LEAD SCORING CASE STUDY

Submitted by:
Uma Sivakumar
Harsh Kushwaha
Guntur Harika

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.

BUSINESS GOAL

- X Education requires a model to be built for selecting the most promising leads.

- The business wants you to create a model in which you give each lead a lead score so that leads with higher lead scores have a better likelihood of converting, while leads with lower lead scores have a lesser chance of converting. The goal lead conversion rate was mentioned as being something in the neighborhood of 80% by the CEO in particular.

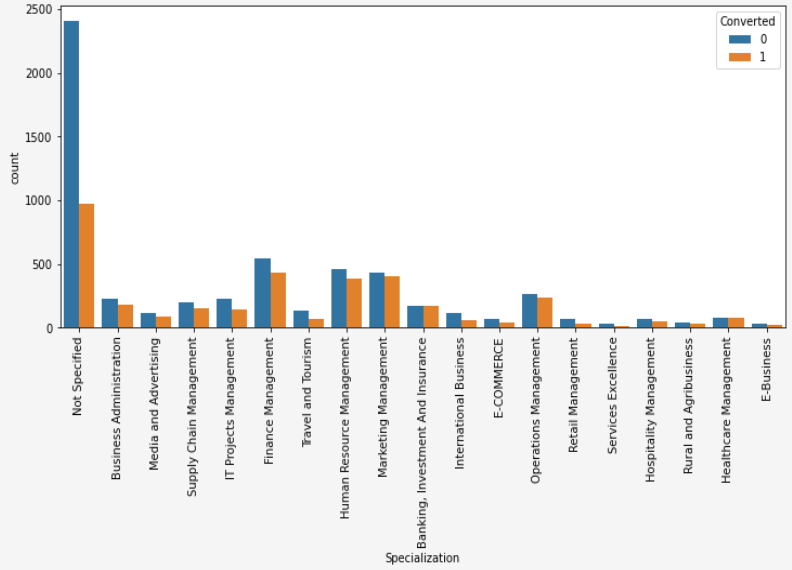
STRATERGY

- Import the data
- Reading and understanding data
- Data cleaning and preparation for the analysis
- Preparing the data for modelling
  - Dummy Variable creation
- Train-Test split
  - Rescaling the variables
- Model building
- Model evaluation
- Prediction on the test set

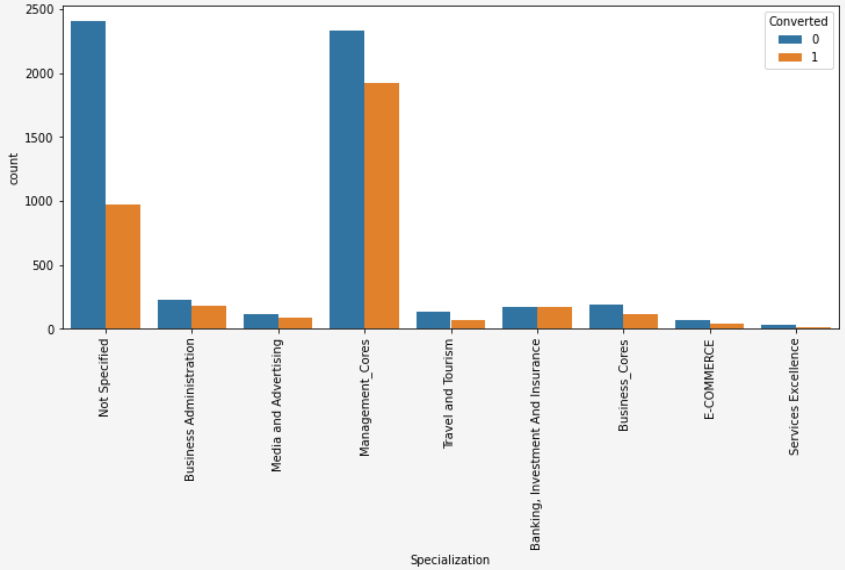
EXPLORATORY DATA ANALYSIS

Viewing the **Specialization** graph to observe the distribution of the categories before and after merging.

Before



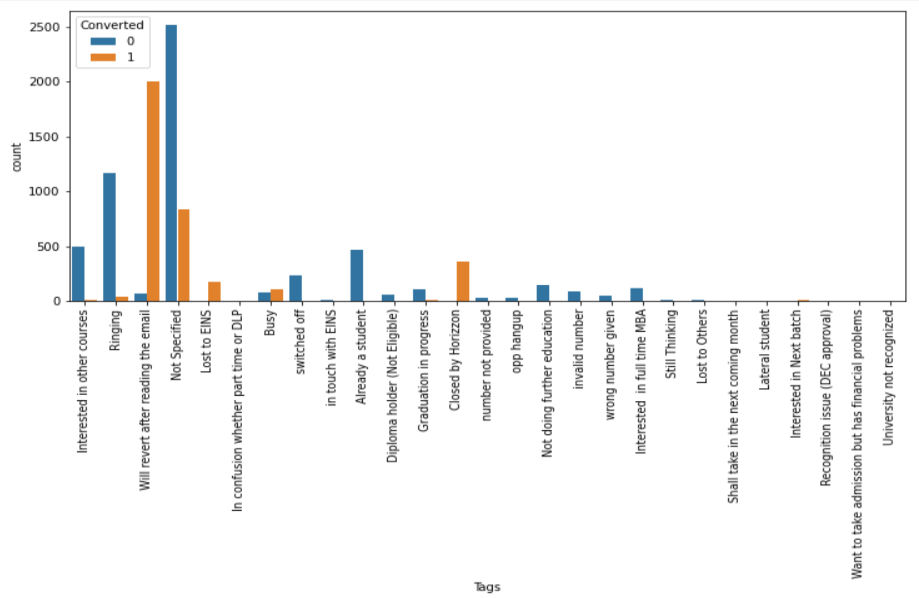
After



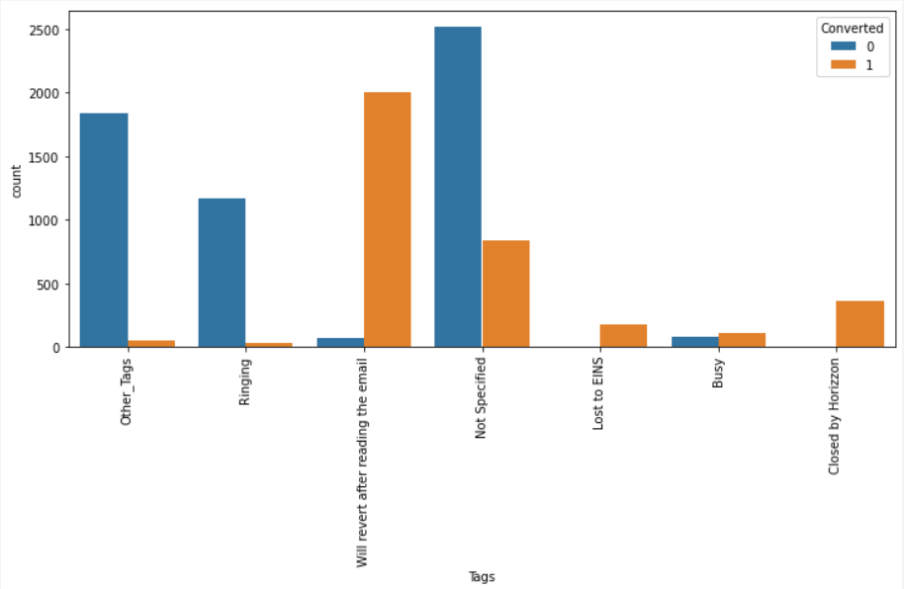
- There are numerous management courses available with a strong conversion rate; it would be advantageous to combine them into one entity for future study. Moreover, a few business courses can be merged.

Viewing the **Tags** graph to observe the distribution of the categories before and after merging.

Before



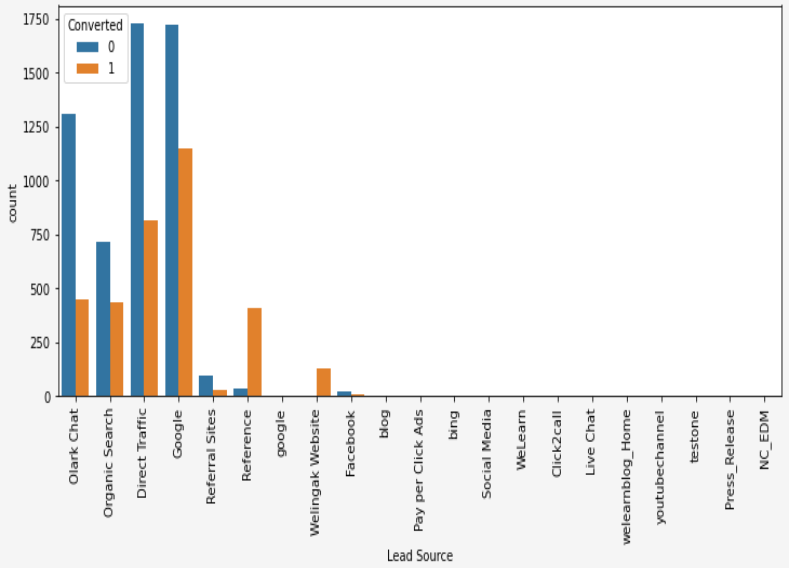
After



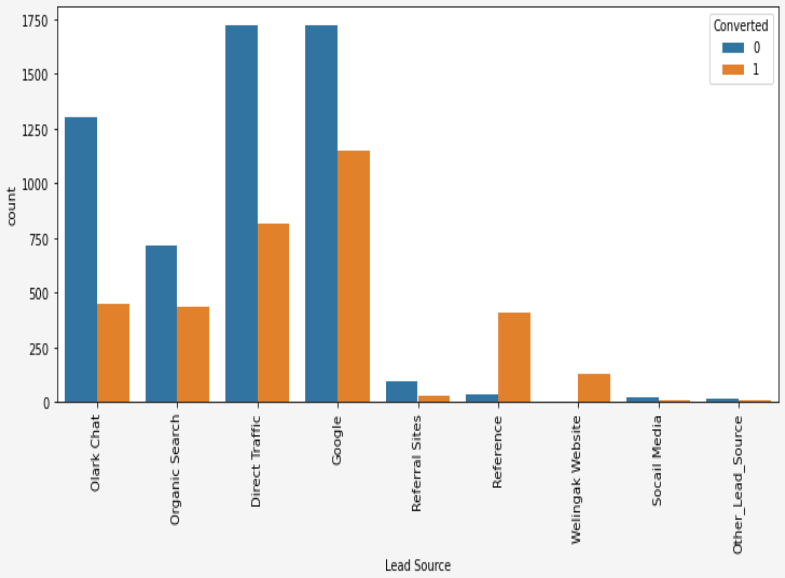
- There are numerous management courses available with a strong conversion rate; it would be advantageous to combine them into one entity for future study. Moreover, a few business courses can be merged.

Observing the **Lead Source** graph to compare the distribution of the categories before and after merging

Before



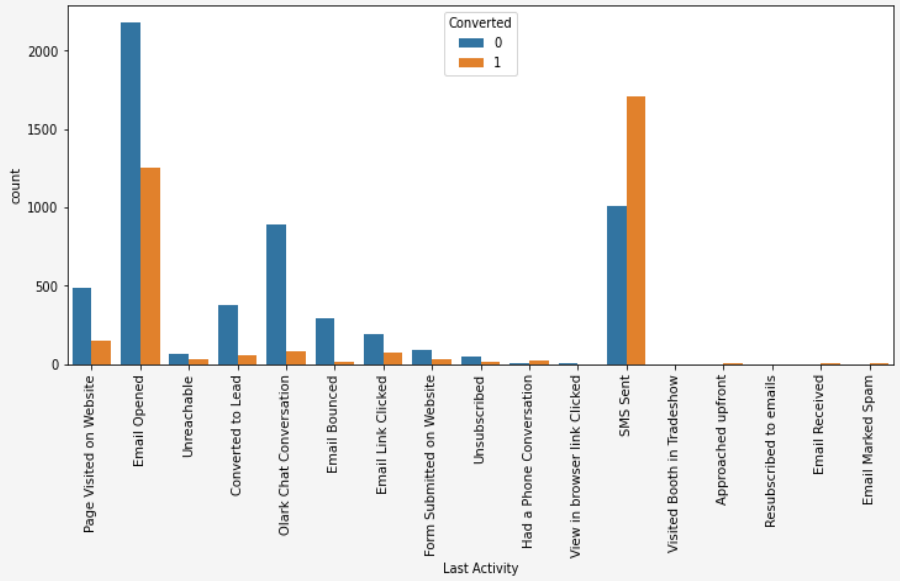
After



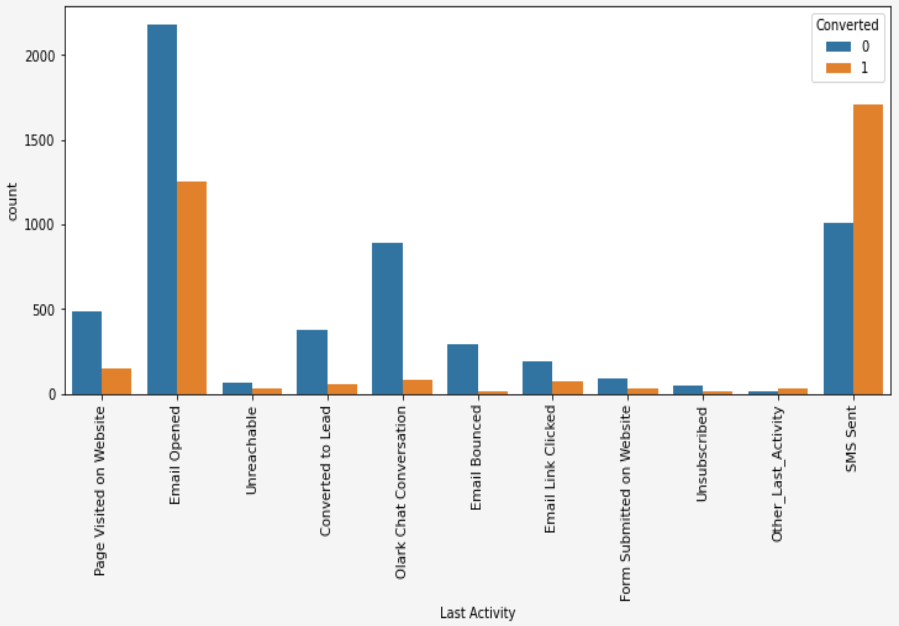
- There are many categories which add no value to the conversion/non-conversion of the leads. Combining them as one entity would help in further analysis while creating dummy variables.

Observing the **Last Activity** graph to compare the distribution of the categories before and after merging

Before



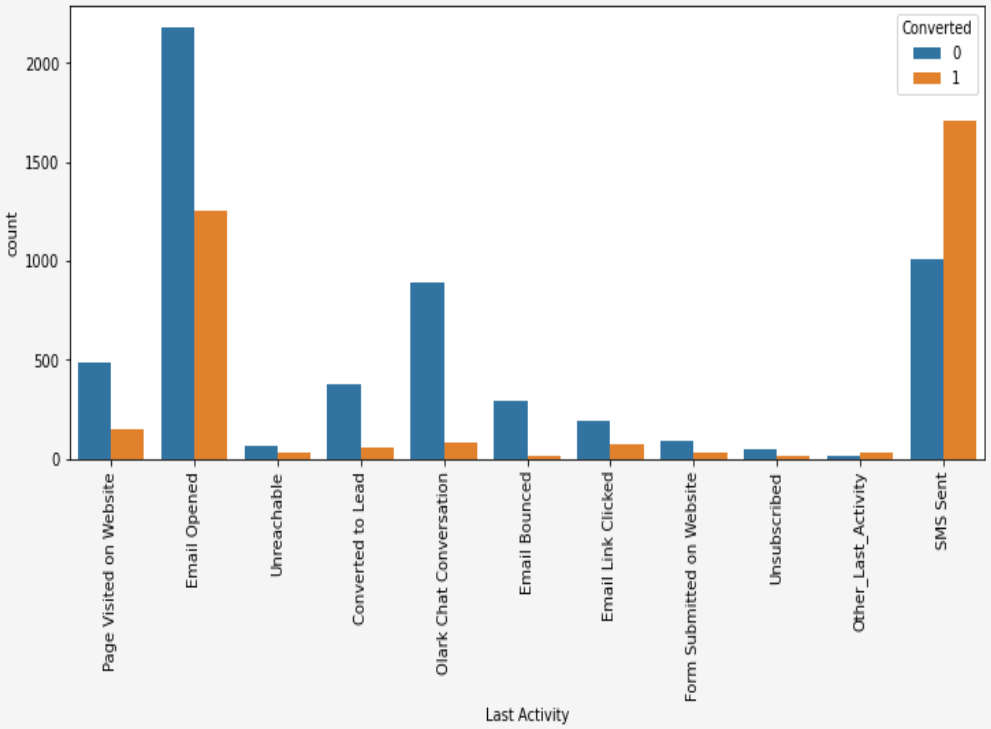
After



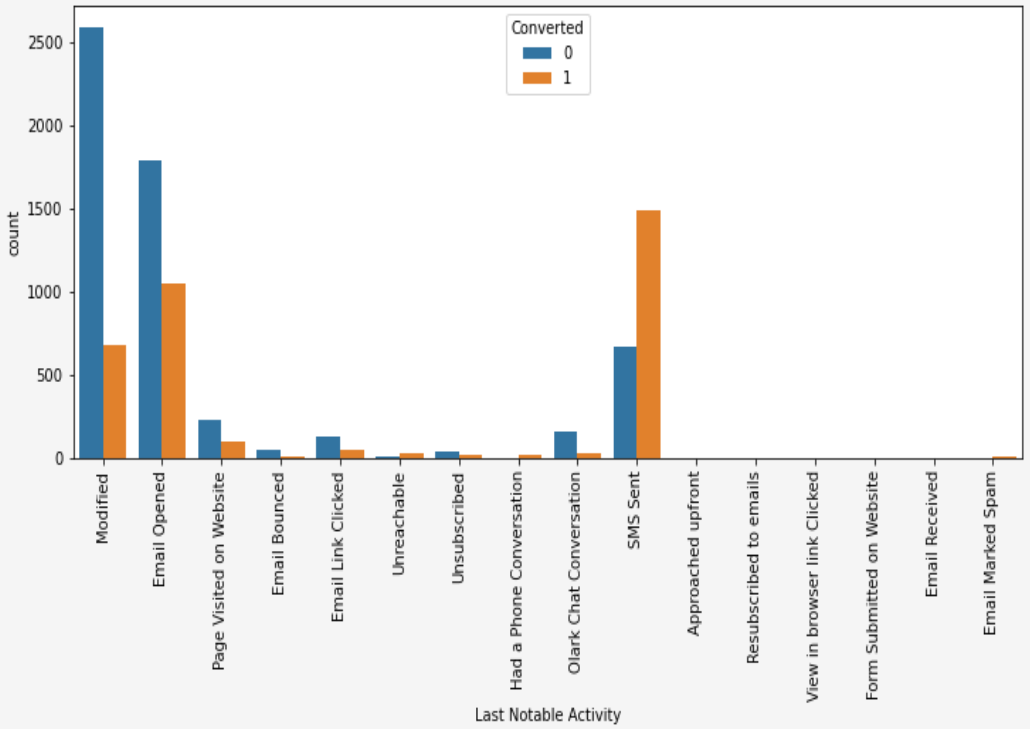
- Combining the many Last Activity categories will improve our analysis because many of them don't add anything to it.

Observing **Last Notable Activity** graph to compare the distribution of the categories before and after merging

Before



After

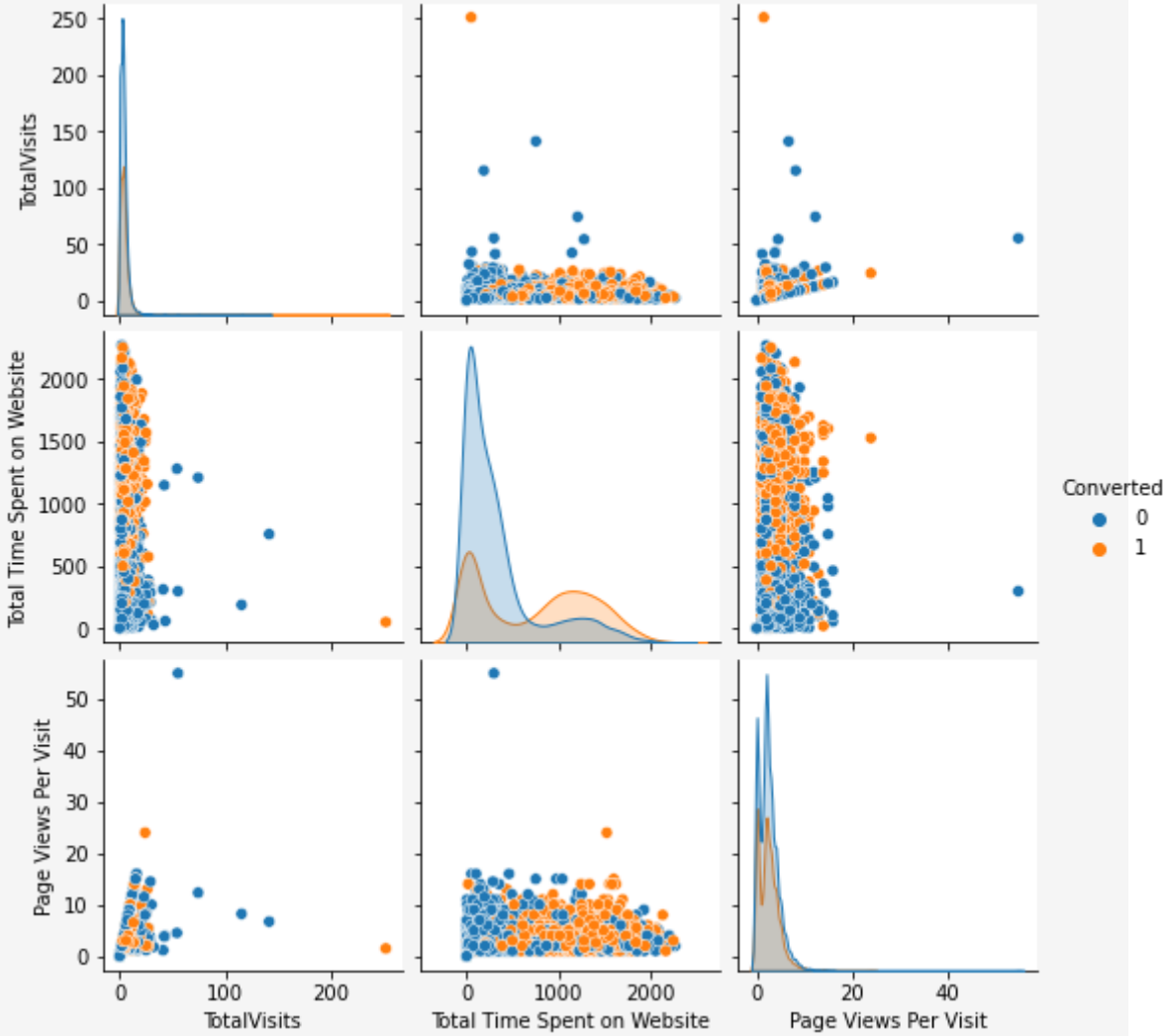


- There are many Last Notable Activity categories which do not add any value to our analysis, so combining them will benefit our analysis further.

VISUALIZING THE NUMERICAL VARIBALES USING PAIRPLOT

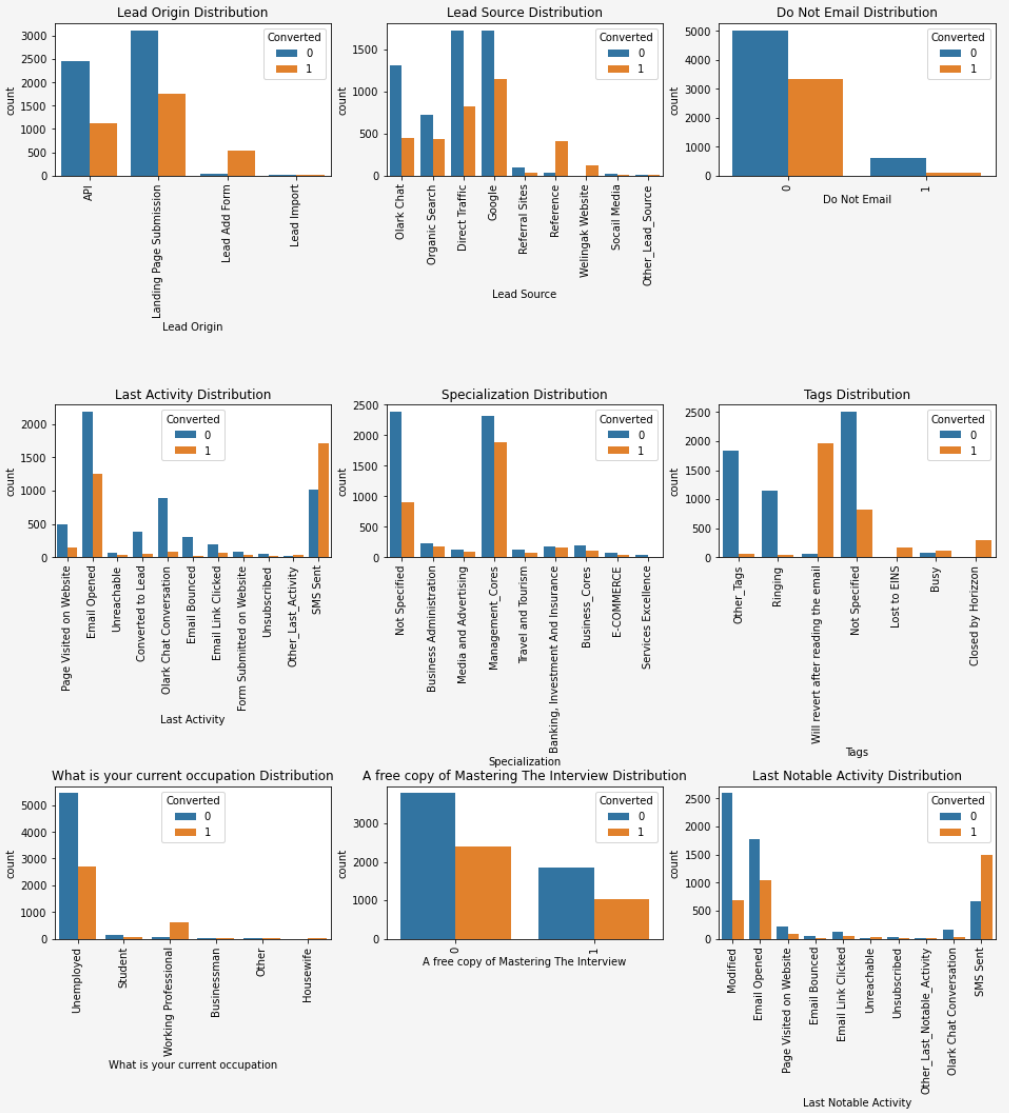
- Page Views Per Visit and Total Visits appear to be slightly correlated.
- Overall Time Spent on Website Seems Proportional to the Conversion.

- The linear association between these variables is not particularly strong.



VISUALIZING CATEGORICAL VARIABLES USING COUNTPLOT

- Lead Origin [Lead Add Form, Landing Page Submission] have a high conversion rate.
- If the Lead Source is from Google, Direct Traffic, Reference they seem to be a potential lead with high conversion rates.
- Customers who choose 'Do Not Email' seem to have a higher conversion than the customers who opt for Email.
- If the Last Activity is SMS Sent or Email Opened they have a higher chance of being converted.
- Customers browsing for various management courses seem to be converted.
- Tags with 'Will revert after reading the email' are highest converted followed by Closed by Horizon tags.
- High number of Working Professionals and Unemployed leads get converted.
- Leads who don't opt for 'A free copy of mastering The Interview' seem to convert more than who opt for it.
- If the Last Notable Activity is SMS Sent or Email Opened then they have a higher chance of being converted.



MODEL BUILDING

- Splitting the data into training and testing sets
- Currently we have around 50 variables in the dataset the data cleaning and preparation process, we cannot possibly use all these variables for model building. We have to choose the most appropriate variables/features that add value to our business goal so as to have a high accuracy and sensitivity and specificity. So we will be using both RFE (to narrow down to a small pool from a large number of features) and stats model for further feature selection.
- Using RFE with 15 variables as output.
- Using stats model to build a logistic regression model.
- Eliminating variables from a model whose p-value is higher than 0.05 and VIF value is higher than 5.
- Predictions on test data
- Obtaining overall accuracy

VISUALIZATION OF THE FINAL MODEL

Generalized Linear Model Regression Results

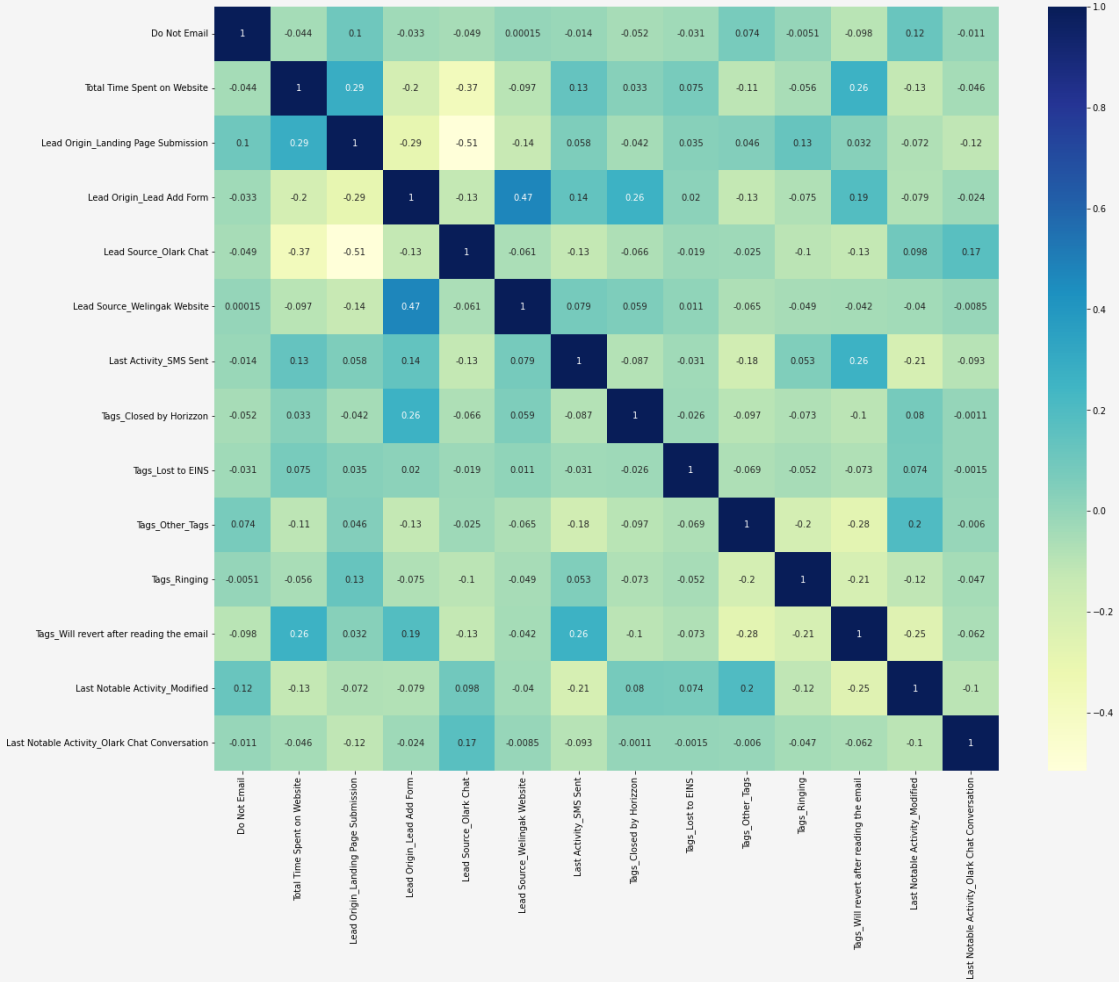
Dep. Variable:	Converted	No. Observations:	6351
Model:	GLM	Df Residuals:	6336
Model Family:	Binomial	Df Model:	14
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-1165.3
Date:	Mon, 20 Mar 2023	Deviance:	2330.7
Time:	20:19:18	Pearson chi2:	8.42e+03
No. Iterations:	8	Pseudo R-squ. (CS):	0.6194
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0692	0.126	-8.500	0.000	-1.316	-0.823
Do Not Email	-1.1083	0.274	-4.043	0.000	-1.646	-0.571
Total Time Spent on Website	1.1495	0.064	18.015	0.000	1.024	1.275
Lead Origin_Landing Page Submission	-0.8718	0.138	-6.324	0.000	-1.142	-0.602
Lead Origin_Lead Add Form	1.3336	0.482	2.765	0.006	0.388	2.279
Lead Source_Olark Chat	0.8729	0.170	5.128	0.000	0.539	1.207
Lead Source_Welingak Website	3.8368	0.871	4.406	0.000	2.130	5.543
Last Activity_SMS Sent	2.1229	0.121	17.573	0.000	1.886	2.360
Tags_Closed by Horizzon	6.9566	0.739	9.418	0.000	5.509	8.404
Tags_Lost to EINS	6.2480	0.740	8.442	0.000	4.797	7.699
Tags_Other_Tags	-2.2731	0.192	-11.835	0.000	-2.650	-1.897
Tags_Ringing	-3.6948	0.261	-14.145	0.000	-4.207	-3.183
Tags_Will revert after reading the email	4.9529	0.211	23.453	0.000	4.539	5.367
Last Notable Activity_Modified	-1.8052	0.132	-13.703	0.000	-2.063	-1.547
Last Notable Activity_Olark Chat Conversation	-1.8130	0.432	-4.194	0.000	-2.660	-0.966

	Features	VIF
2	Lead Origin_Landing Page Submission	2.25
3	Lead Origin_Lead Add Form	1.98
11	Tags_Will revert after reading the email	1.79
12	Last Notable Activity_Modified	1.60
6	Last Activity_SMS Sent	1.57
4	Lead Source_Olark Chat	1.52
1	Total Time Spent on Website	1.45
9	Tags_Other_Tags	1.43
5	Lead Source_Welingak Website	1.36
10	Tags_Ringing	1.28
7	Tags_Closed by Horizzon	1.24
0	Do Not Email	1.13
8	Tags_Lost to EINS	1.07
13	Last Notable Activity_Olark Chat Conversation	1.07

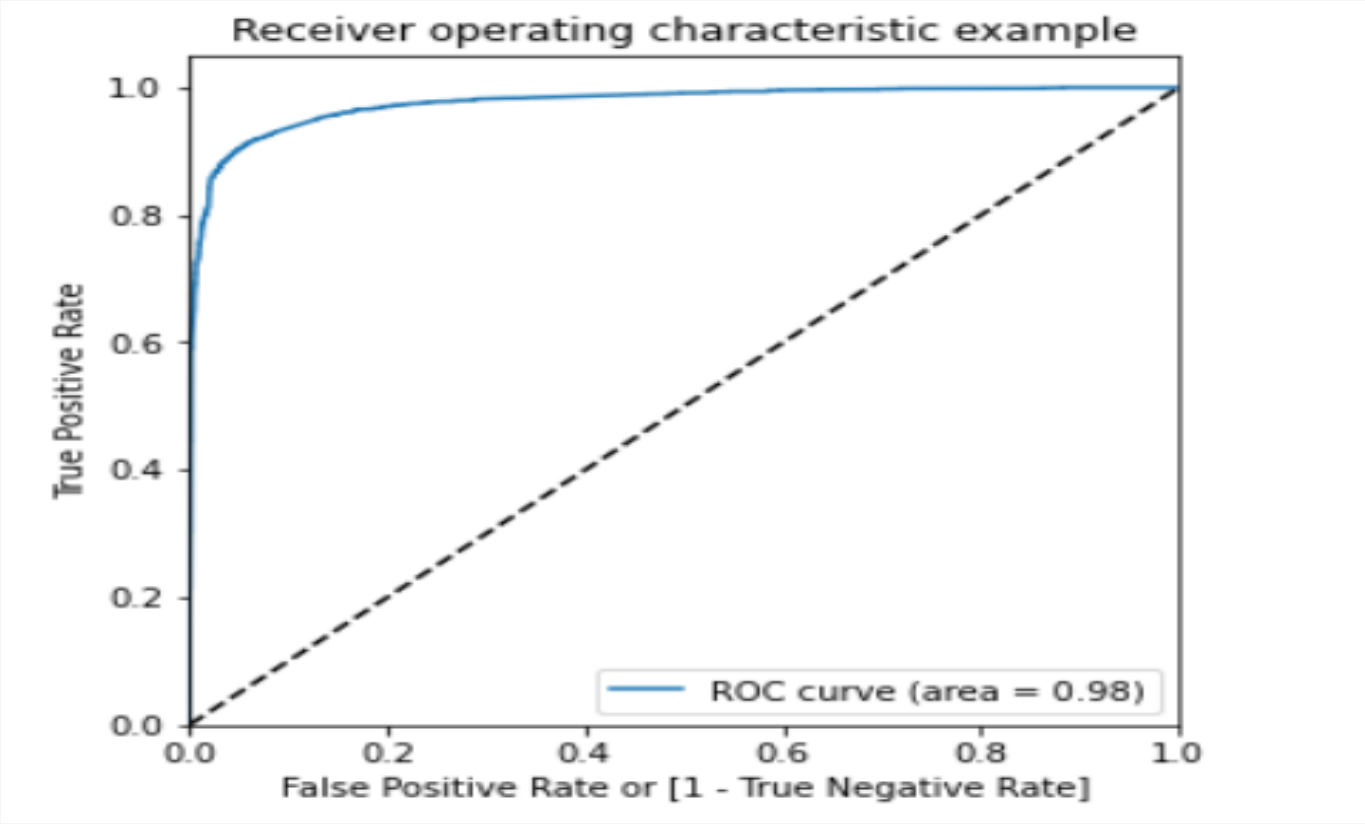
- The Model built seems to have a very low p-value (< 0.005) and VIF (multi-collinearity : < 3)

CORRELATION MATRIX AFTER FEATURE SELECTION



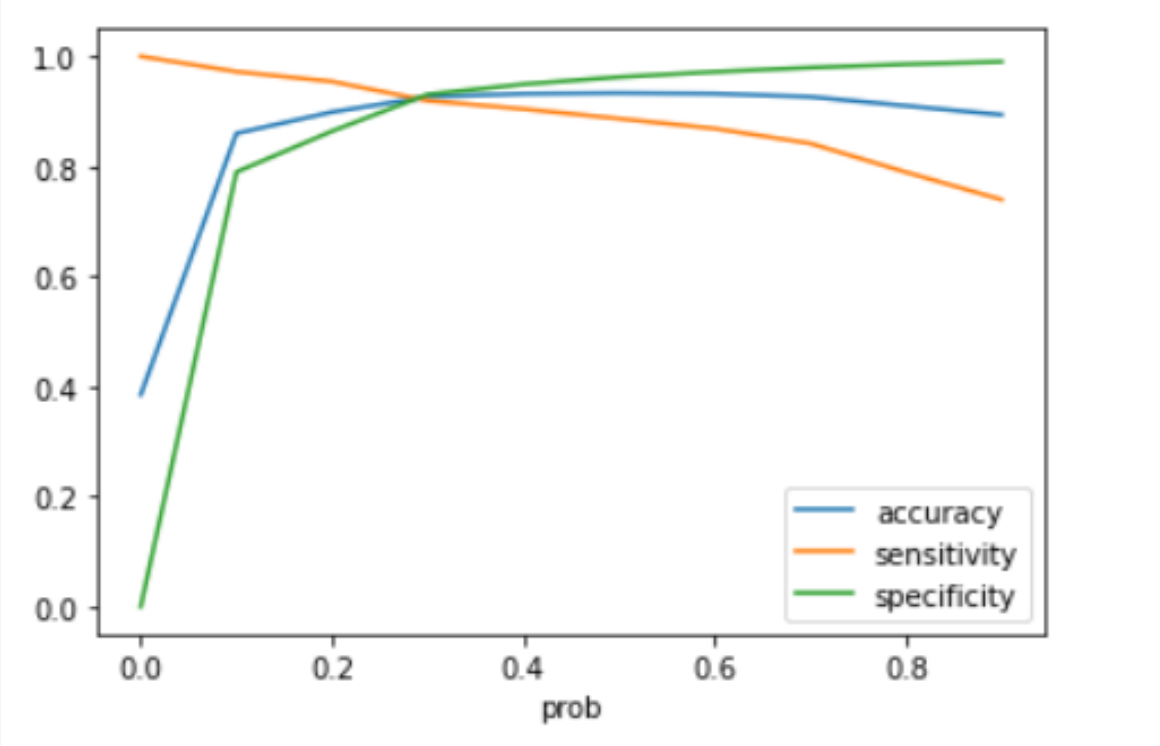
- There doesn’t seem to be much collinearity between the features to affect the prediction of the model.

ROC CURVE



- The Area under the ROC curve is 0.98, which means the model built is a good predictive model

FINDING OPTIMAL THRESHOLD



- According to the curve above, 0.3 is the ideal cutoff threshold value.

FINAL OBSERVATIONS ON TRAIN AND TEST DATA:

Final Observations (Train and Test Data):

Train Data:

1. Confusion matrix - [[3635 270]
   [197 2249]]
2. Accuracy - 92.64 %
3. Sensitivity - 91.94 %
4. Specificity - 93.08 %
5. Precision - 89.28 %
6. Recall - 91.94 %

Test Data:

1. Confusion matrix - [[1606, 128],
   [96, 893]]
2. Accuracy - 91.77 %
3. Sensitivity - 90.29 %
4. Specificity - 92.61 %
5. Precision - 87.46 %
6. Recall - 90.29 %

- There is a decrease in the values of the metrics for the test data set compared to the train data set, but it is within 5% This is an acceptable rate. Hence we can conclude that it is a good generic model.

CONCLUSION

VALUABLE INSIGHTS:

- We can see that when the leads are marked with tags such as 'Closed by Horizzon', 'Lost to EINS', and 'Will revert after reading the email' are likely to get converted to paid customers and are considered as hot/potential leads. Likewise, leads from sources like Welingak Website or the last activity is an SMS Sent are likely to be converted to paid customers. The sales team must monitor the leads from these sources and occupations and be quick with their first approach or send out brochures to lock the client as soon as possible. Also, if a user is spending more time on the website browsing for courses, that user is likely a potential lead, so the sales team has to keep an eye out for users spending a high total time on their website as well. These leads would be the potential leads that will be converted if given the time and effort from the sales team. This will increase the conversion from 30% up to or more than 80%.

- The sales team can come to a conclusion internally that if the lead score is greater than 90 then they can invest more time in a day to these customers for a better conversion rate and gradually decrease the time invested in each customer as the lead score decreases.

THANK YOU