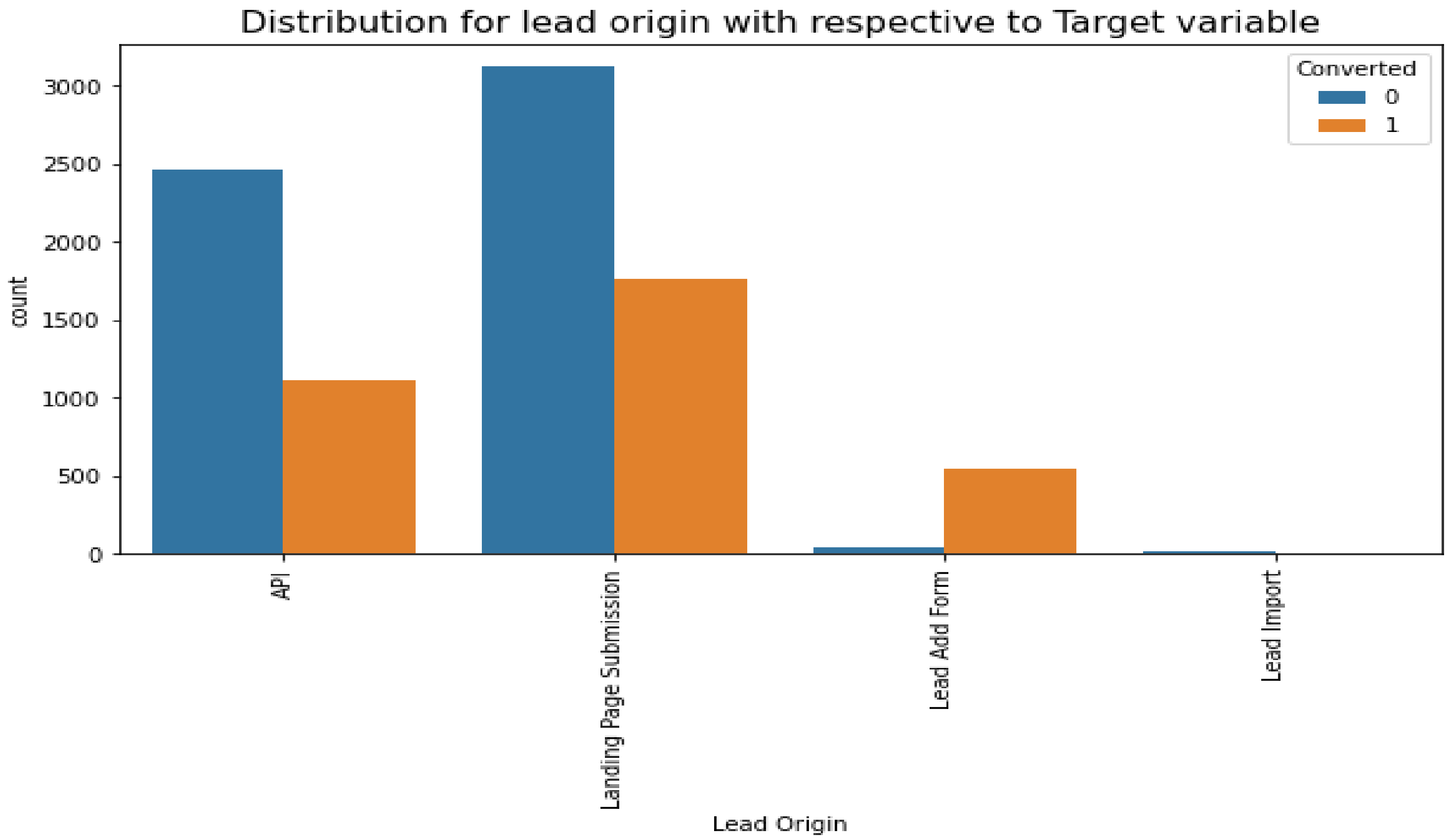
EDA INSIGHTS



Insights:

1. More than 90% of the data is through 2 lead origins only - API based on Landing Page submission.
2. The remaining 2 category values make up about 7% - but they could still be significant in their impact on Converted, these values will be retained.
3. It is also important to understand how this 90% data plays a role w.r.t values of the target variable Converted. We shall look at this next.

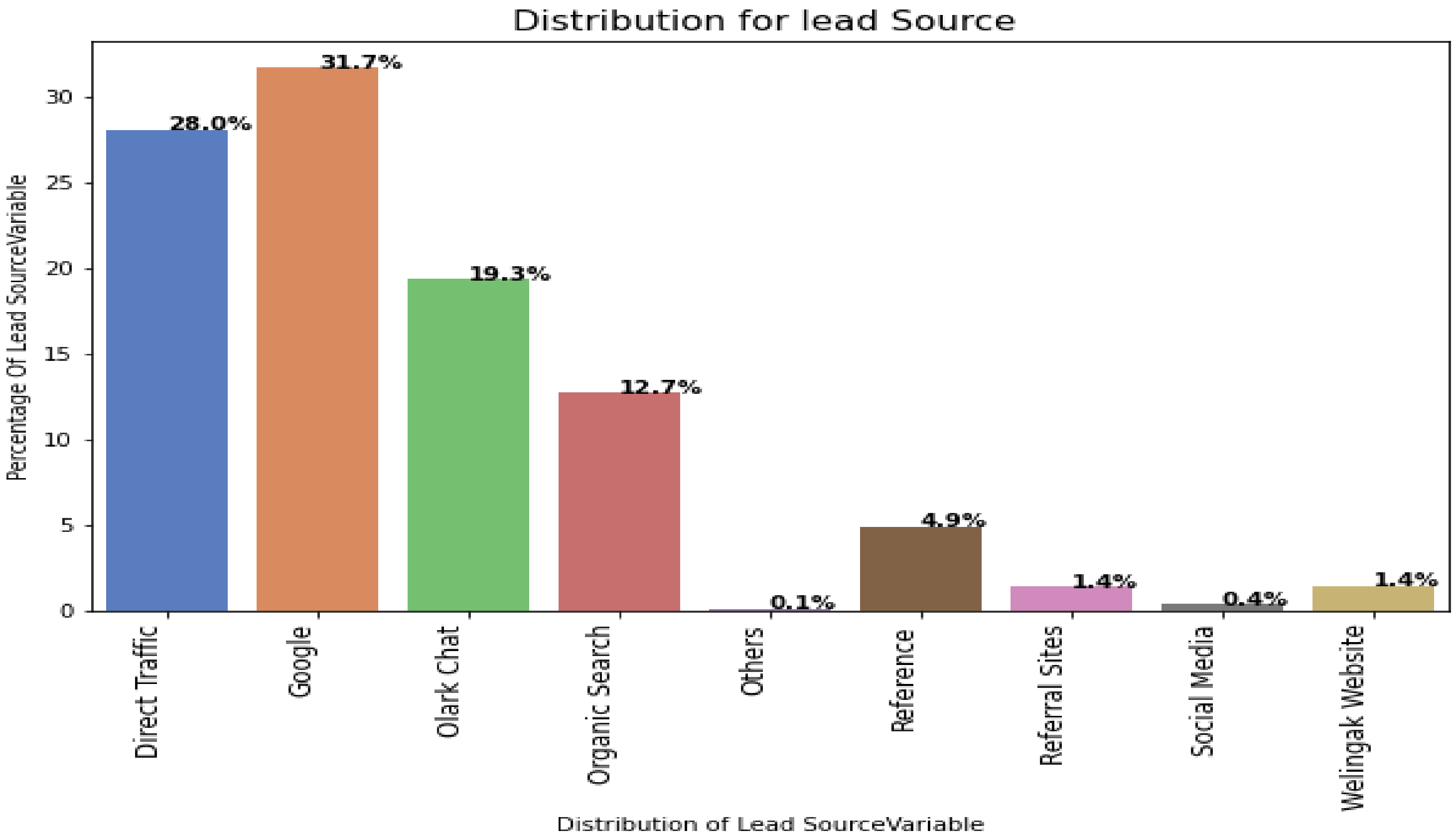
EDA INSIGHTS



Insights:

1. While API and Landing page submission have more data (93% data across the dataset), only about 50% of those convert.
2. Contrary to these, the Lead Add form, contributing to only 6% data set across the dataset, converts more than 80% of those.
3. The Lead Add form could be a potential focus area to spend more efforts on, and convert more prospects successfully.

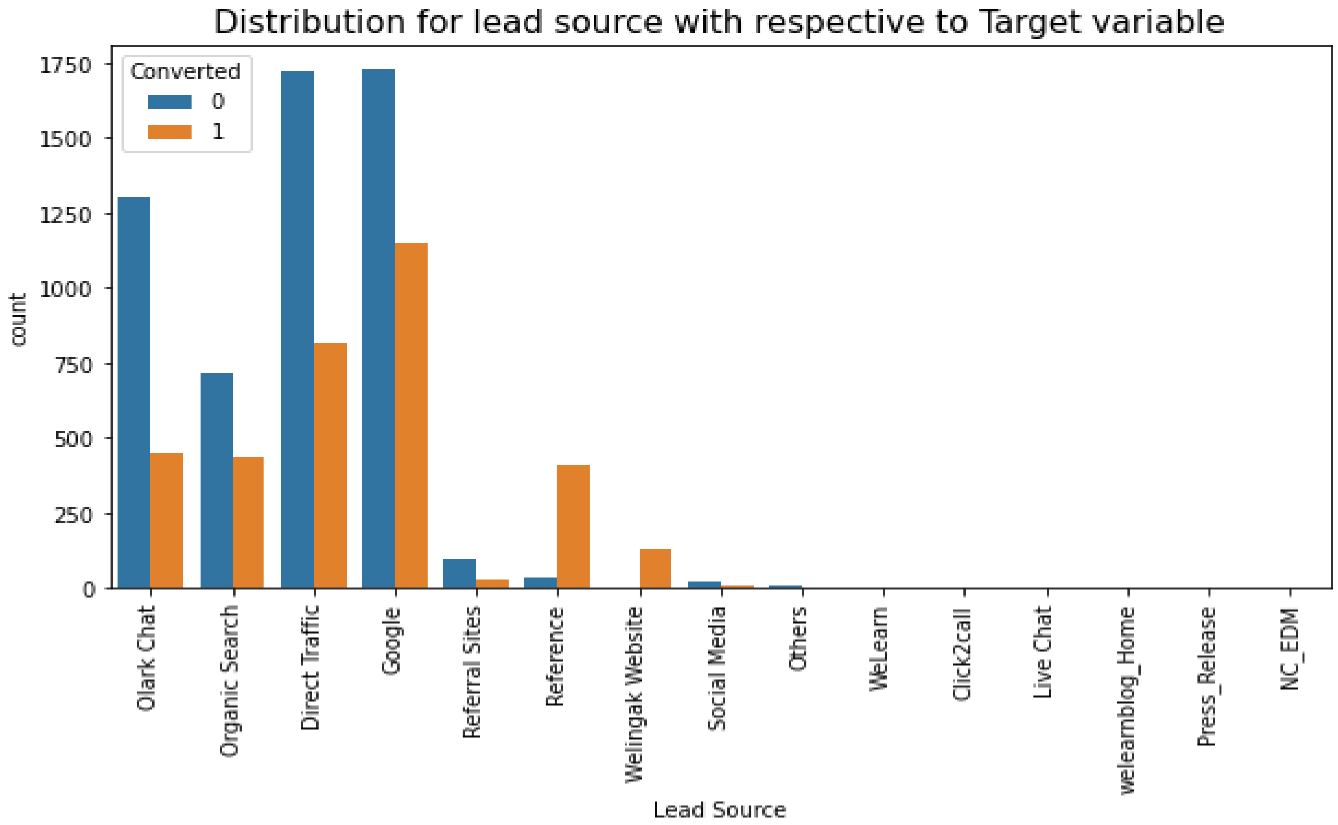
EDA INSIGHTS



Insights:

1. 4 Categories - Direct Traffic, Google, Olark Chat and Organic Search make up for more 90% of the dataset.
2. Categories clubbed into Others make up only 0.1% of the total dataset.
3. We will retain the remaining values which make up approx 10%. These will be evaluated against the target variable Converted.

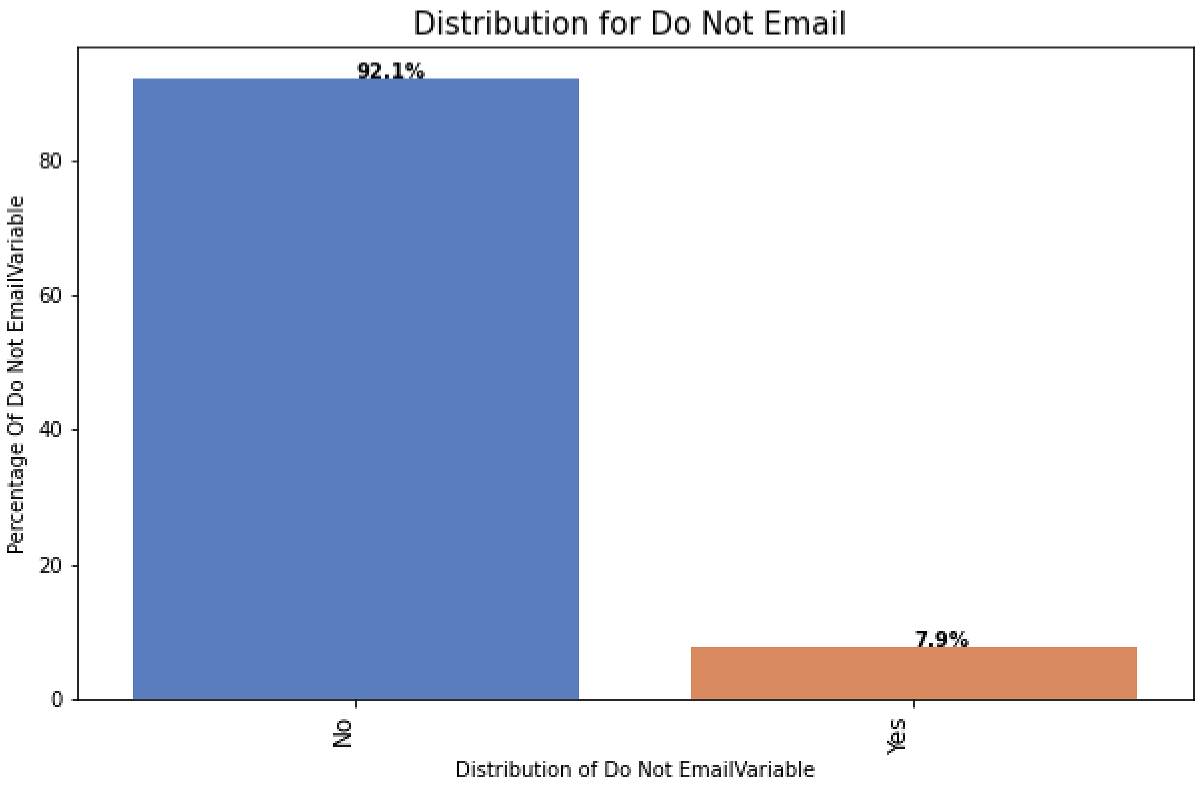
EDA INSIGHTS



Insights:

1. Of the 4 categories contributing to 90% in the dataset, Organic Search and Google look to have a relatively higher conversion ratio.
2. Reference and Welingak Website, while <10% of cases, seem to have a very strong conversion rate. These should be focused on more.
3. Olark Chat and Direct Traffic has a lot of prospects but conversion ratios are about 30% and 50% respectively. This could still be utilized when the company wants to go aggressive.

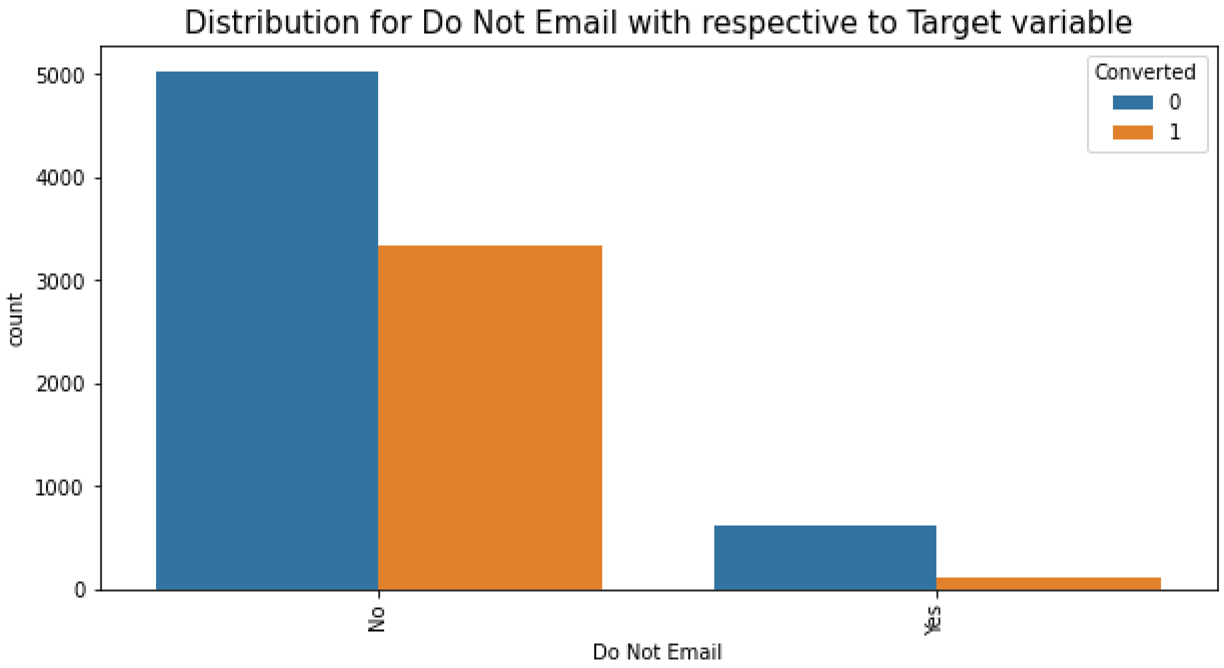
EDA INSIGHTS



Insights:

- The field name is Do Not Email. Here, YES would indicate selecton. 8% are not interested in receiving emails. Almost 92% of the users want an email related to the course.

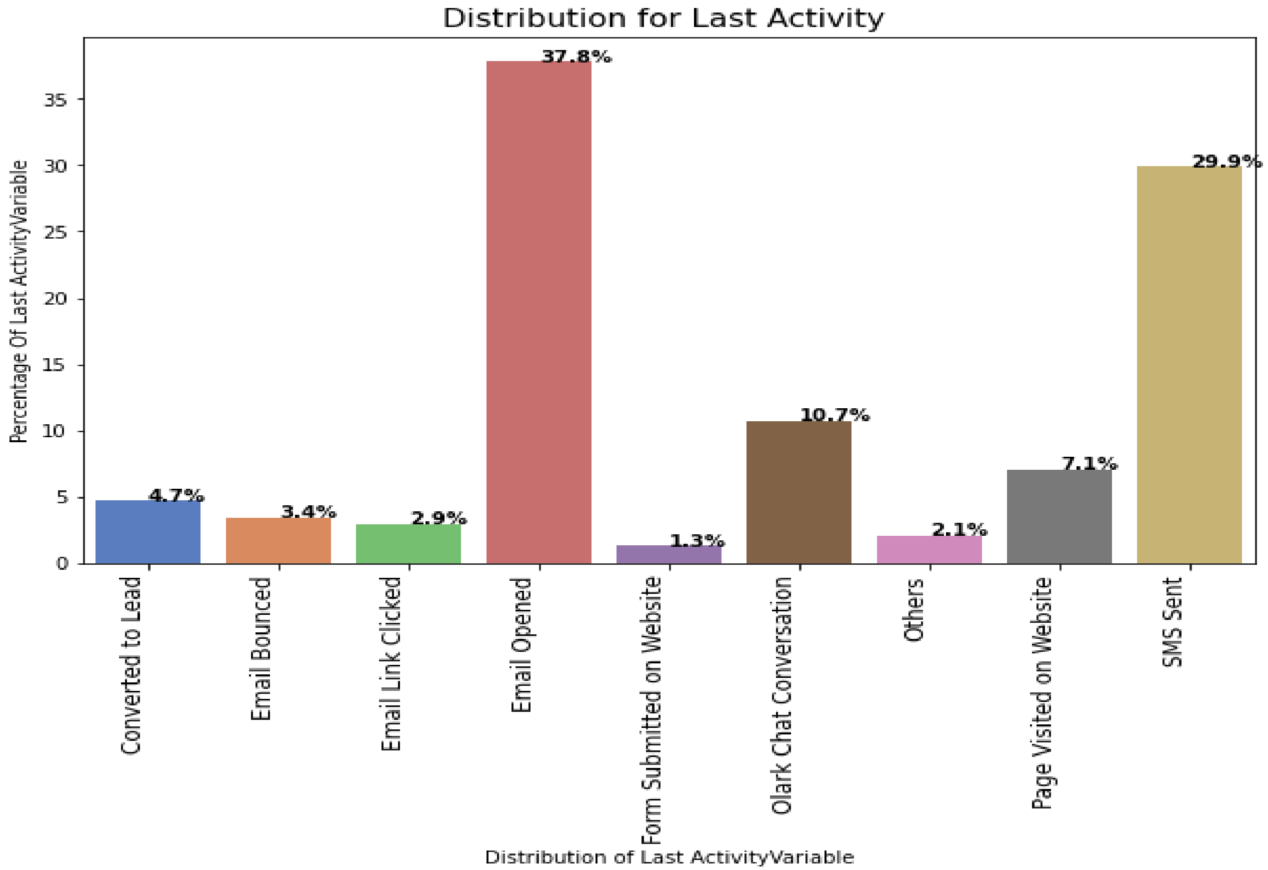
EDA INSIGHTS



Insights:

1. With users choosing to receive email, the % of conversion is approx 70%. This is a good focus area for conversion.
2. Additionally where users do not want to receive email, the % of conversion is low but still significant. In times of aggressive marketing and conversion, this should be picked up.

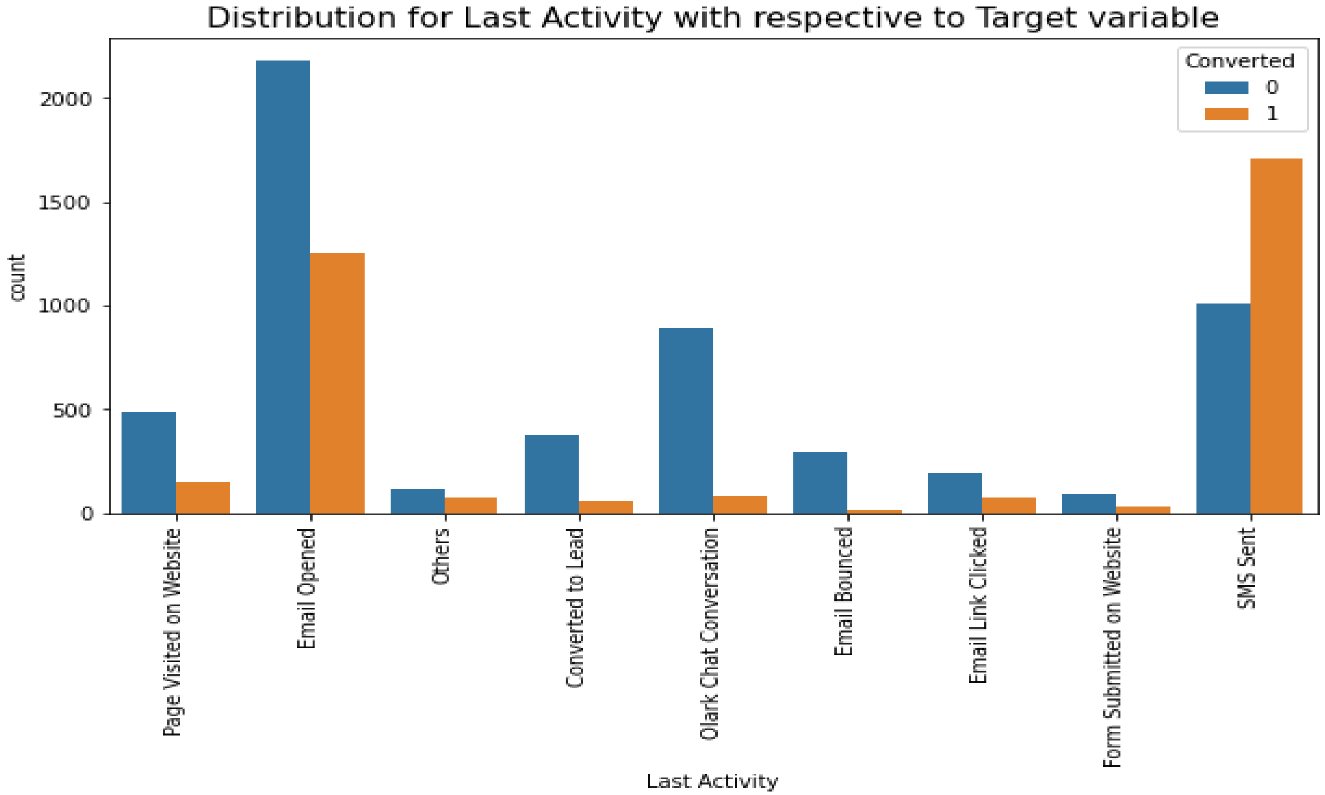
EDA INSIGHTS



Insights:

1. SMS Sent, Email Opened and Olark Chat Conversation together make up for about 70% of the dataset.
2. Others as a category (made up of smaller value categories) accounts for only 2% of the data.
3. Other instances of Last Activity - are smaller overall about 15-20% however will need to evaluate these using the Converted values to check how significant these are.

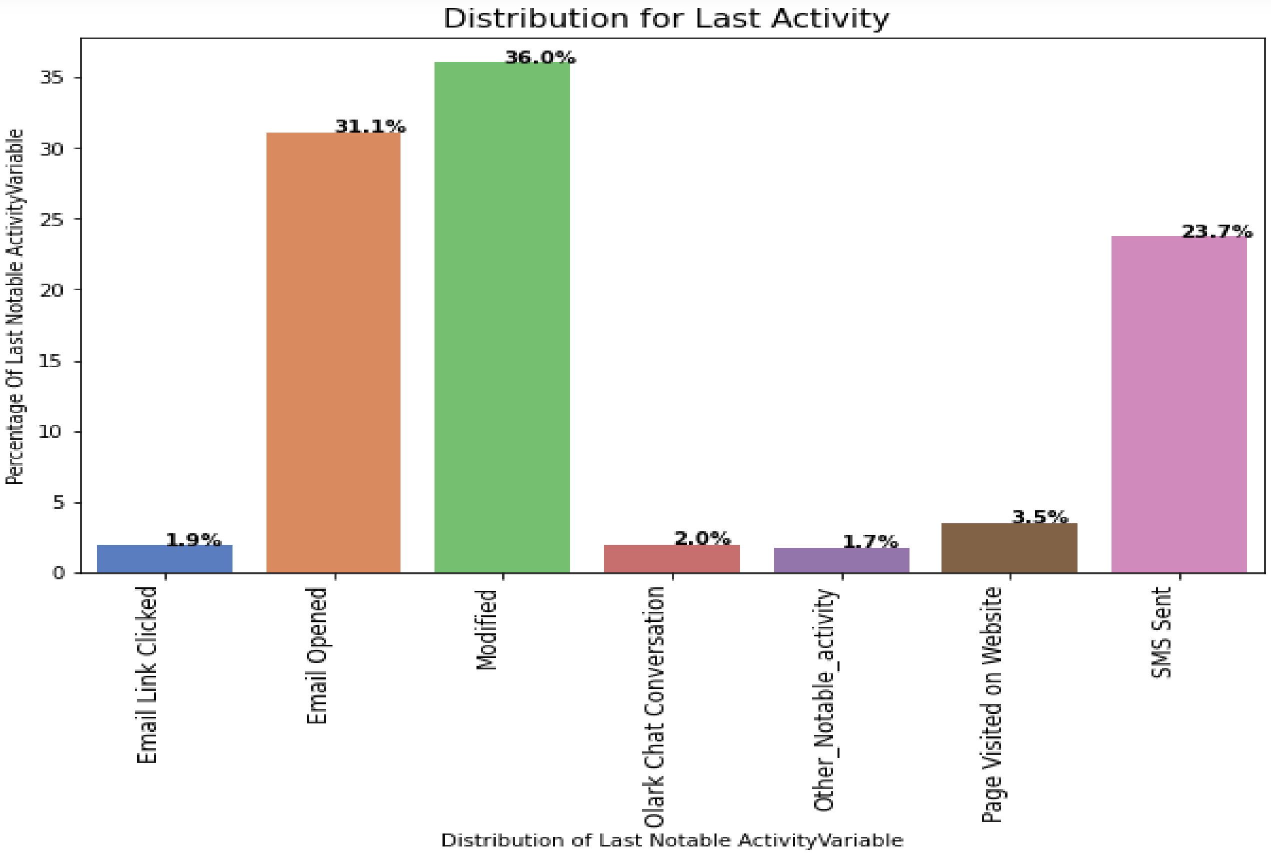
EDA INSIGHTS



Insights:

1. SMS Sent has a huge conversion ratio. Even though it contributes to only 30% prospects, its conversion rate is the best. This shows that more push strategy should be leveraged, since it has higher conversion opportunity.
2. Olark Chat Conversation, while having about 10% of prospects, has very low conversion rate - as also confirmed from Lead Source insights.
3. Email Opened while has much lower contribution to dataset, but has relative higher conversion ratios of almost 60-65%. Sending emails should also be leveraged more instead of focusing on Olark Chat etc.

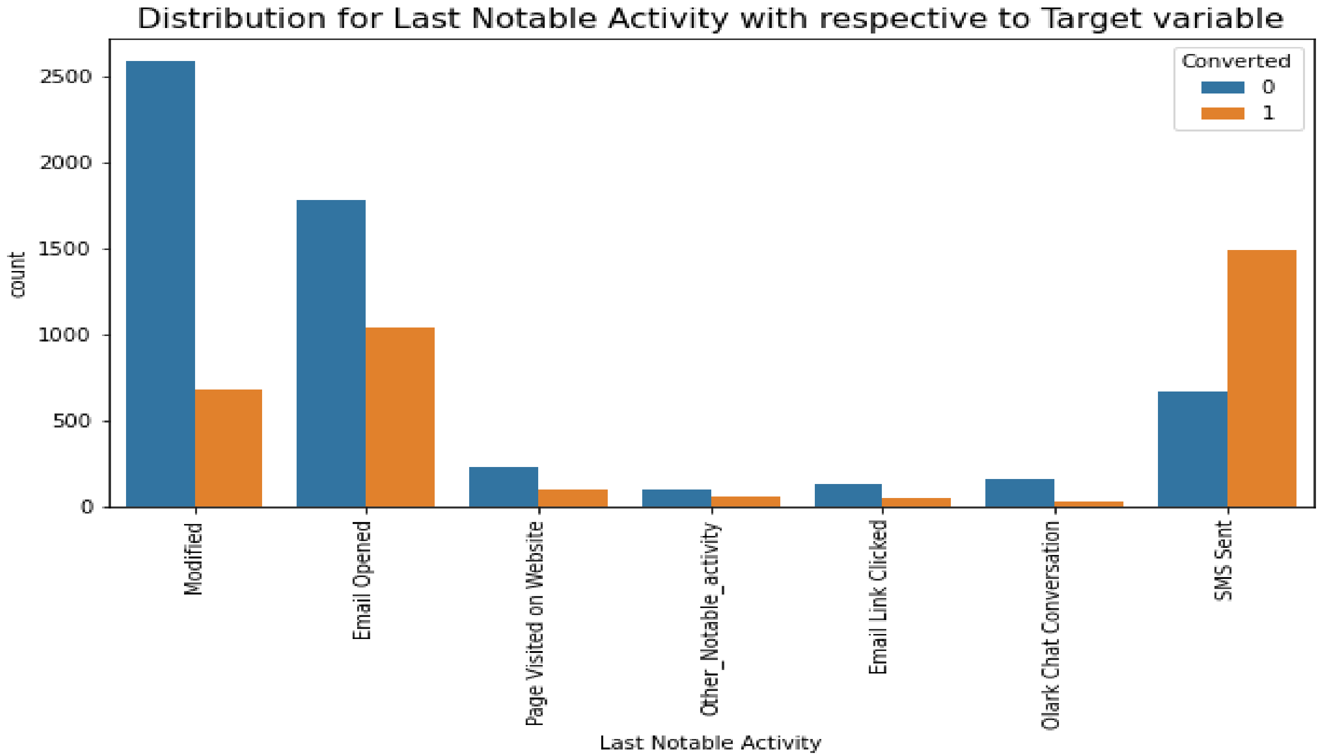
EDA INSIGHTS



Insights:

1. Email Opened, SMS Sent and Modified make up for 90% of the prospects.
2. However Modified, which makes up 36% of the total, is not very self explanatory as to the exact action that was performed. This will make it difficult to be leveraged as a focus area.
3. Other activities, though smaller sections of the pie, however - we will check distribution against Converted variable to see how they affect the conversion.

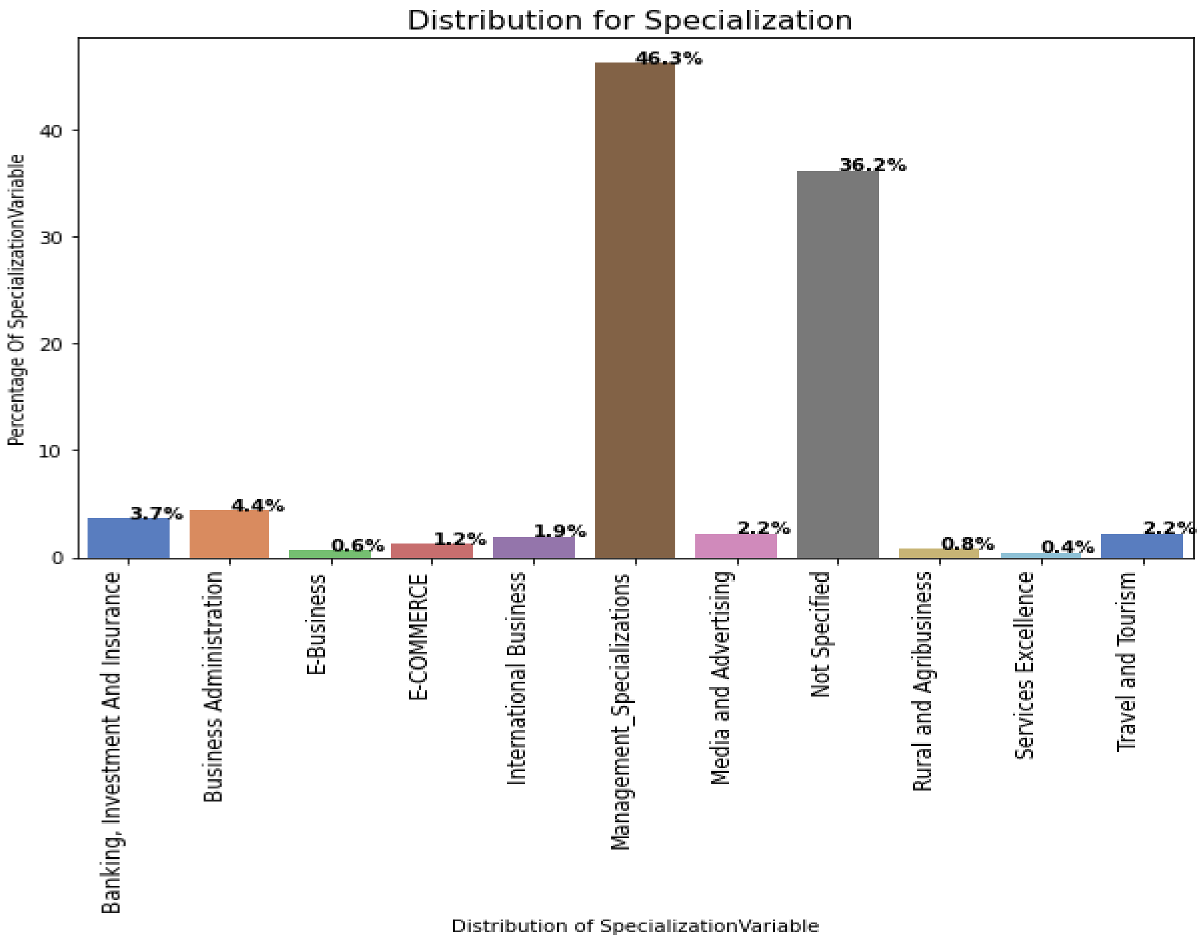
EDA INSIGHTS


Distribution for Last Notable Activity with respective to Target variable

Insights:

1. As also evidenced above, SMS sent continues to be a strong approach to converting prospects into customers and should be heavily leveraged.
2. Email Opened is also a good contender and should be given more focus.
3. Modified is low on conversion, but also not very clear on what gets included here. This should be expanded to include further detail and exact activities - which might then influence the conversion depending on specific activity.
4. Page Visited from Website, Email Link clicked are more pull operations that perform decent conversion but are more user dependent and less on control of company.

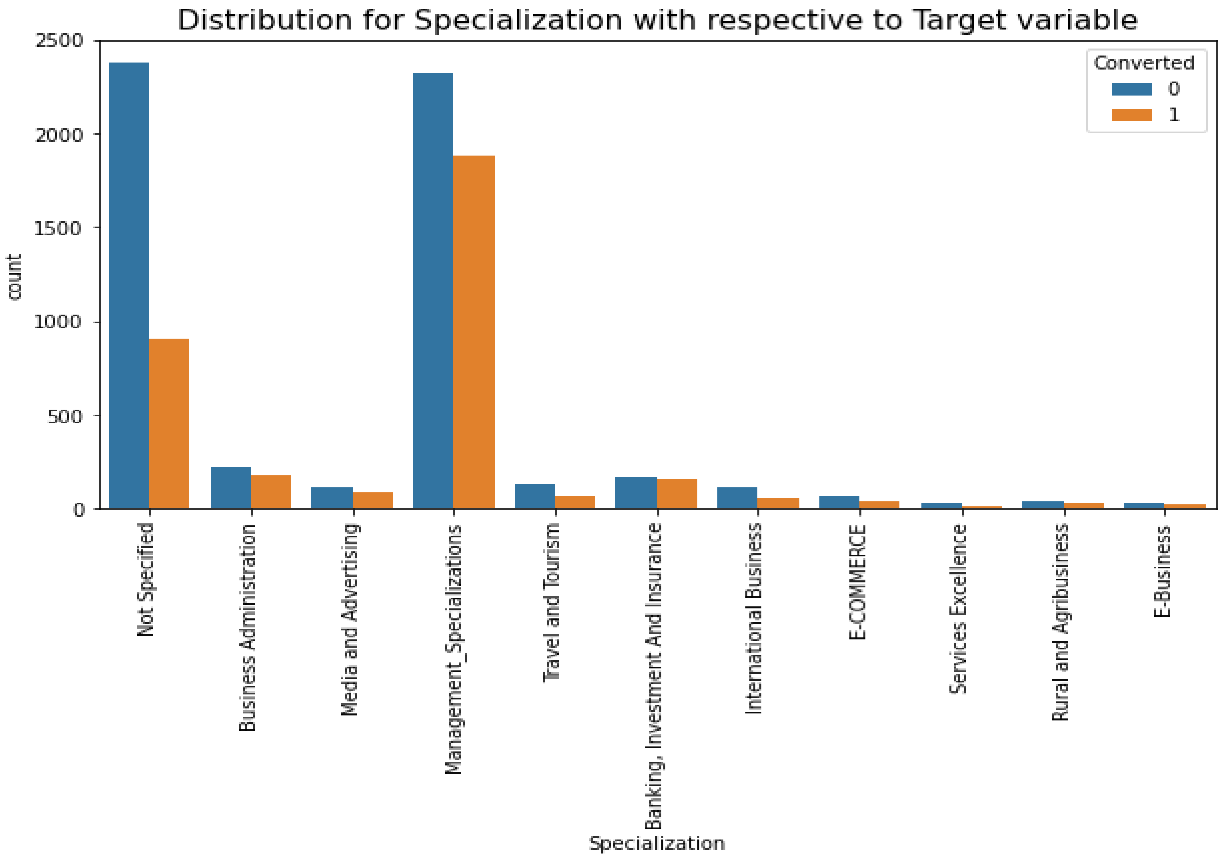
EDA INSIGHTS



Insights:

1. 'Not specified' and management specializations unfortunately, together, make up for the bulk of the dataset.
2. Conversion rate is also pretty high for Management specializations - about 75-80%. This is a huge opportunity.

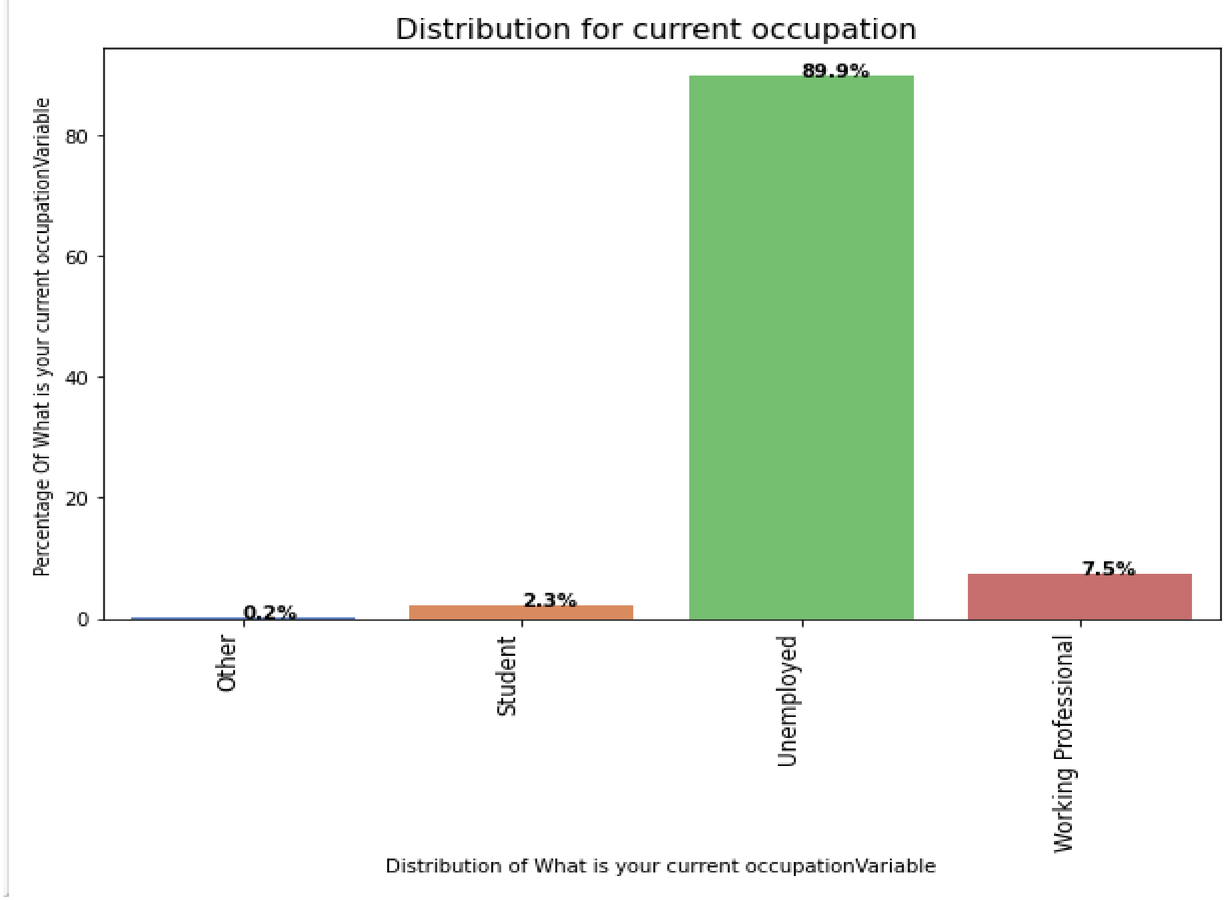
EDA INSIGHTS



Insights:

1. Management specialization individuals have a very strong ~ 75-80% conversion ratio. This should be capitalized on.
2. While individuals from 'Business Administration', 'Banking, Investment and Insurance', and 'Media and Advertising' are less represented in the dataset, their conversion ratio seems to be high. These individuals should be aggressively followed up on as prospects.

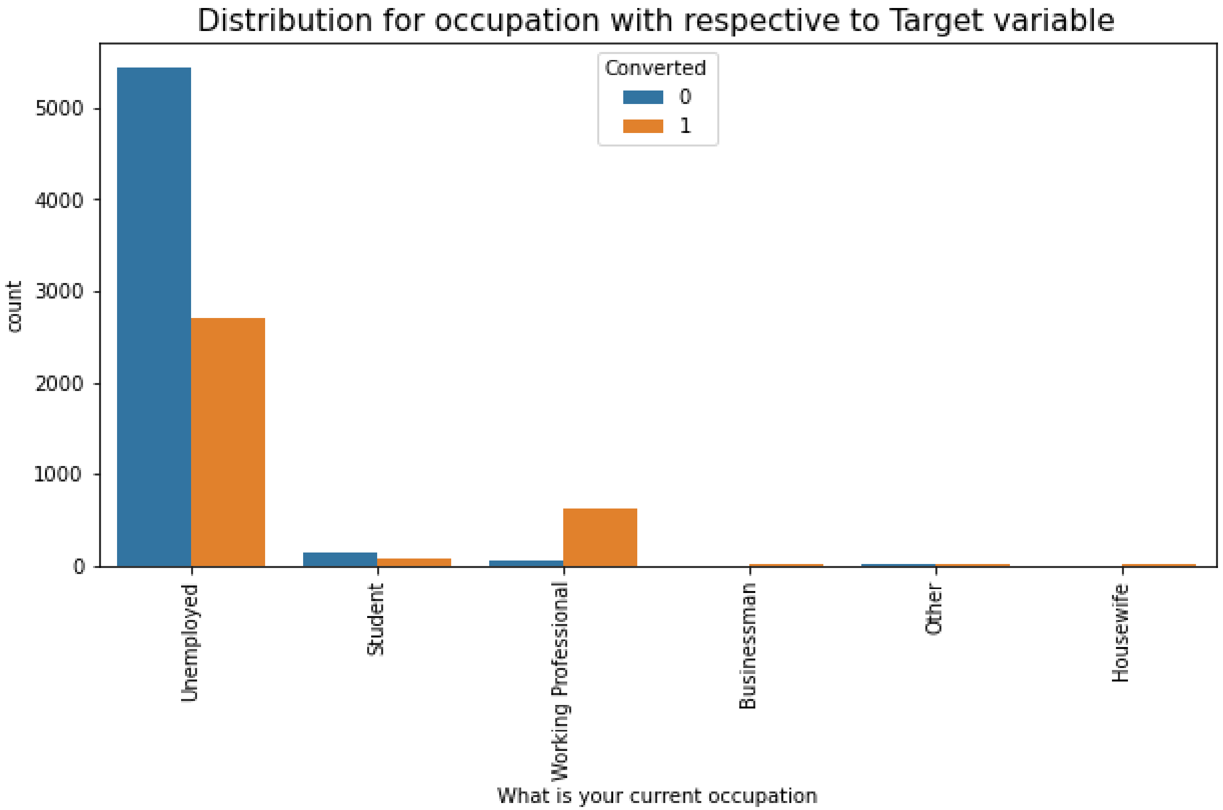
EDA INSIGHTS



Insights:

1. The bulk of the individuals that X Education is being marketed to are mostly unemployed individuals.
2. However, working professionals also figure, though quite less in number.
3. Students and other categories are negligible.

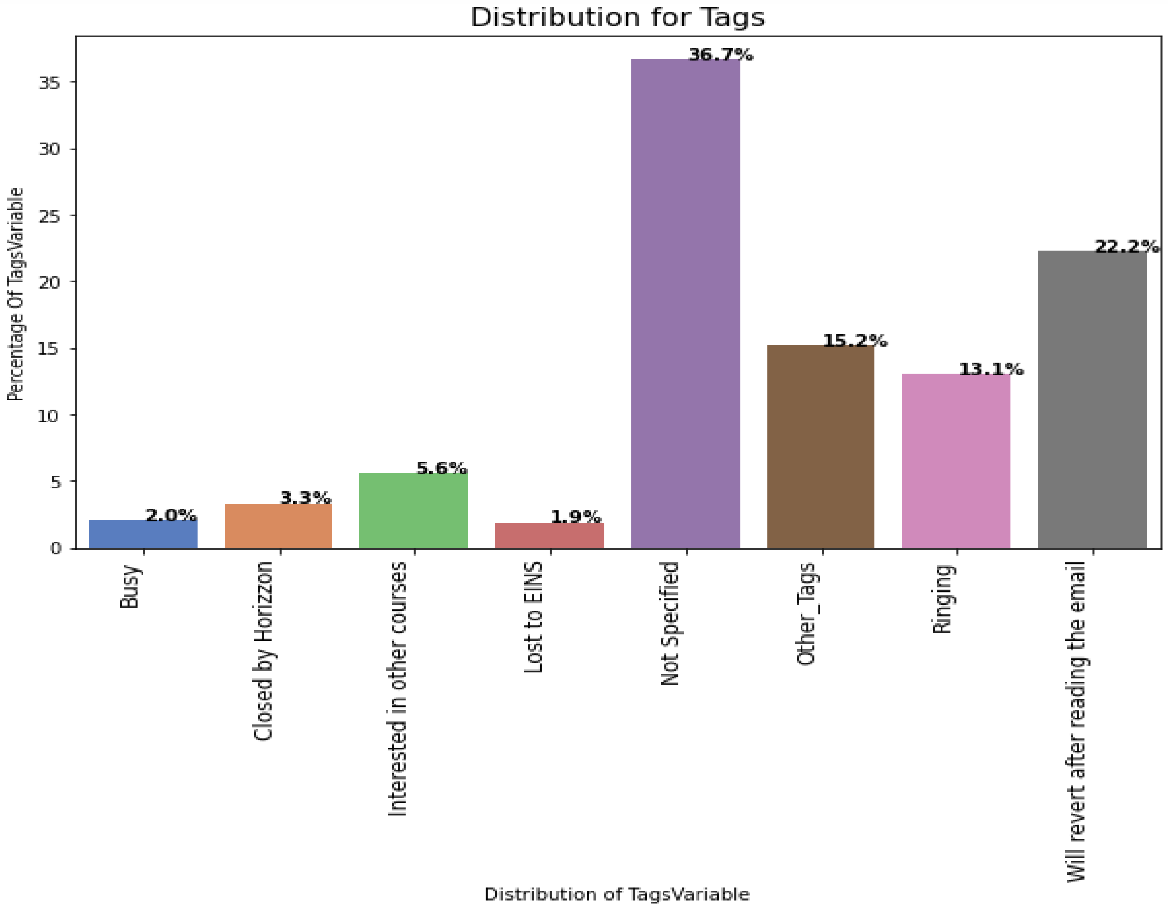
EDA INSIGHTS



Insights:

1. The highest conversion ratio here is for working professionals. Even though they're less in number within the dataset. This shows that marketing needs to attract more working professionals to increase their revenue through conversion.
2. The chances of conversion within unemployed seem to only be about 50%
3. More focus should be spent on working individuals as number of conversions is easily 4-5 times that of not converted.

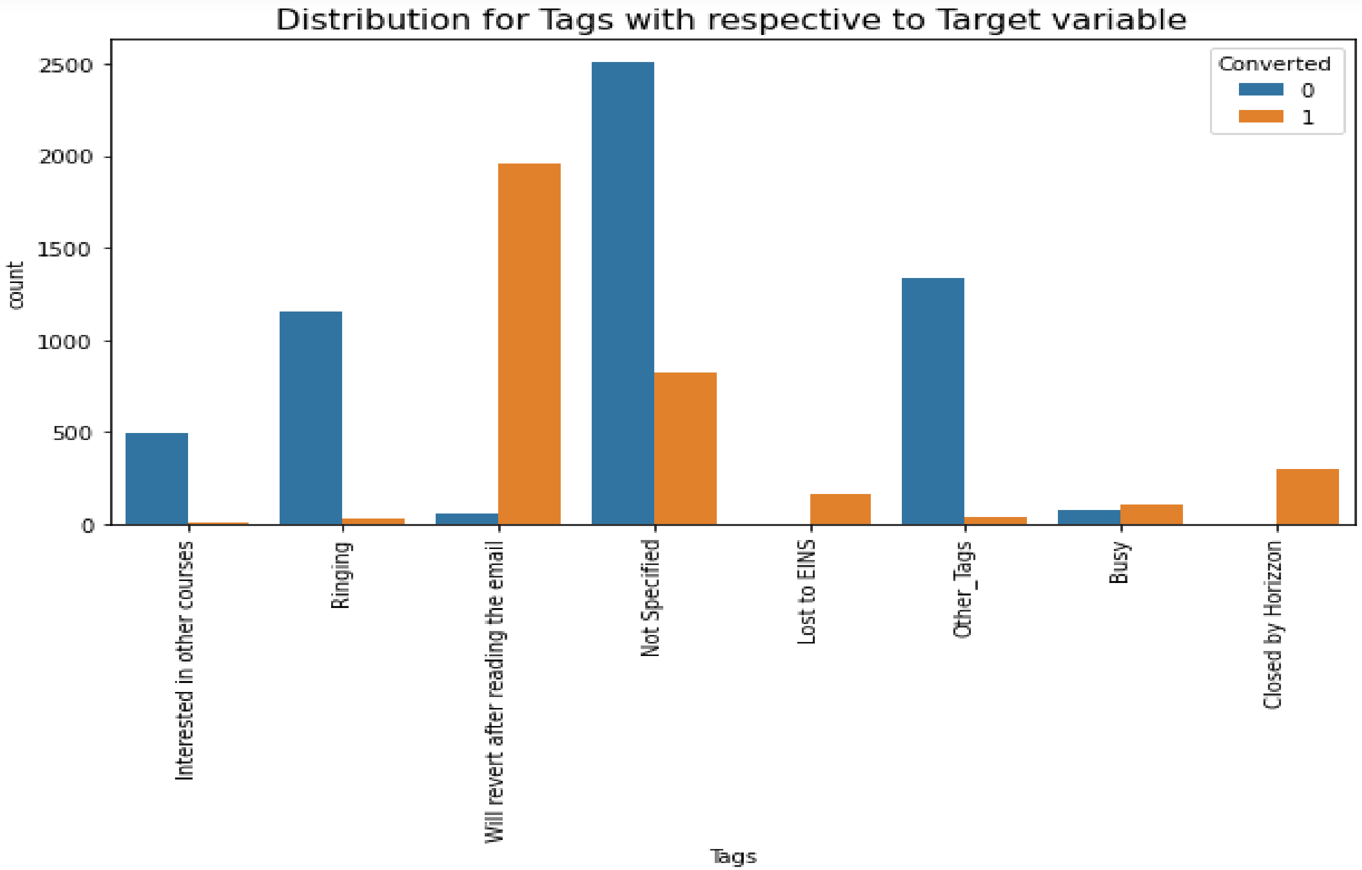
EDA INSIGHTS



Insights:

1. 'Not Specified' unfortunately forms a major portion of the dataset. It is difficult to focus on a specific strategy if the tag itself is not specified.
2. 'Will revert after reading the email' is a close contender with 22% of the share.
3. Several other smaller categories have been clubbed as Other - this forms about 15% of the data.

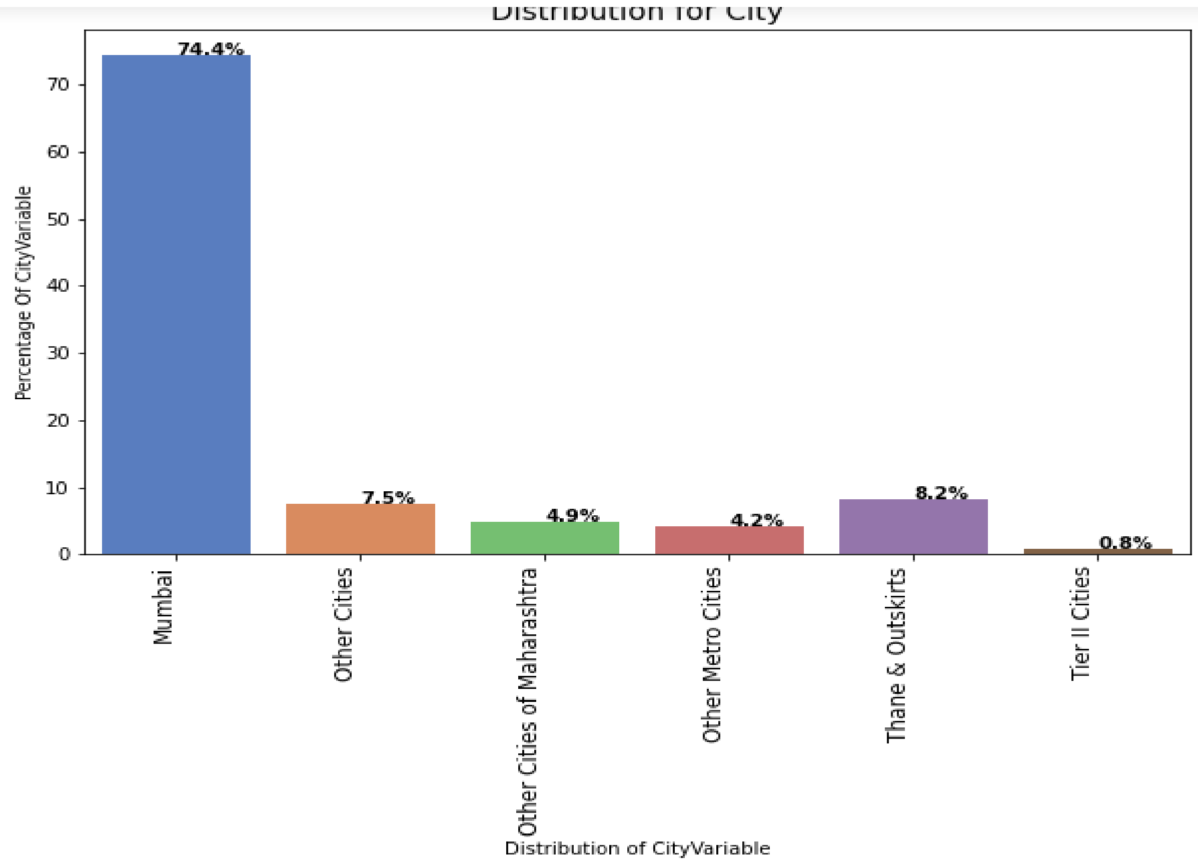
EDA INSIGHTS



Insights:

1. Not specified though forming a major portion of the dataset, falls low in conversion - only about 30-40% converting.
2. In contrast, however, email recipients who mark as 'Will revert after reading the email' have converted far more.
3. More focus and marketing efforts to be spent on email activities and following up through email - even call ringing does not convert as effectively.

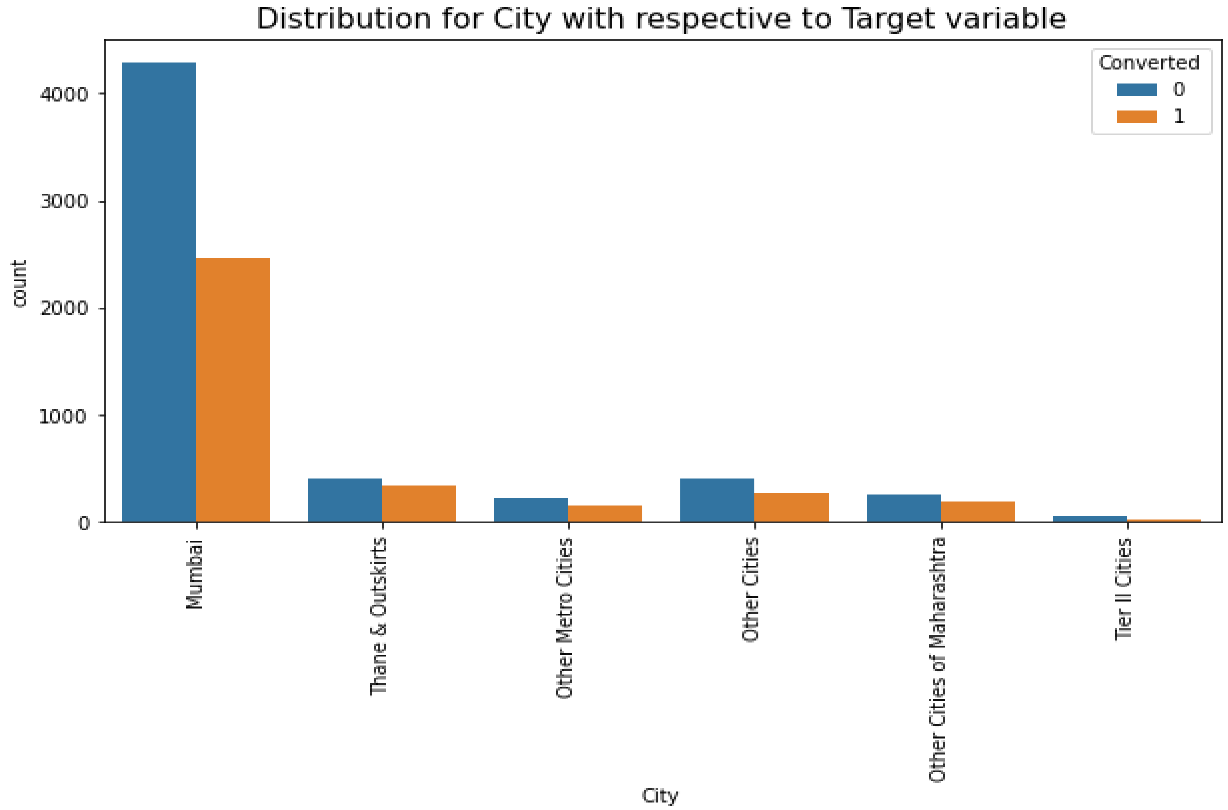
EDA INSIGHTS

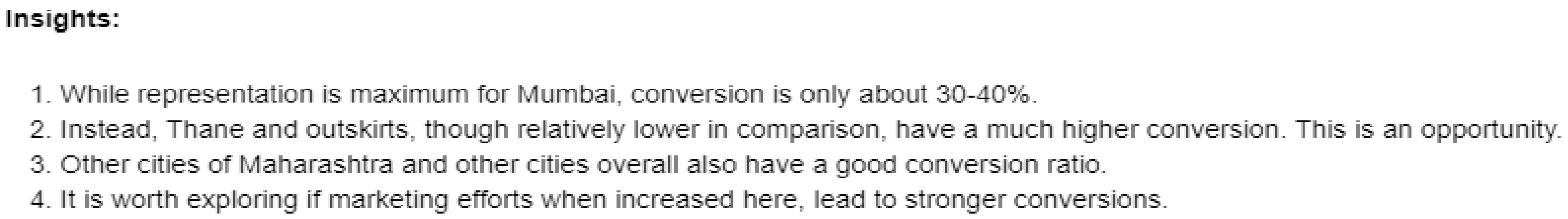


Insights:

1. Maximum data is available for individuals from Mumbai, Thane and outskirts
2. Tier II cities and other Metro cities, and other cities are much lesser in representation.
3. Other cities of Maharashtra is also a considerable number following other cities.
4. This indicates that most of the data being focused upon for the training model pertains to mostly Maharashtra state of India, and specifically Mumbai with neighbouring areas.

EDA INSIGHTS

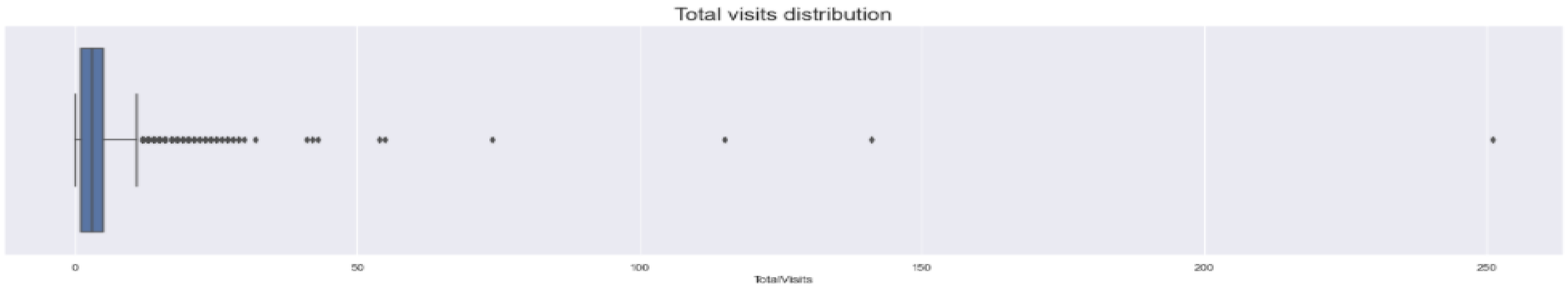


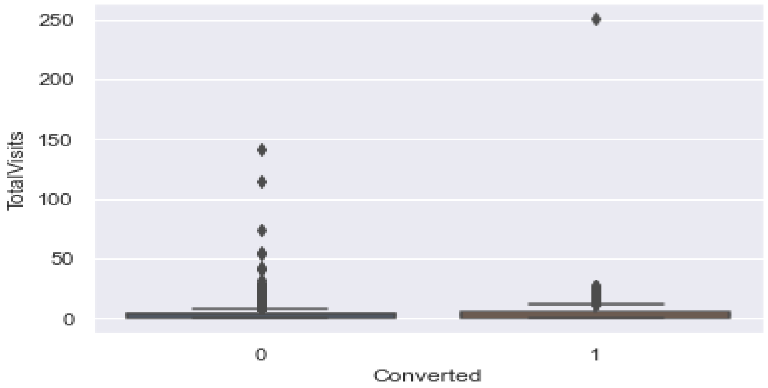


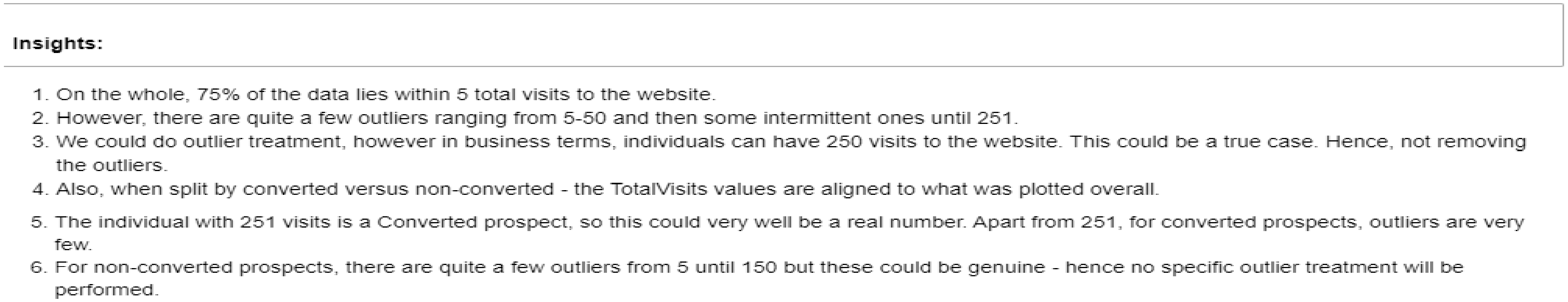
Insights:

1. While representation is maximum for Mumbai, conversion is only about 30-40%.
2. Instead, Thane and outskirts, though relatively lower in comparison, have a much higher conversion. This is an opportunity.
3. Other cities of Maharashtra and other cities overall also have a good conversion ratio.
4. It is worth exploring if marketing efforts when increased here, lead to stronger conversions.

EDA INSIGHTS



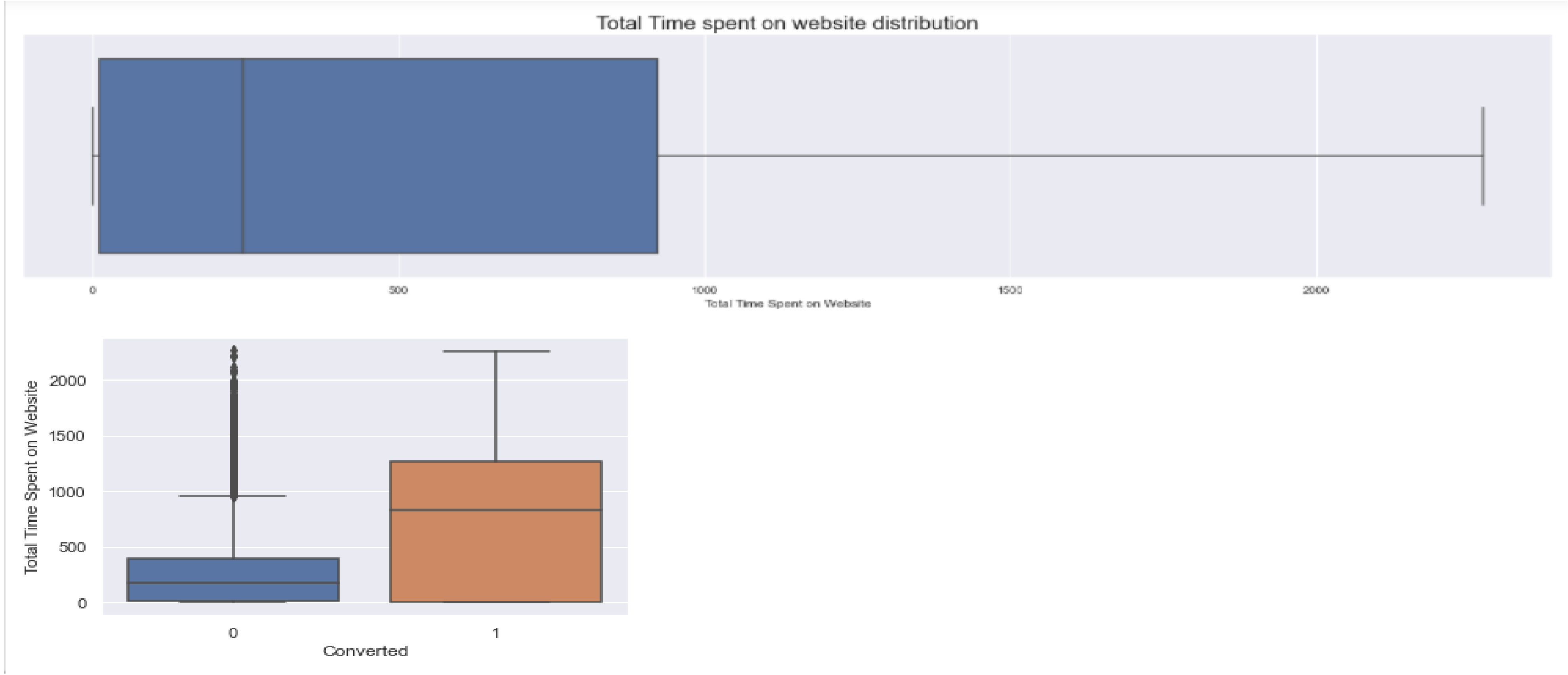




Insights:

1. On the whole, 75% of the data lies within 5 total visits to the website.
2. However, there are quite a few outliers ranging from 5-50 and then some intermittent ones until 251.
3. We could do outlier treatment, however in business terms, individuals can have 250 visits to the website. This could be a true case. Hence, not removing the outliers.
4. Also, when split by converted versus non-converted - the TotalVisits values are aligned to what was plotted overall.
5. The individual with 251 visits is a Converted prospect, so this could very well be a real number. Apart from 251, for converted prospects, outliers are very few.
6. For non-converted prospects, there are quite a few outliers from 5 until 150 but these could be genuine - hence no specific outlier treatment will be performed.

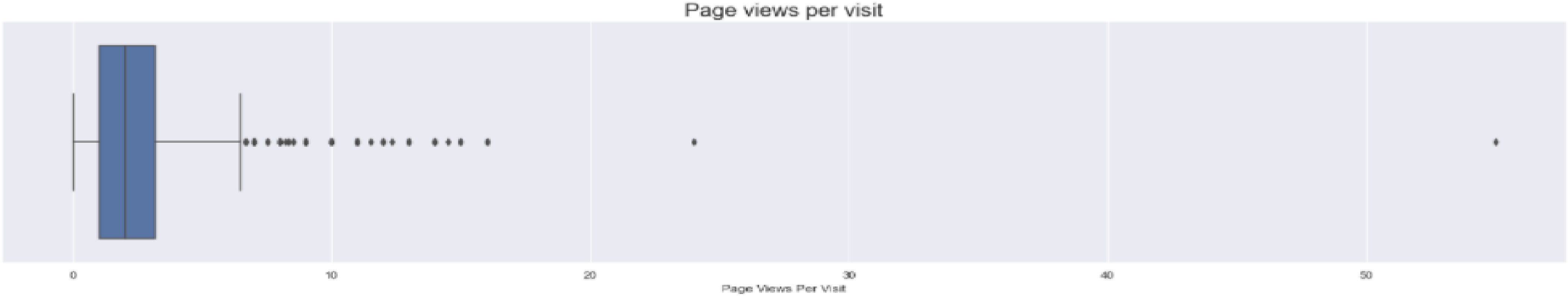
EDA INSIGHTS

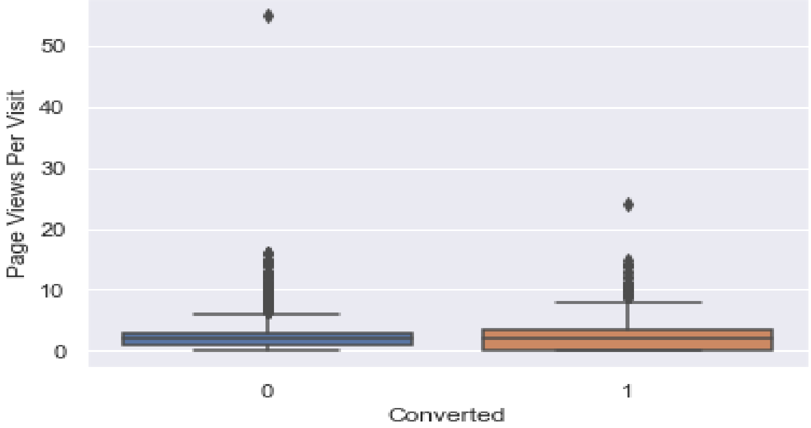


Insights:

1. The total time spent by most customers on the website is about 922 (assuming this is seconds - the data dictionary does not specify. For the purposes of documentation, we shall indicate this as 922 m where m stands for unit of measure.
2. The median is around 246 m; whereas there's still certain individuals between 922 m and 2272 m.
3. When split by converted and non-converted prospects, the median and 75 percentile is much higher for those who converted indicated that those who spent more time on the site have a higher probability of conversion.
4. In contrast, those who spent less time on the site have not converted.

EDA INSIGHTS



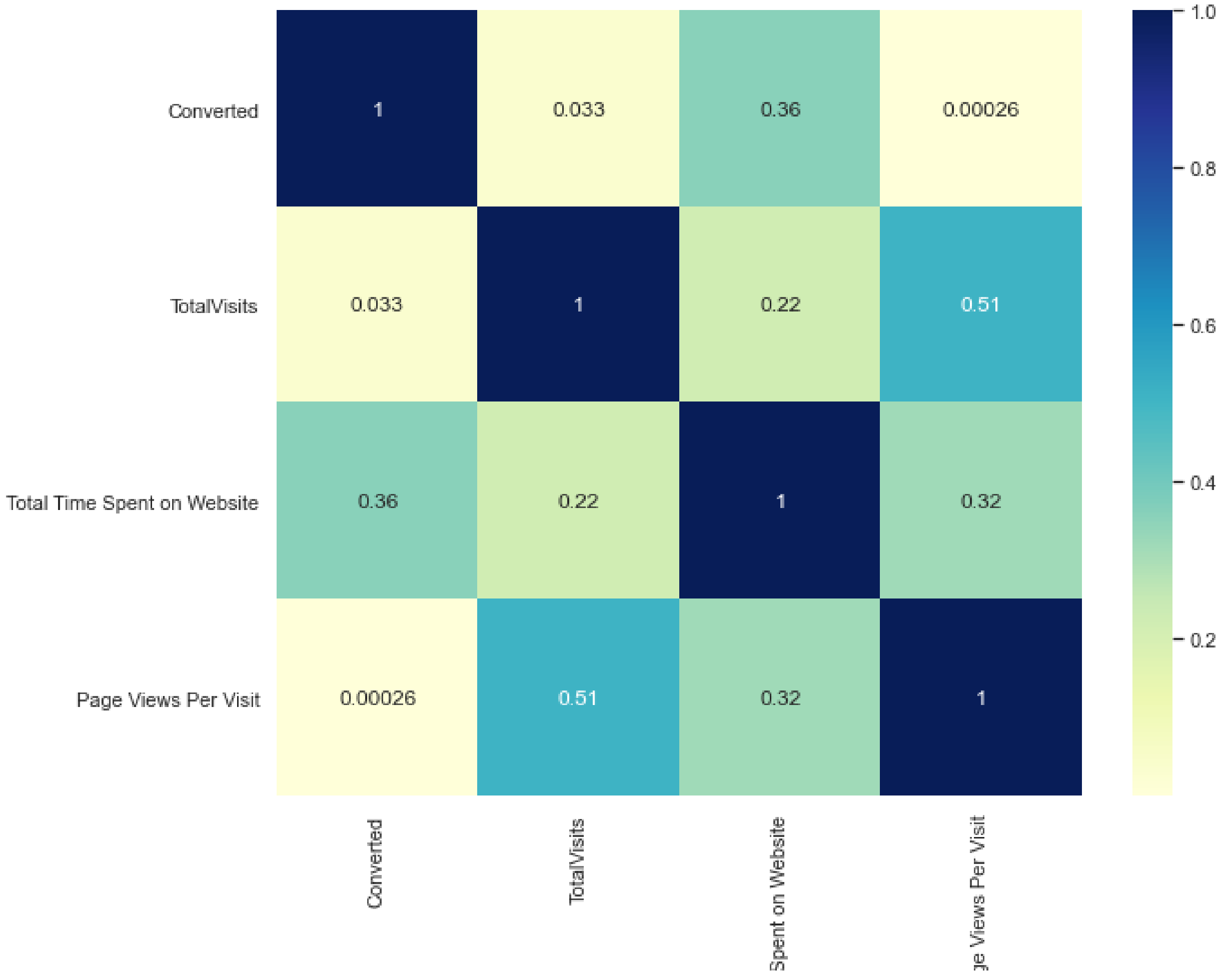


Insights:

1. Typically it is observed that, individuals have about 0-3 views per visit.
2. However, there are quite a few outliers taking the max value to 55.
3. Since this could also be a genuine case, hence outlier treatment will not be done - no eliminations will be performed.
4. There is no major significant difference in the page per views between those of converted versus non converted, so this may not necessarily be a factor in conversion.

CORELATIONS

- Numerical to numerical corelations are identified using heatmap.

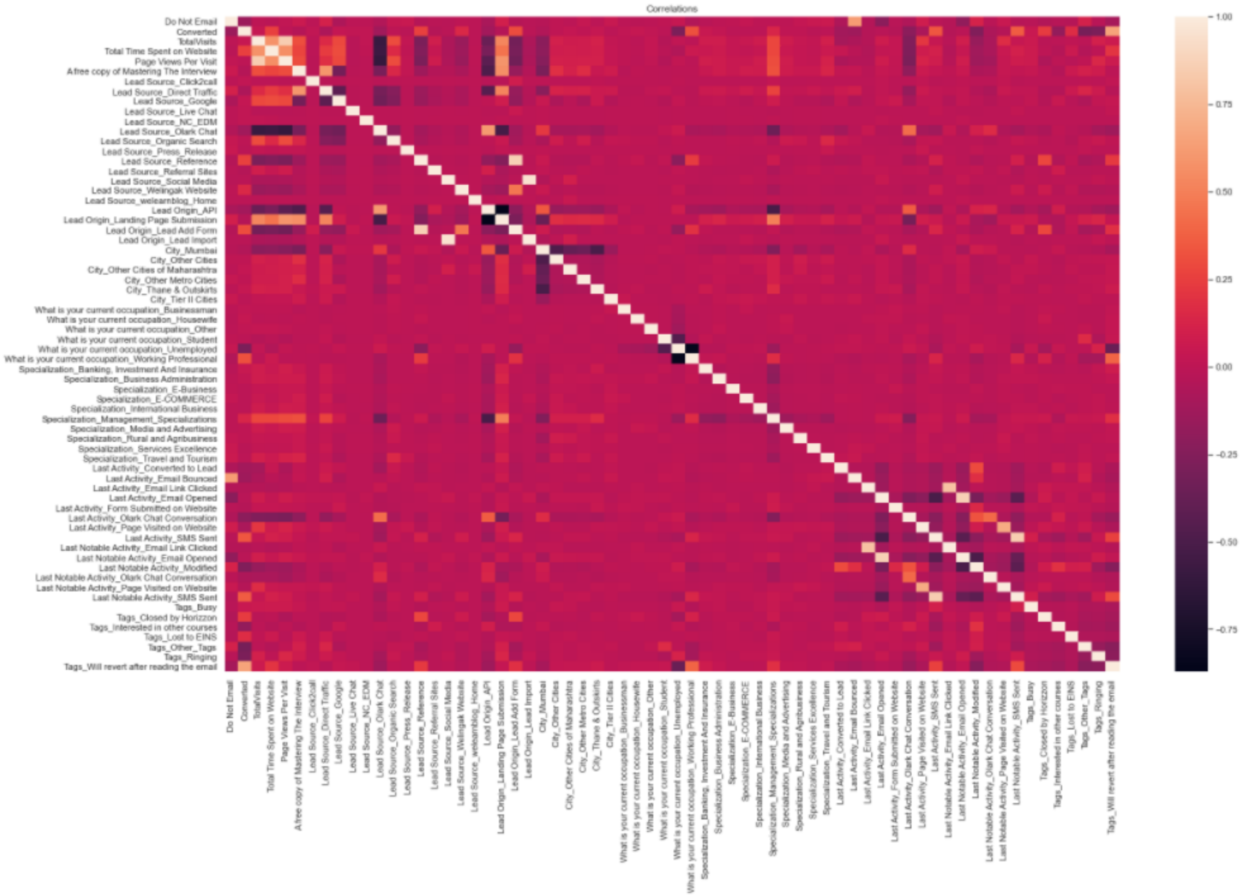


DUMMY VARIABLE CREATION

- Categorical variables having multiple values are split into dummy variables – one hot encoding process.
- This is to identify which of the values have high correlation and are highly significant to the target variable Converted.
- Upon dummy variable creation, certain dummy variables which do not have business sense are dropped or clubbed as Others.
- Once all dummy variables are created, and treated according to business sense or clubbed as Others – we then proceed with the train and test data creation.

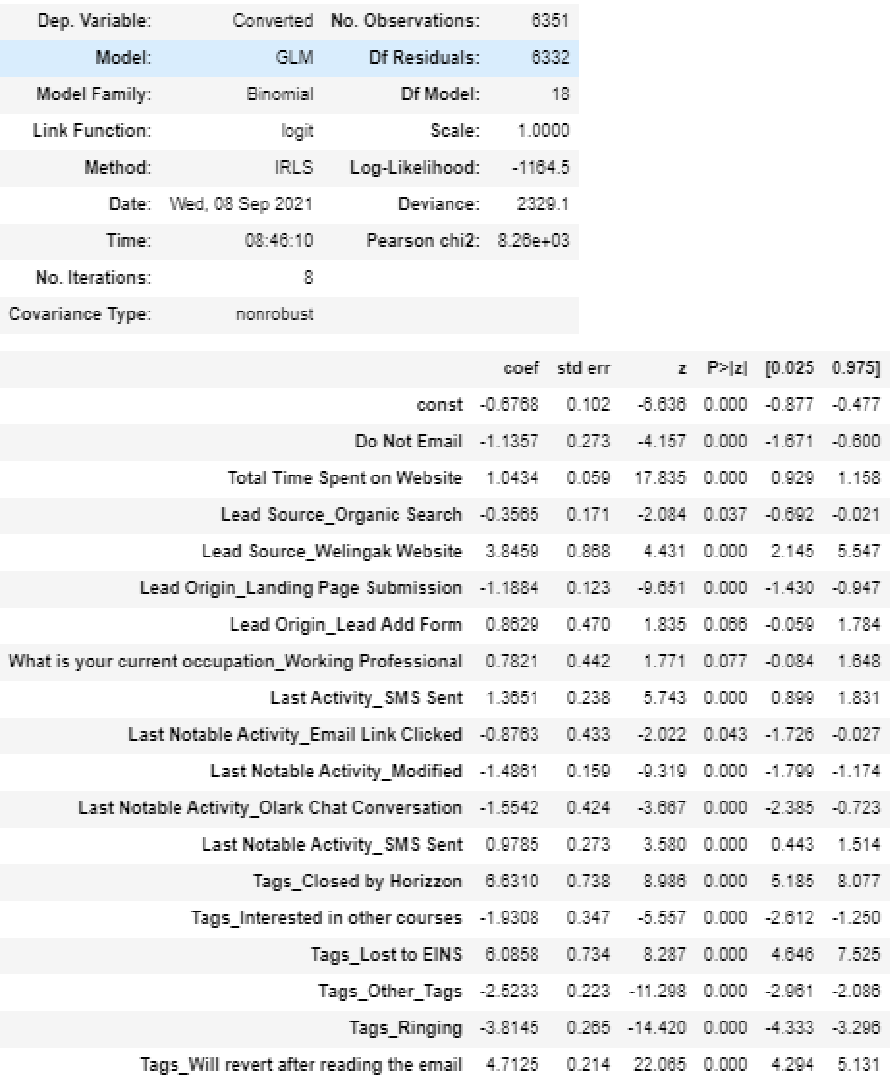
MODEL CREATION

- The dataset is now split into train and test data sets (both X and Y) on a random 70-30 split.
- X datasets consist of the index value Lead Number and all variables except the target variable Converted.
- Y dataset consists of the index value Lead Number and only the target variable Converted.
- Post this, we now check the correlation between all variables as well as all variables with the target variable Converted.
- Post this, we also drop the variable with max collinearity.



MODEL CREATION AND ASSESSMENT

- The first model on the training data is then run using GLM and the summary statistics are obtained.
- Based on this feature selection is running using 19 variables as an initial starting point.
- Once RFE is done, the next model is done only with those variable which have RFE support as true. The ranking also helps establish significance.
- The model version 2 is then generated.

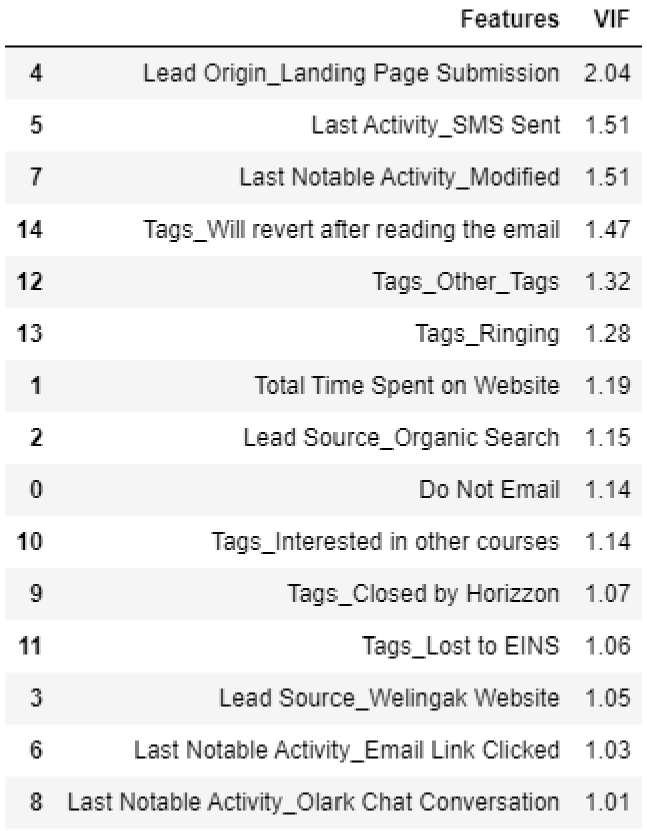


Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6332
Model Family:	Binomial	Df Model:	18
Link Function:	logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1164.5
Date:	Wed, 08 Sep 2021	Deviance:	2329.1
Time:	08:46:10	Pearson chi2:	8.26e+03
No. Iterations:	8		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.6768	0.102	-6.636	0.000	-0.877	-0.477
Do Not Email	-1.1357	0.273	-4.157	0.000	-1.671	-0.600
Total Time Spent on Website	1.0434	0.059	17.835	0.000	0.929	1.158
Lead Source_Organic Search	-0.3565	0.171	-2.084	0.037	-0.692	-0.021
Lead Source_Welingak Website	3.8459	0.868	4.431	0.000	2.145	5.547
Lead Origin_Landing Page Submission	-1.1884	0.123	-9.651	0.000	-1.430	-0.947
Lead Origin_Lead Add Form	0.8629	0.470	1.835	0.066	-0.059	1.784
What is your current occupation_Working Professional	0.7821	0.442	1.771	0.077	-0.084	1.648
Last Activity_SMS Sent	1.3651	0.238	5.743	0.000	0.899	1.831
Last Notable Activity_Email Link Clicked	-0.8763	0.433	-2.022	0.043	-1.726	-0.027
Last Notable Activity_Modified	-1.4861	0.159	-9.319	0.000	-1.799	-1.174
Last Notable Activity_Olark Chat Conversation	-1.5542	0.424	-3.667	0.000	-2.385	-0.723
Last Notable Activity_SMS Sent	0.9785	0.273	3.580	0.000	0.443	1.514
Tags_Closed by Horizzon	6.6310	0.738	8.986	0.000	5.185	8.077
Tags_Interested in other courses	-1.9308	0.347	-5.557	0.000	-2.612	-1.250
Tags_Lost to EINS	6.0858	0.734	8.287	0.000	4.646	7.525
Tags_Other_Tags	-2.5233	0.223	-11.298	0.000	-2.961	-2.086
Tags_Ringing	-3.8145	0.265	-14.420	0.000	-4.333	-3.296
Tags_Will revert after reading the email	4.7125	0.214	22.065	0.000	4.294	5.131

MODEL CREATION AND ASSESSMENT

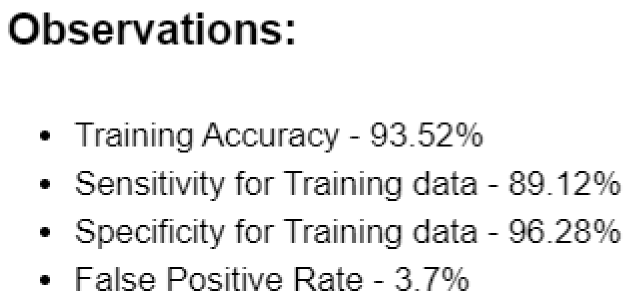
- Variable collinearity is checking using VIF. Variables having P > 0.05 or high VIF are dropped and the model is iteratively developed and summary statistics generated again.
- The iterative process is done until P is as close to 0.000 and VIF is < 5.
- Final model has 15 variables.



	Features	VIF
4	Lead Origin_Landing Page Submission	2.04
5	Last Activity_SMS Sent	1.51
7	Last Notable Activity_Modified	1.51
14	Tags_Will revert after reading the email	1.47
12	Tags_Other_Tags	1.32
13	Tags_Ringing	1.28
1	Total Time Spent on Website	1.19
2	Lead Source_Organic Search	1.15
0	Do Not Email	1.14
10	Tags_Interested in other courses	1.14
9	Tags_Closed by Horizzon	1.07
11	Tags_Lost to EINS	1.06
3	Lead Source_Welingak Website	1.05
6	Last Notable Activity_Email Link Clicked	1.03
8	Last Notable Activity_Olark Chat Conversation	1.01

MODEL EVALUATION

- Evaluation of the model is important to understand how well it fits the test data.
- Predicted variable is created based on the model.
- Once the predicted variable is available, it can be used against the converted variable to evaluate the model. Based on the predicted score, we're using a cut off of 0.5 initially to come up with 1-0 number for predicted. Scores > 0.5 are marked as 1, others as 0.
- The confusion matrix is then generated to derive metrics like accuracy, specificity, sensitivity etc.
- This gives us a measure of how well our model predicted the conversion.

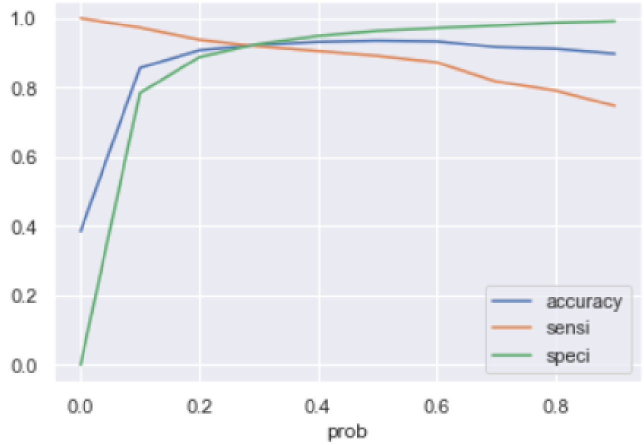


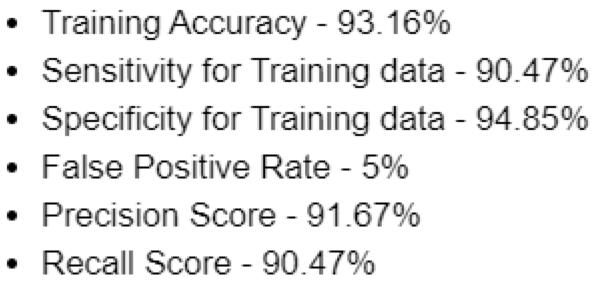
Observations:

- Training Accuracy - 93.52%
- Sensitivity for Training data - 89.12%
- Specificity for Training data - 96.28%
- False Positive Rate - 3.7%

RE-CALCULATING CUTOFF AND PLOTTING ROC

- Based on the confusion matrix based metrics, the specificity, sensitivity and accuracy are plotted to see where they meet.
- Optimal cut off is 0.3 from the graph.
- Based on this optimal cutoff, the predicted value is recalculated.
- While the ROC chart shows 0.3, however from the probability table we can see that at 0.3 cutoff, both accuracy and specificity have a dip. So, we instead go with 0.4 as a cutoff where there is no drop of any of the metrics.
- Based on the new cutoff and the new predicted values, once again the confusion matrix based metrics are derived including precision and recall.

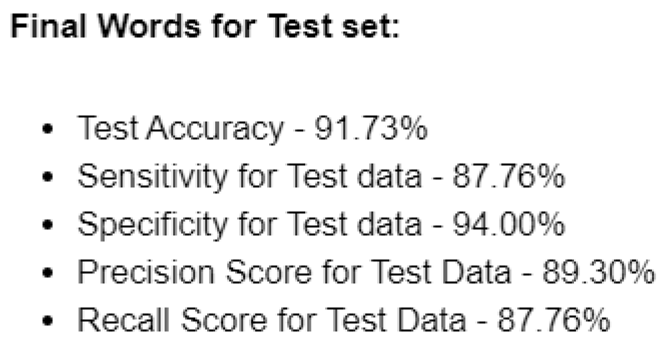




- Training Accuracy - 93.16%
- Sensitivity for Training data - 90.47%
- Specificity for Training data - 94.85%
- False Positive Rate - 5%
- Precision Score - 91.67%
- Recall Score - 90.47%

PREDICTIONS ON THE TEST SET

- The model is now ready and evaluated to be used on the test data.
- Same set of functions are performed on the test set, so that it only contains the required variables as are a part of the training set.
- Predicted variable is calculated directly on the test set using the final model.
- The confusion matrix metrics are derived again to check effectiveness on test data.



Final Words for Test set:

- Test Accuracy - 91.73%
- Sensitivity for Test data - 87.76%
- Specificity for Test data - 94.00%
- Precision Score for Test Data - 89.30%
- Recall Score for Test Data - 87.76%

THE MODEL HAS A STRONG ACCURACY, SENSITIVITY AND SPECIFICITY BOTH ON TRAINING AND TEST DATA. HENCE, WE CAN CONFIDENTLY SAY THIS MODEL PREDICTS CONVERSION WELL AND CAN BE USED TO DRIVE LEAD CONVERSIONS.

CONCLUSION

- All the confusion matrix related metrics of accuracy, sensitivity (recall), specificity, precision are showing strong numbers indicating the strength of the model.
- In business terms, this model also has an ability to adjust to changing requirements.
- The model is stable.
- There is not more than 2% difference in scores between the training data outcomes and test data outcomes.
- The key variables in the model contributing most towards the probability of a lead getting converted are

1. Total Time spent on Website
2. Tag closed by Horizzon
3. Tag Lost to EINS
4. Tag will revert after reading email
5. Lead Source Welingak Website
6. Last Activity SMS Sent