Lead Score Case Study

Presenter Name: Truong Tuyet Lam

Period: January 2025

Agenda

1. Problem Statement & Business Goals
2. Key Findings from Exploratory Data Analysis
3. Model Building & Evaluation
4. Conclusions and Recommendations

1. Problem Statement & Business Objectives

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.
- X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Business Objectives

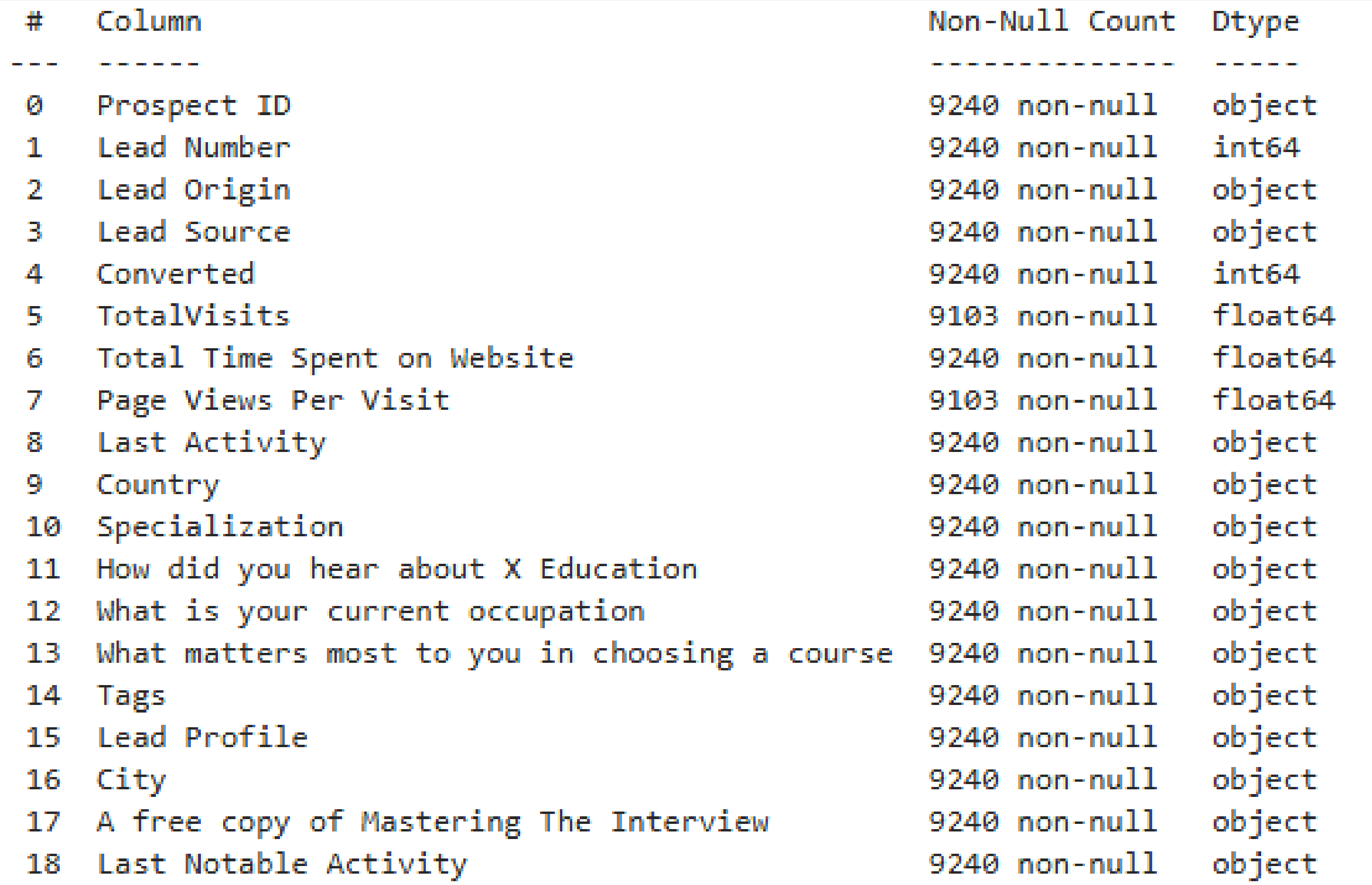
X Education would like to:

1. Select the most promising leads (i.e. the leads that are most likely to convert into paying customers by building a logistic regression model wherein each of the leads will be assigned a lead score (between 0 and 100) such that the customers with a higher lead score having higher coversion chance whereas the customers with a lower lead score having a lower conversion chance. The CEO in particular also has given a ballpark target leader conversion rate at around 80%.
2. Identify the top three variables which contribute the most towards the probabillity of a lead getting converted
3. Identify the top 3 most important categorical/dummy variables to focus on in order to increase the probability of lead conversion
4. Define a good strategy to maximize the conversion rate during the 2 months they have more intern headcounts as well as minimize the rate of useless phone-calls during the idle period.

2. Key Findings From EDA

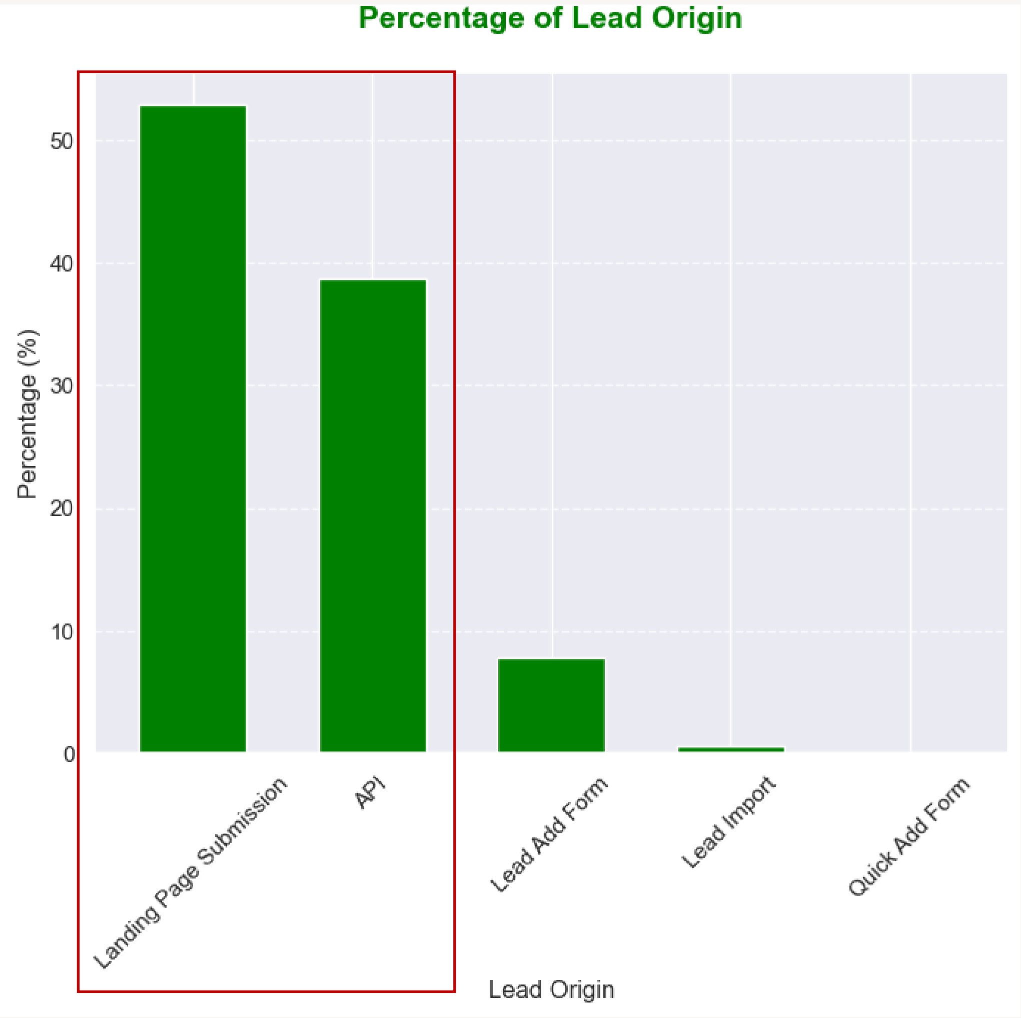
Key factors for analysis

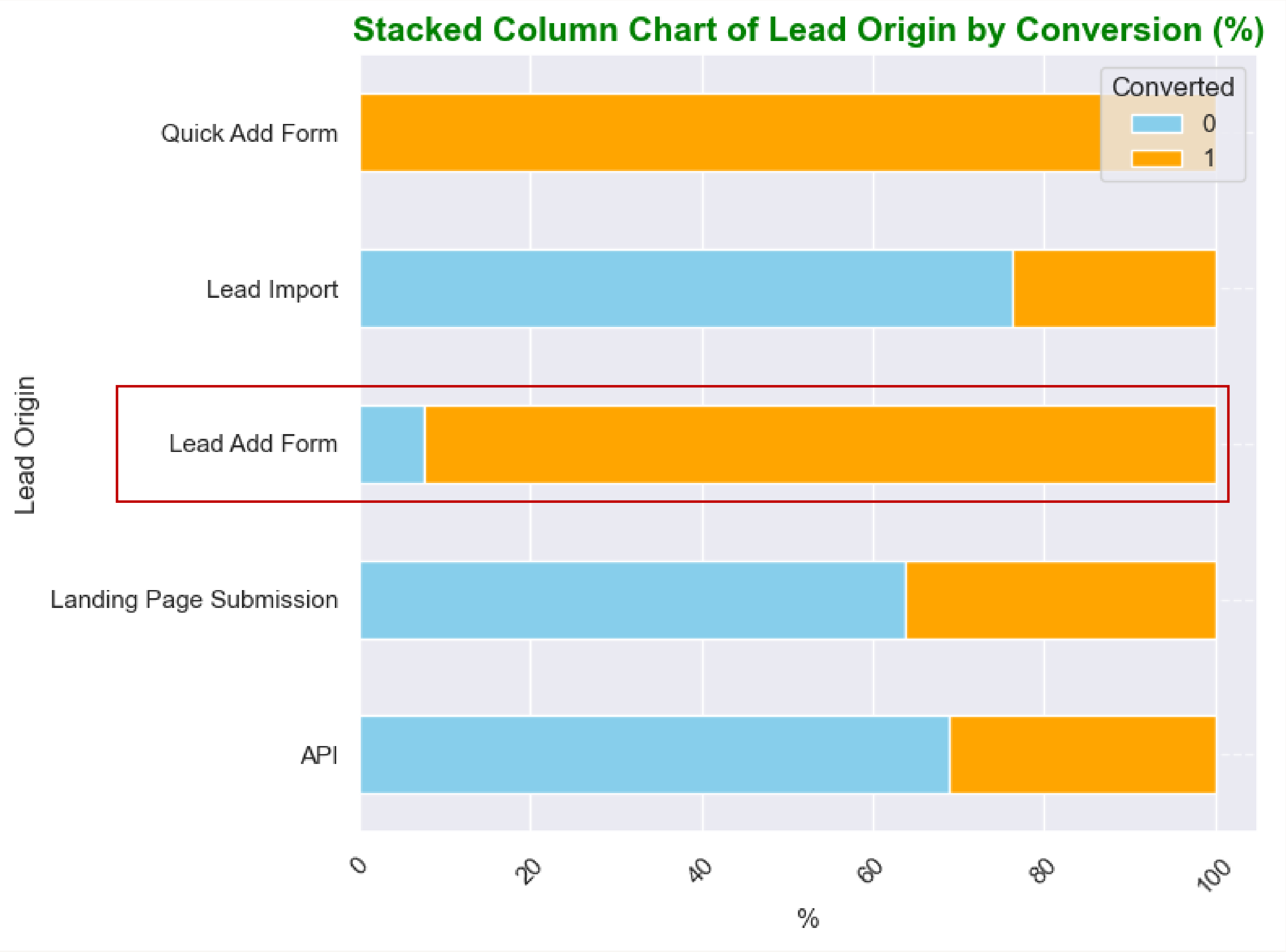
After checking data and treating missing values, below are key columns with good data quality to proceed further for EDA



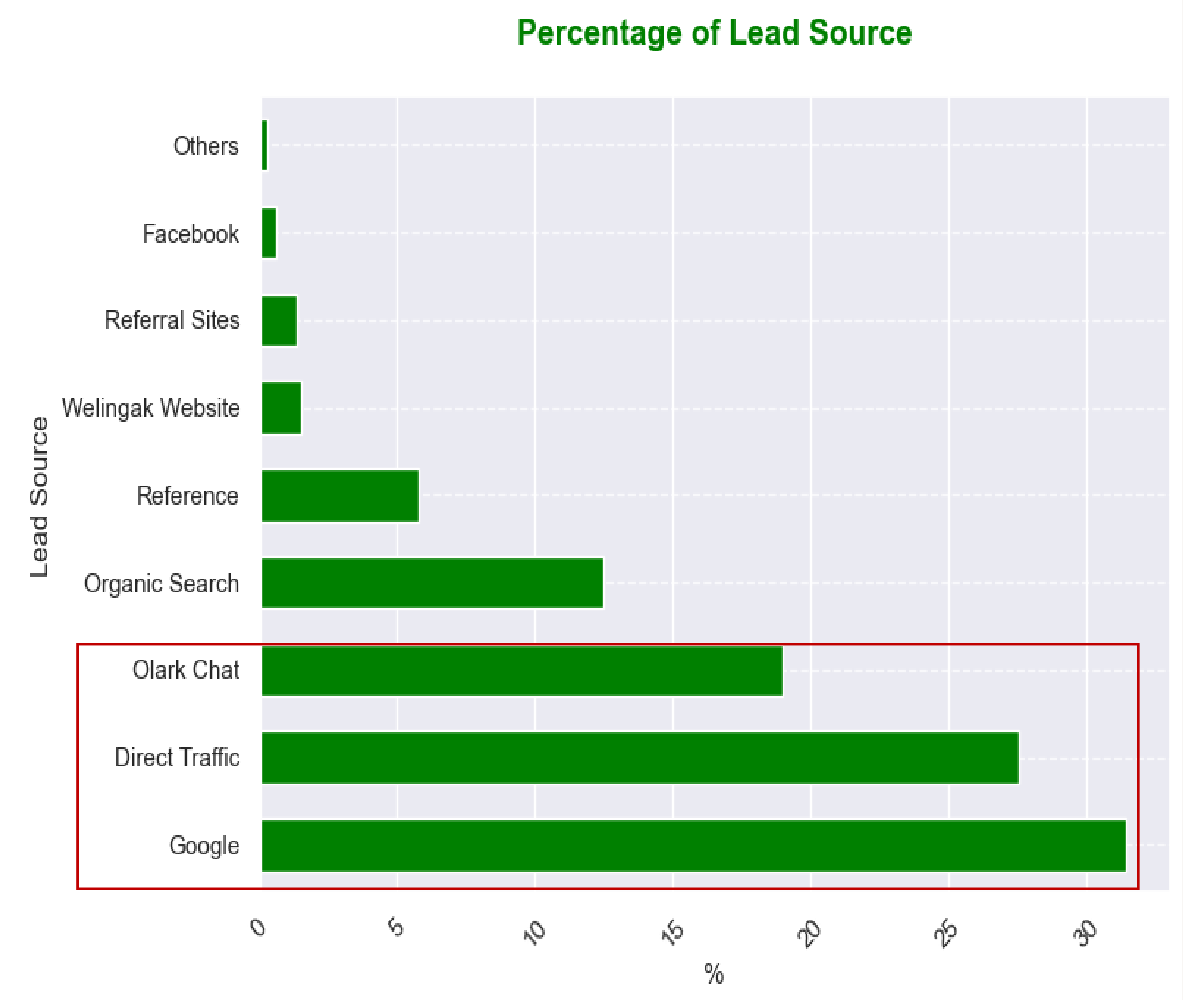
```
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Prospect ID                                   9240 non-null   object
 1   Lead Number                                   9240 non-null   int64
 2   Lead Origin                                   9240 non-null   object
 3   Lead Source                                   9240 non-null   object
 4   Converted                                     9240 non-null   int64
 5   TotalVisits                                   9103 non-null   float64
 6   Total Time Spent on Website                   9240 non-null   float64
 7   Page Views Per Visit                          9103 non-null   float64
 8   Last Activity                                 9240 non-null   object
 9   Country                                       9240 non-null   object
 10  Specialization                                9240 non-null   object
 11  How did you hear about X Education            9240 non-null   object
 12  What is your current occupation               9240 non-null   object
 13  What matters most to you in choosing a course 9240 non-null   object
 14  Tags                                          9240 non-null   object
 15  Lead Profile                                  9240 non-null   object
 16  City                                          9240 non-null   object
 17  A free copy of Mastering The Interview        9240 non-null   object
 18  Last Notable Activity                         9240 non-null   object
```

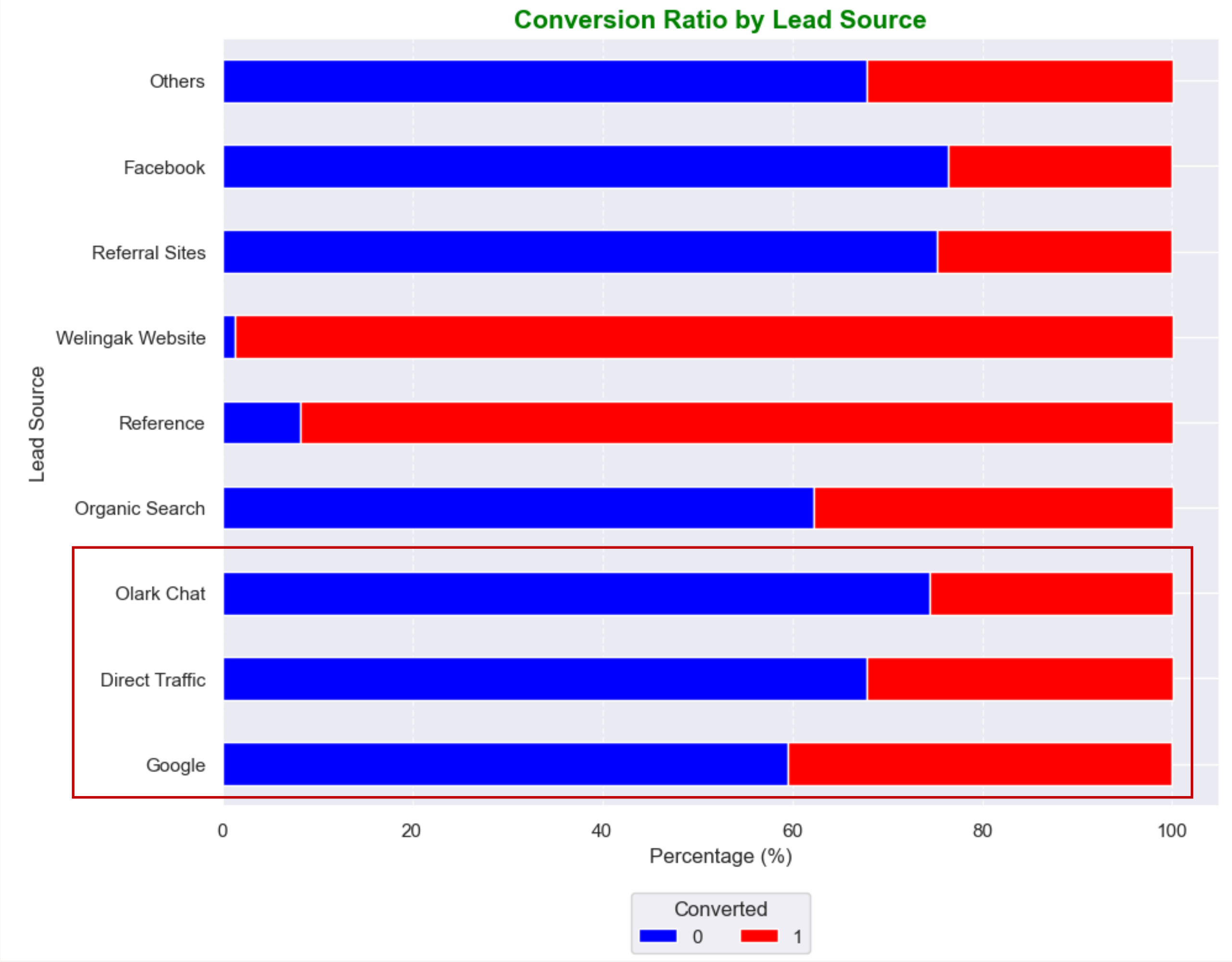
Lead Origin: "Landing page submission" (53%) and "API" (39%) are the most 2 popular origins of lead. Next is "Lead Add Form" (8%). "Landing page submission" also have higher conversion rate than "API". "Lead Add Form" though is less popular, yet the ratio of conversion rate is very high.



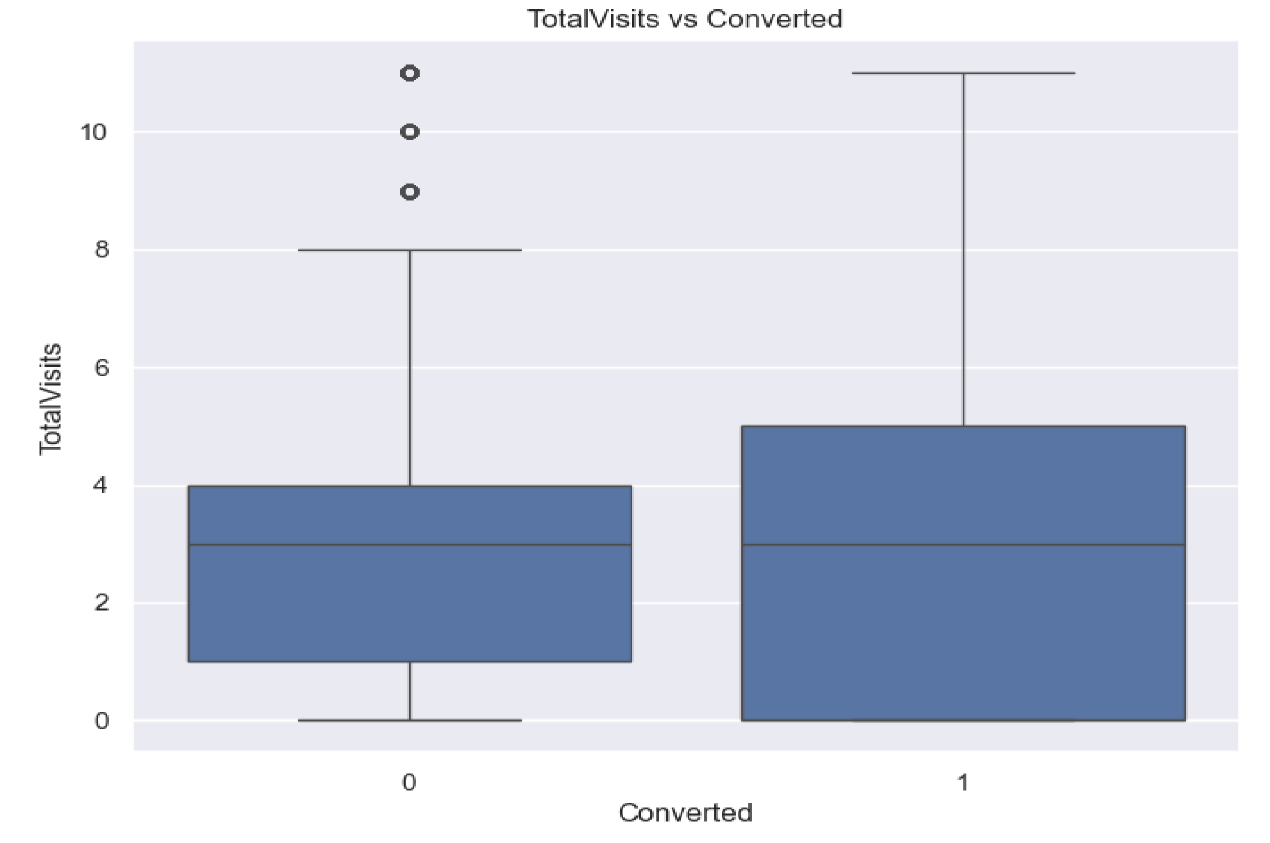


Lead Source: "Google", "Direct Traffic" and "Olark Chat" are the top 3 most common channel of Lead Source. Another channels that are also potential to focus with lower percentage of Lead Source include "Organic Search", "Reference" and "Welingak Website". Among those channel, "Reference" and "Welingak Website" have much higher conversion rate though being less popular.

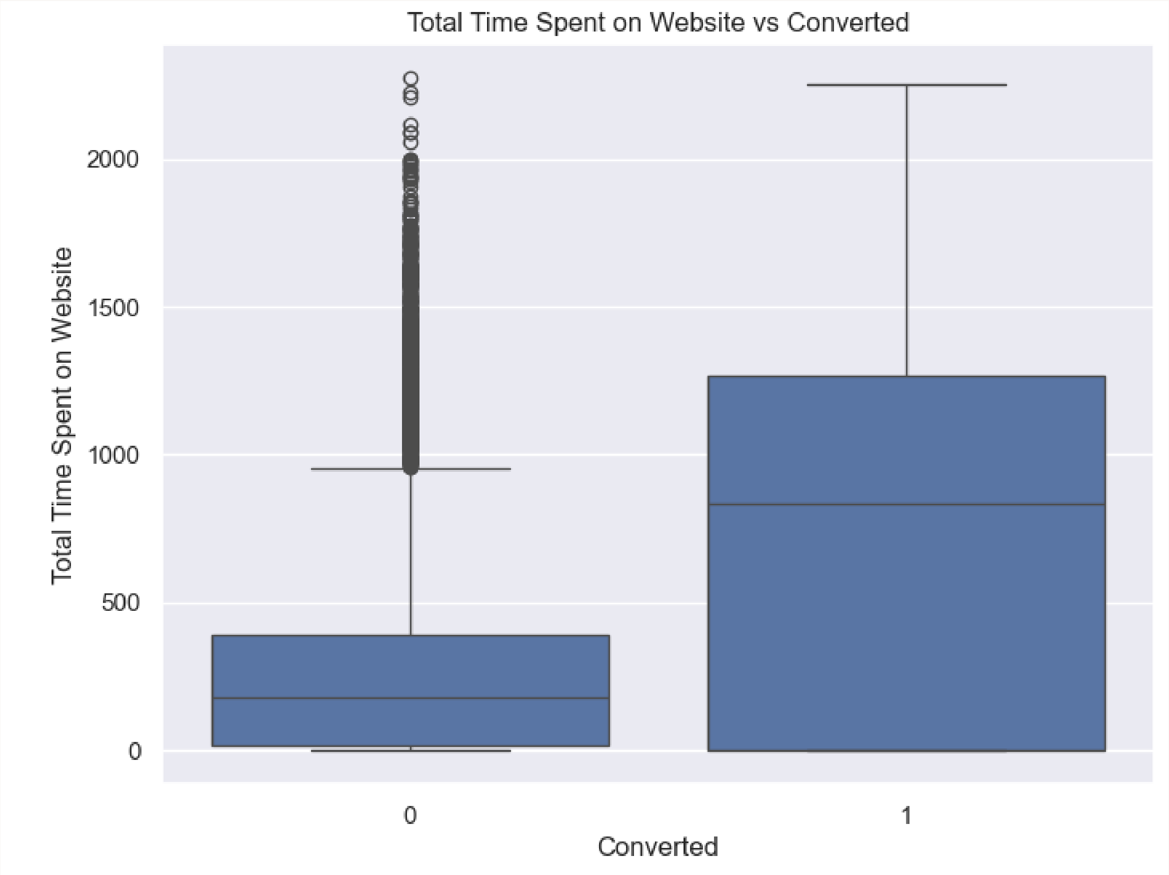


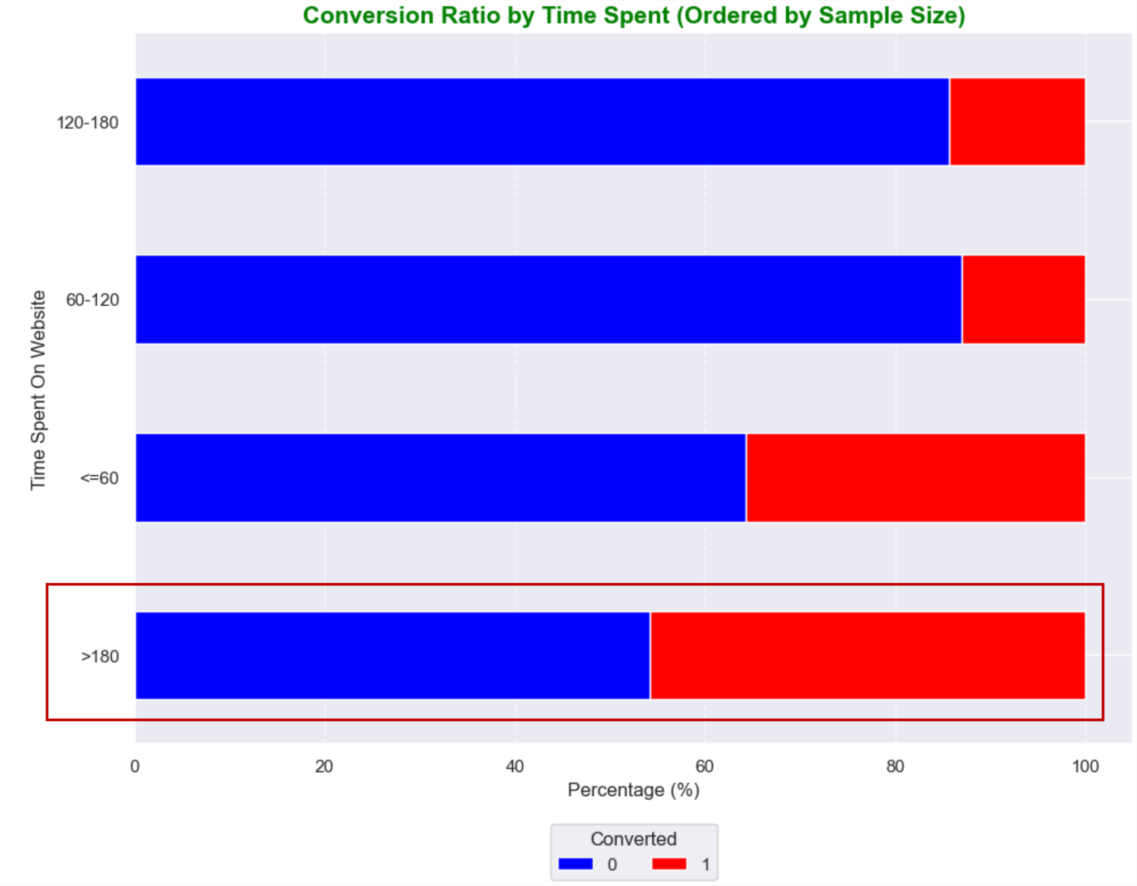


Total Visits: On average, the customers tend to visit 3.2 times and the frequency varies from a range of 1 - 5 times. The mean score between converted and non-converted is quite similar, however, the range of 2 groups is quite different, given the frequency of visits among those being converted is higher than those not converted. Thus, ‘TotalVisits” might be considered as one of important factors when defining potential lead

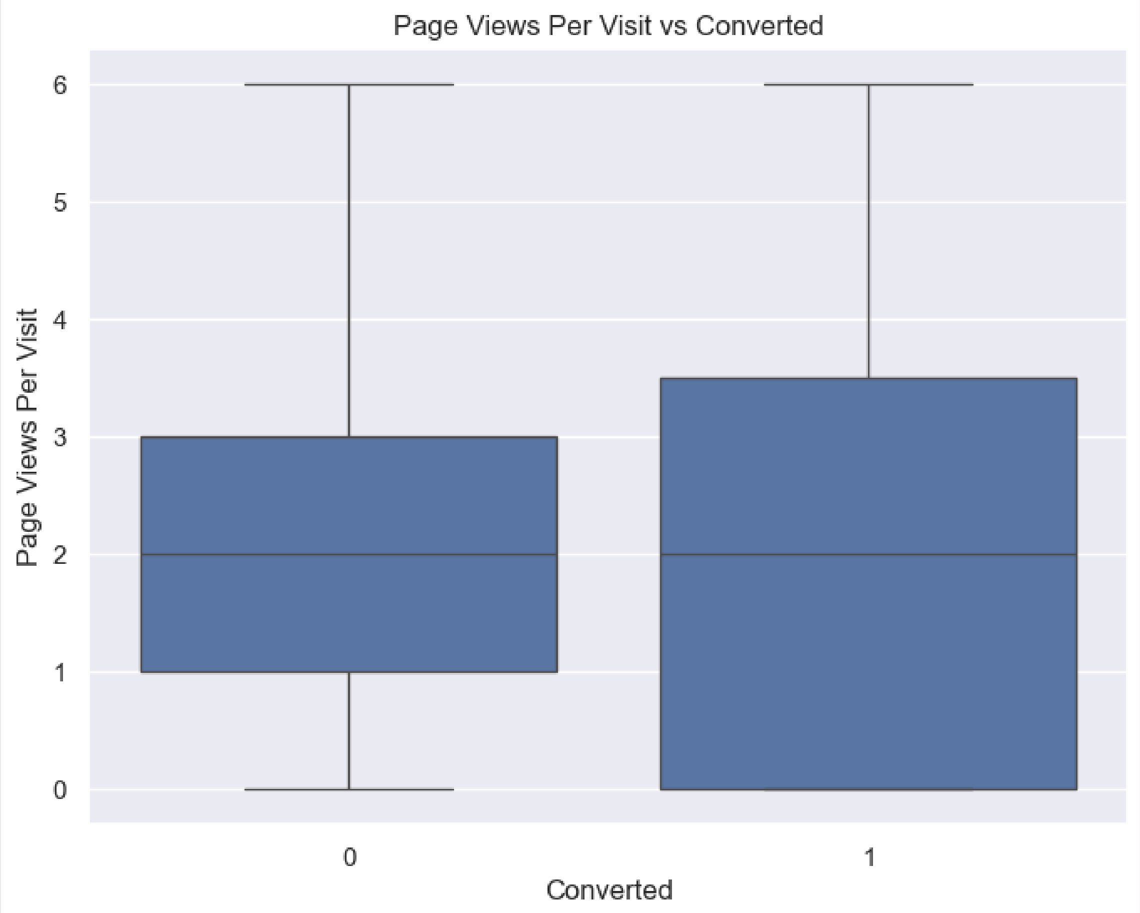


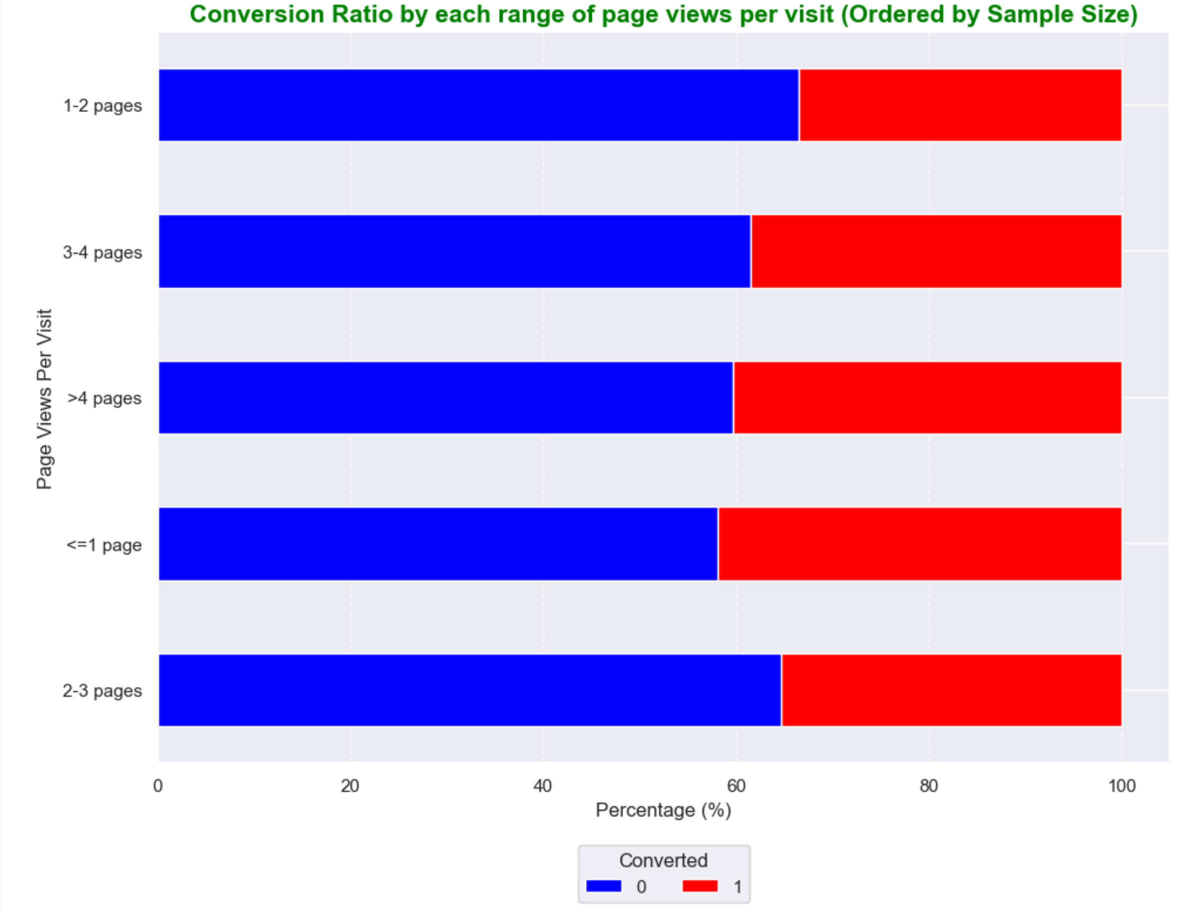
Total time spent on website: 56% of students spend more than 3 hours on the website. Among those who spent >180, 46% of them are converted. Therefore, the conversion rate is much higher among those spending more time on website. Therefore, this factor can also be considered as another critical factor contributing to conversion rate.



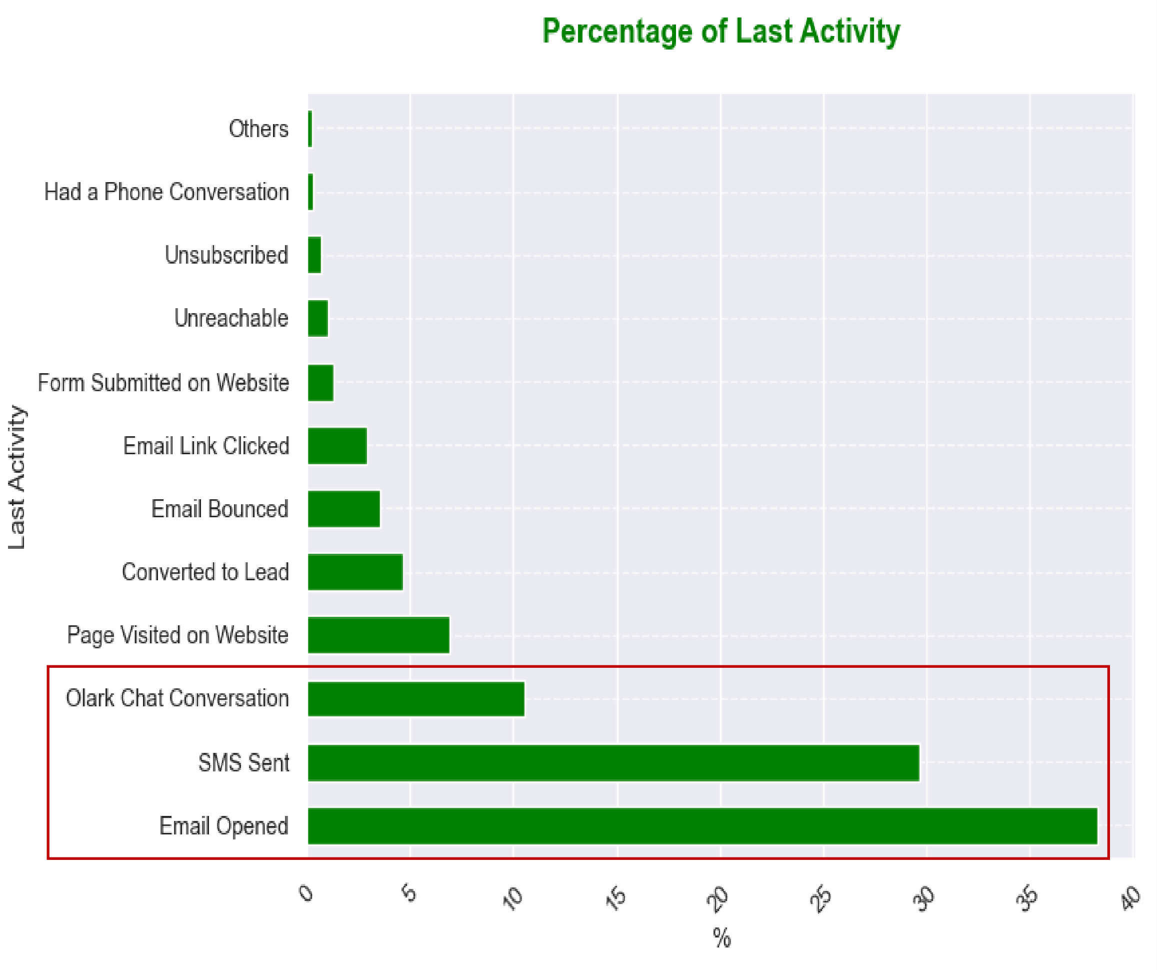


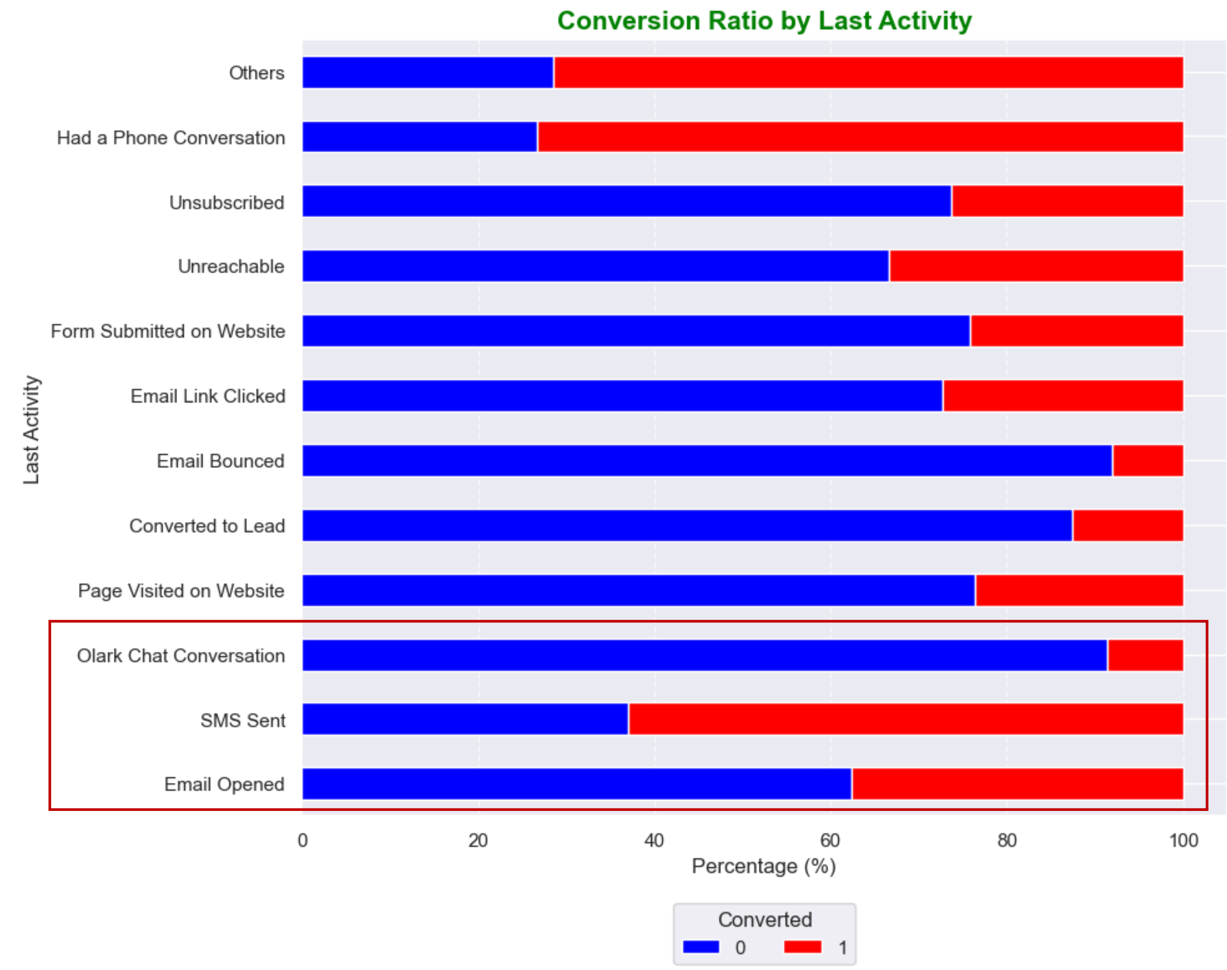
Page Views Per Visit: There is no significant difference in terms of number of pages view per visit of customers across ranges. Thus, this factor seems not be a key factor to differentiate the behaviors for conversion.



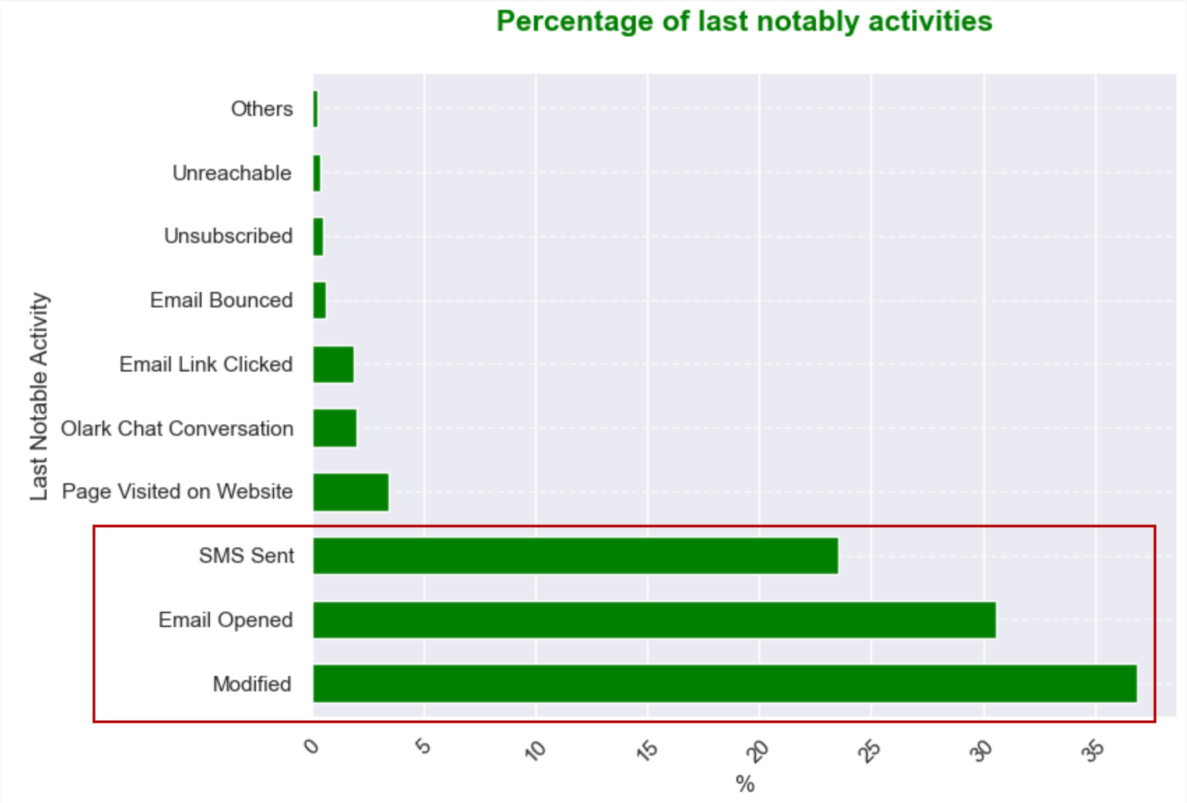


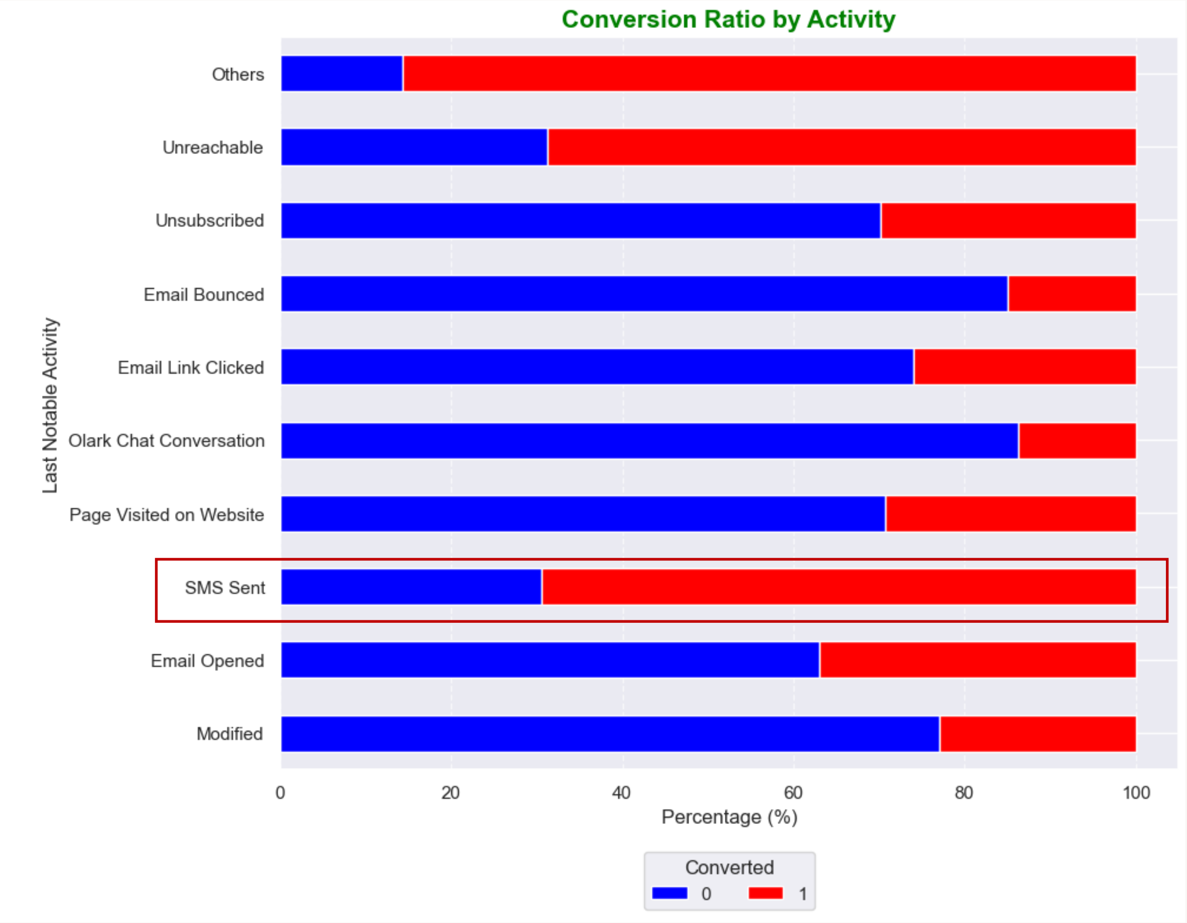
Last Activity: "Email Opened", "SMS Sent" and "Olark Chat Conversion" are the top 3 most common channels that customers interacted lastly. Though "Email Opened" got the highest percentage among the top 3 channels, its conversion rate is not as high as the second most popular "SMS Sent". Therefore, "SMS Sent" is seen as an important factor to consider.



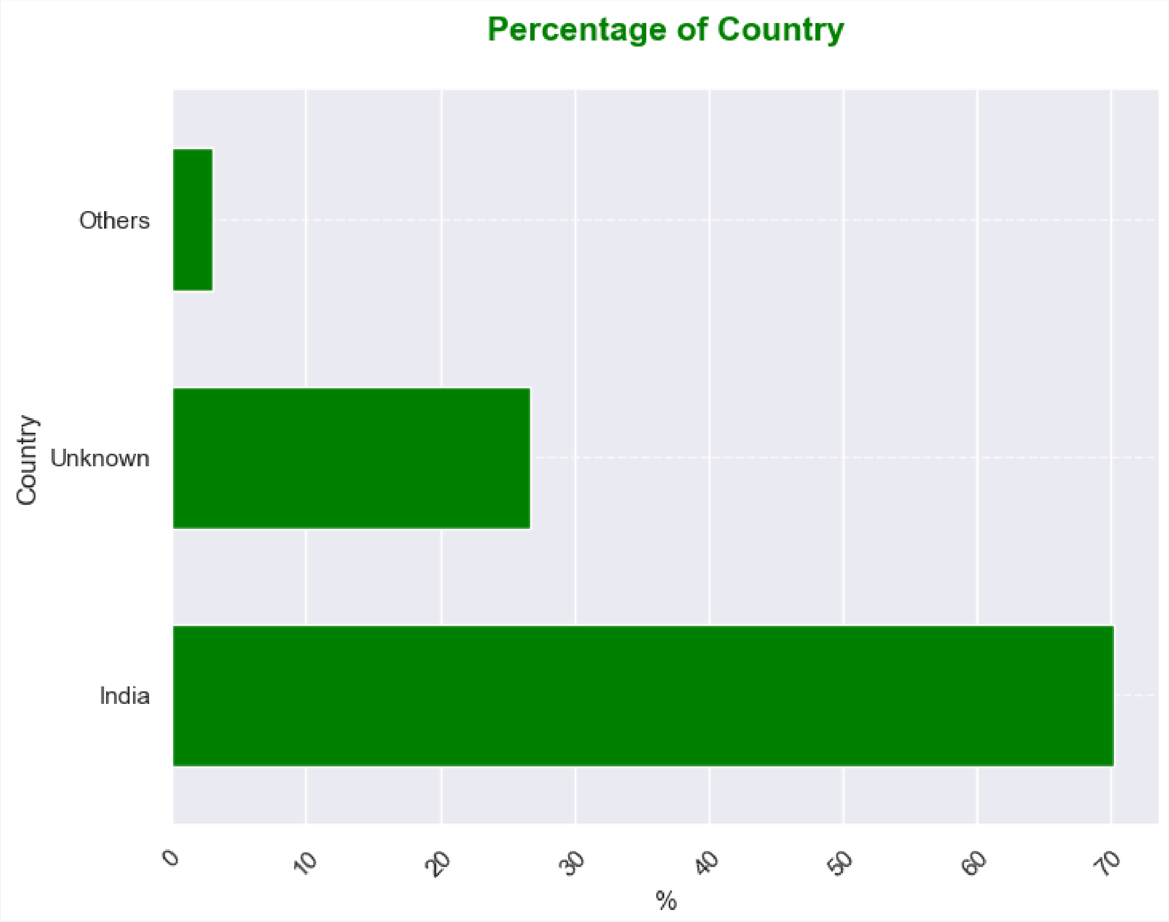


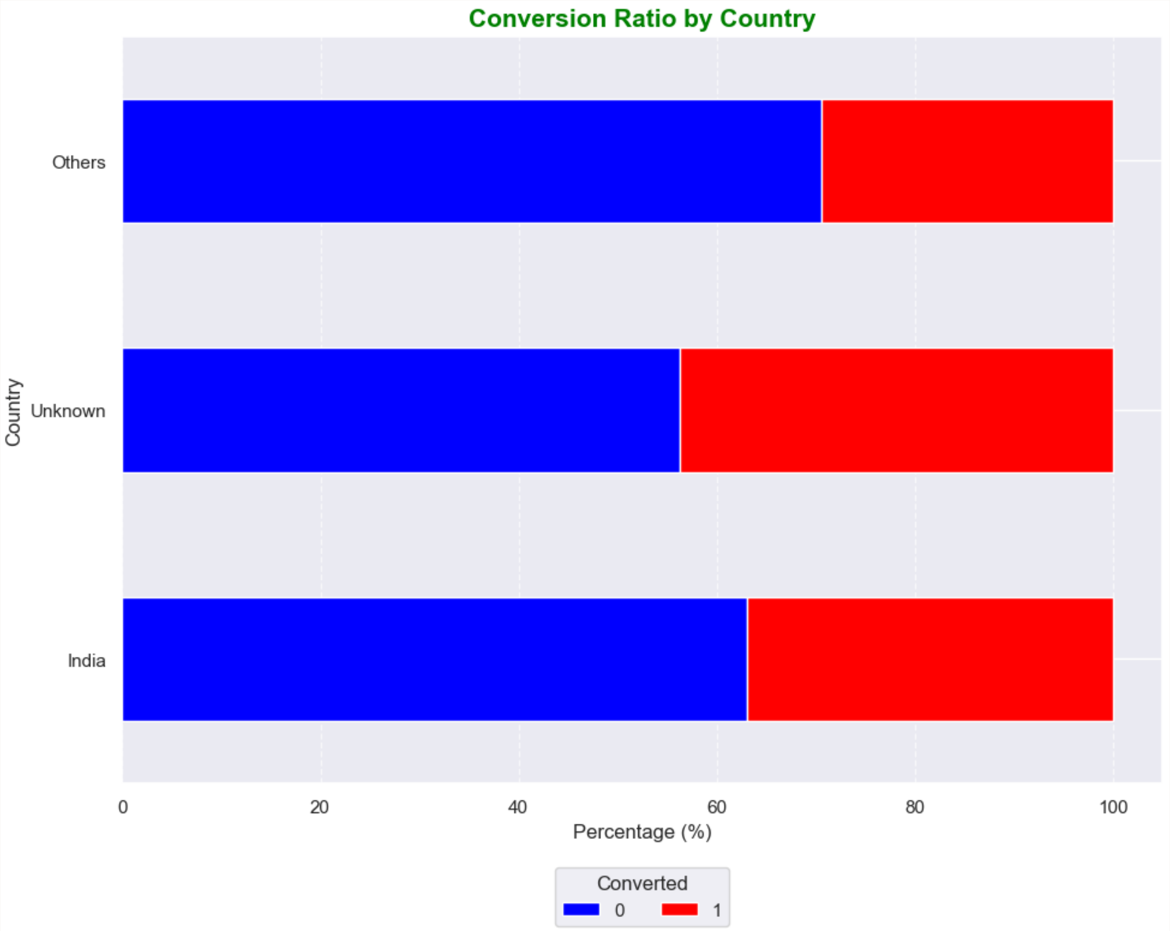
Last Notable Activity: Top 3 most common last notably activities include "Modified", "Email Opened" and "SMS Sent". Among these 3 activities, "SMS Sent" got the highest ratio of conversion rate (63%). This fact is also similar with the finding from 'Last Activity'.



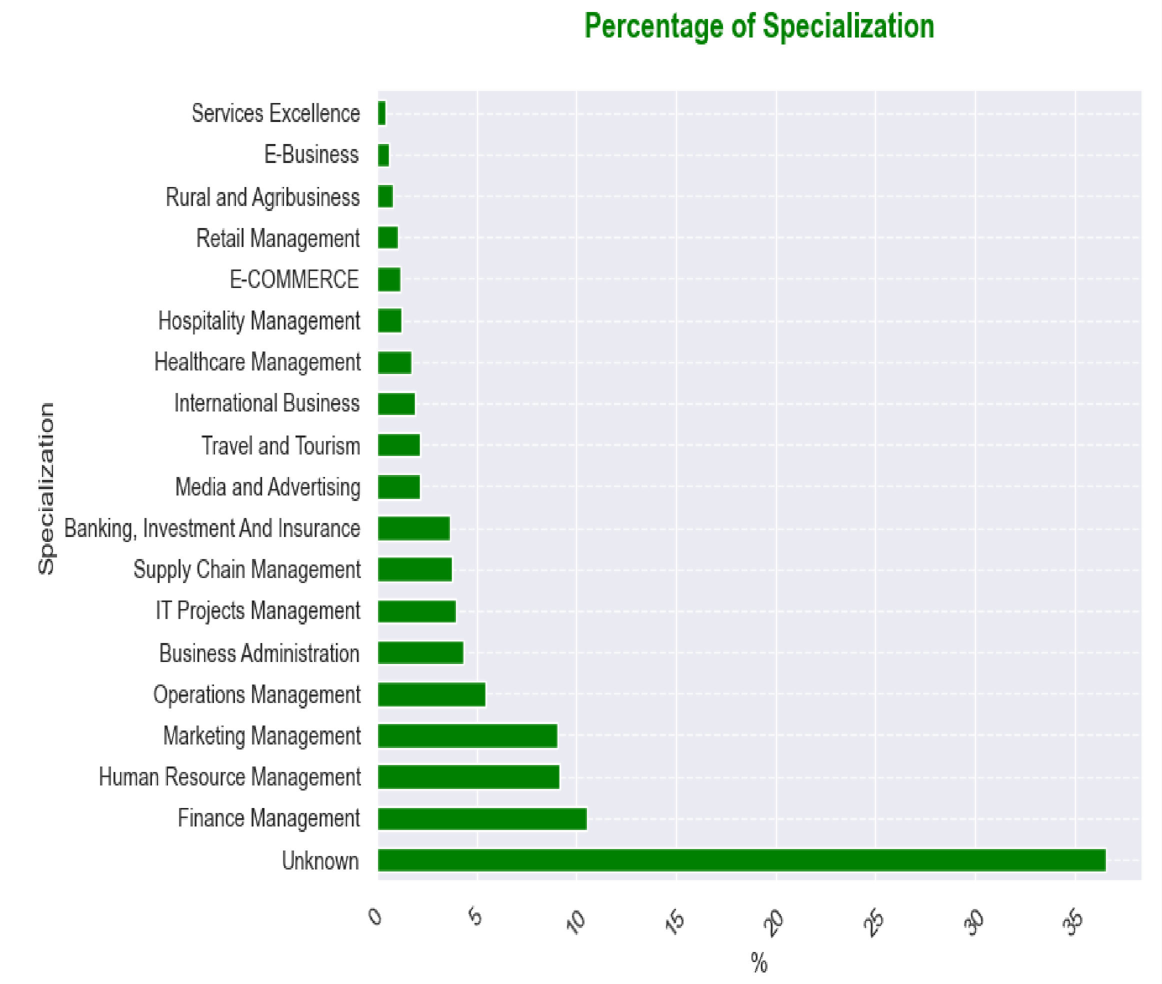


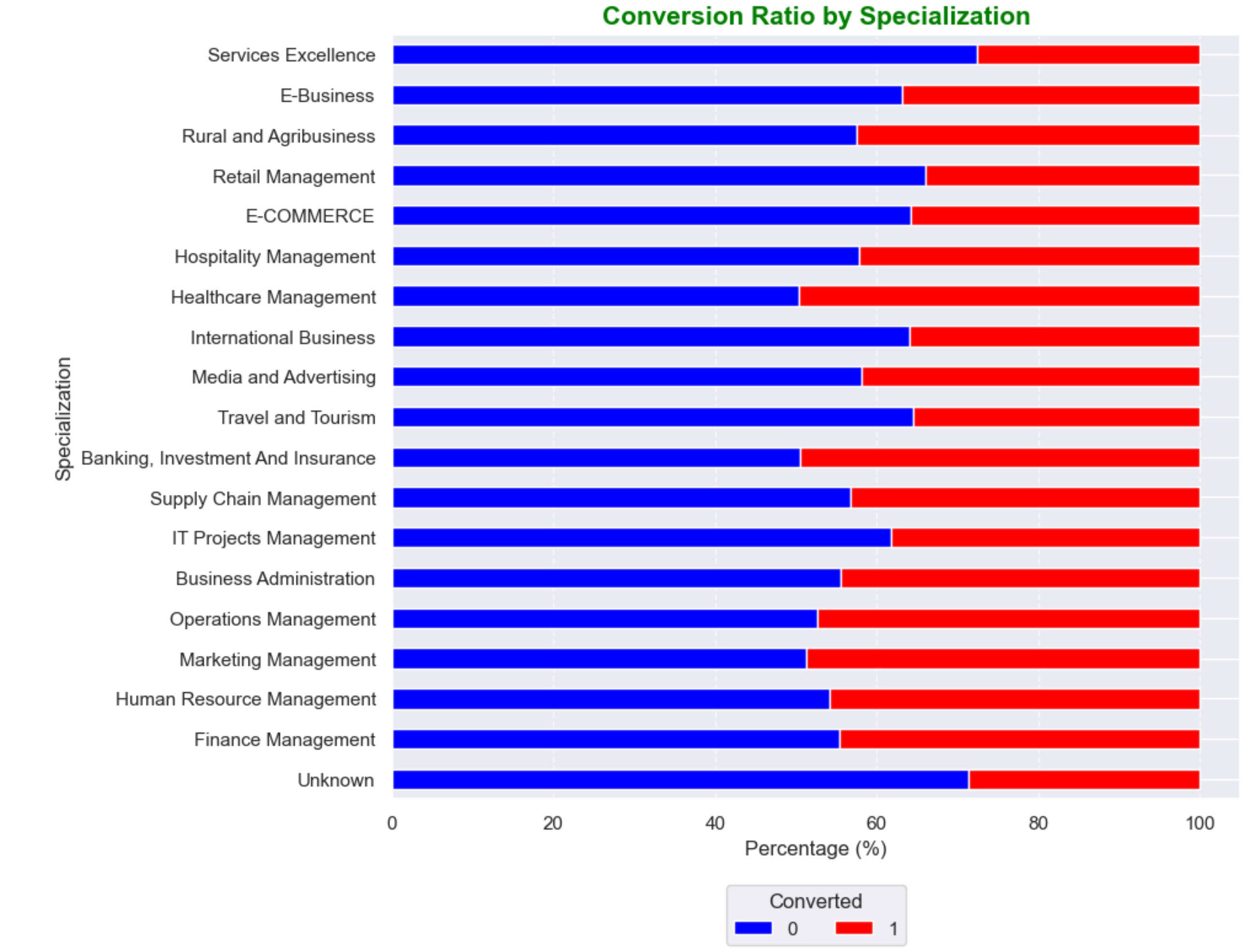
Country: Majority of students are from India (70%), 27% of customers are unknown. Meanwhile, the conversion rate among unknown customers are higher than Indian ones.



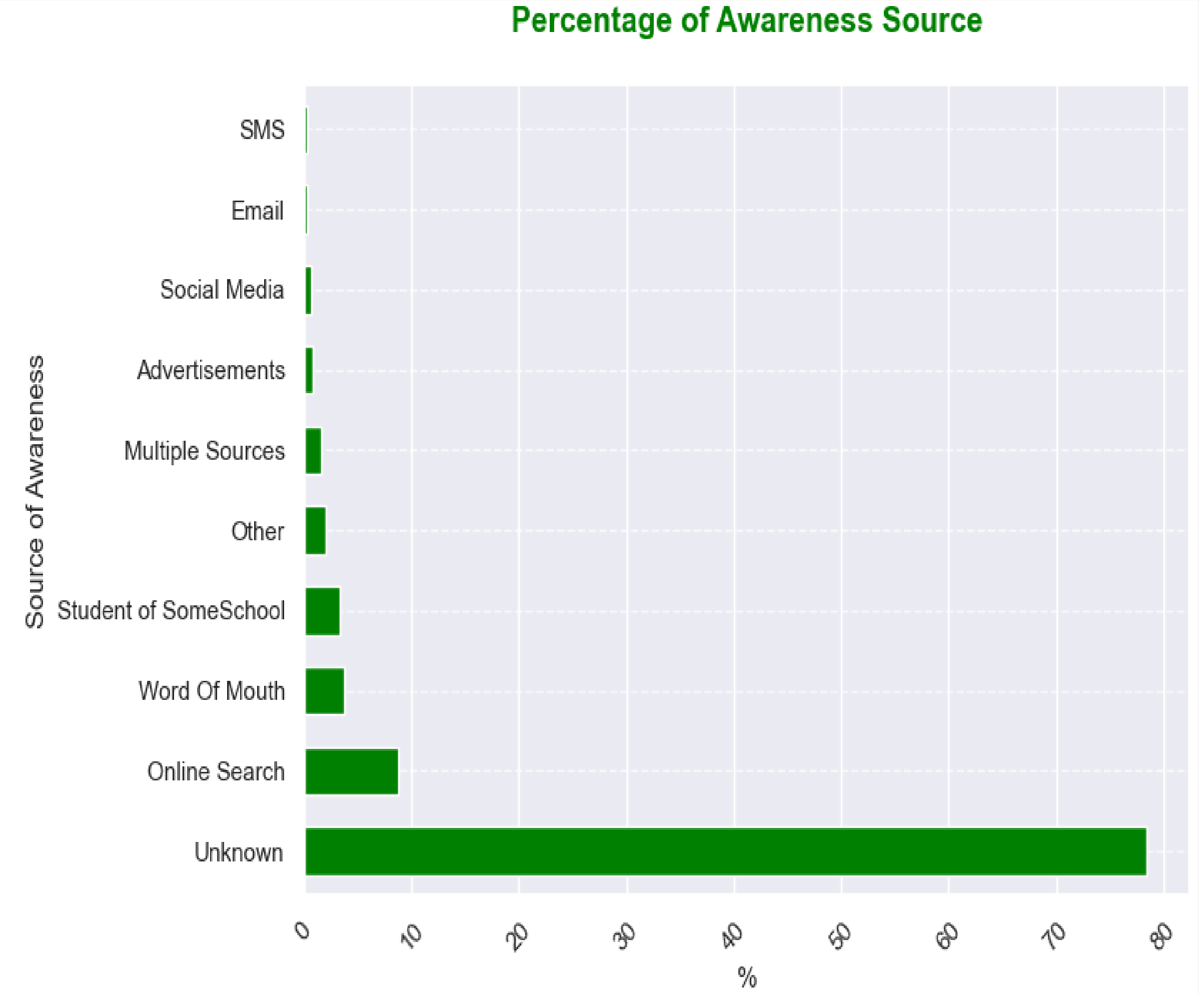


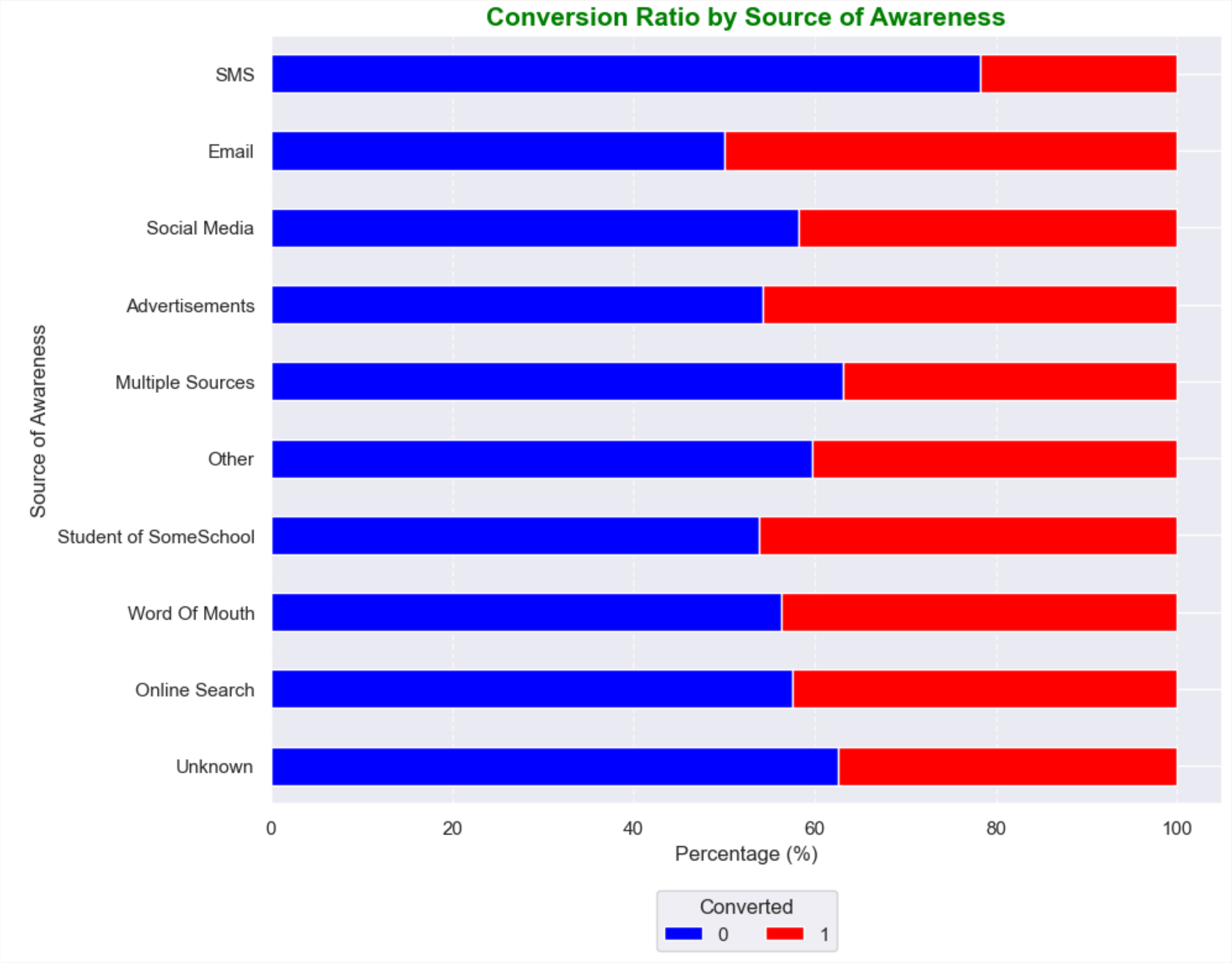
Specialization: "Unknown" occupies 37% of data of this column. Among known data for this column, customers choose a wide range of specializations from various domains including marketing, finance, HR, operation, etc. Among those, top 3 most common specializations are finance management, HR management, Marketing management.



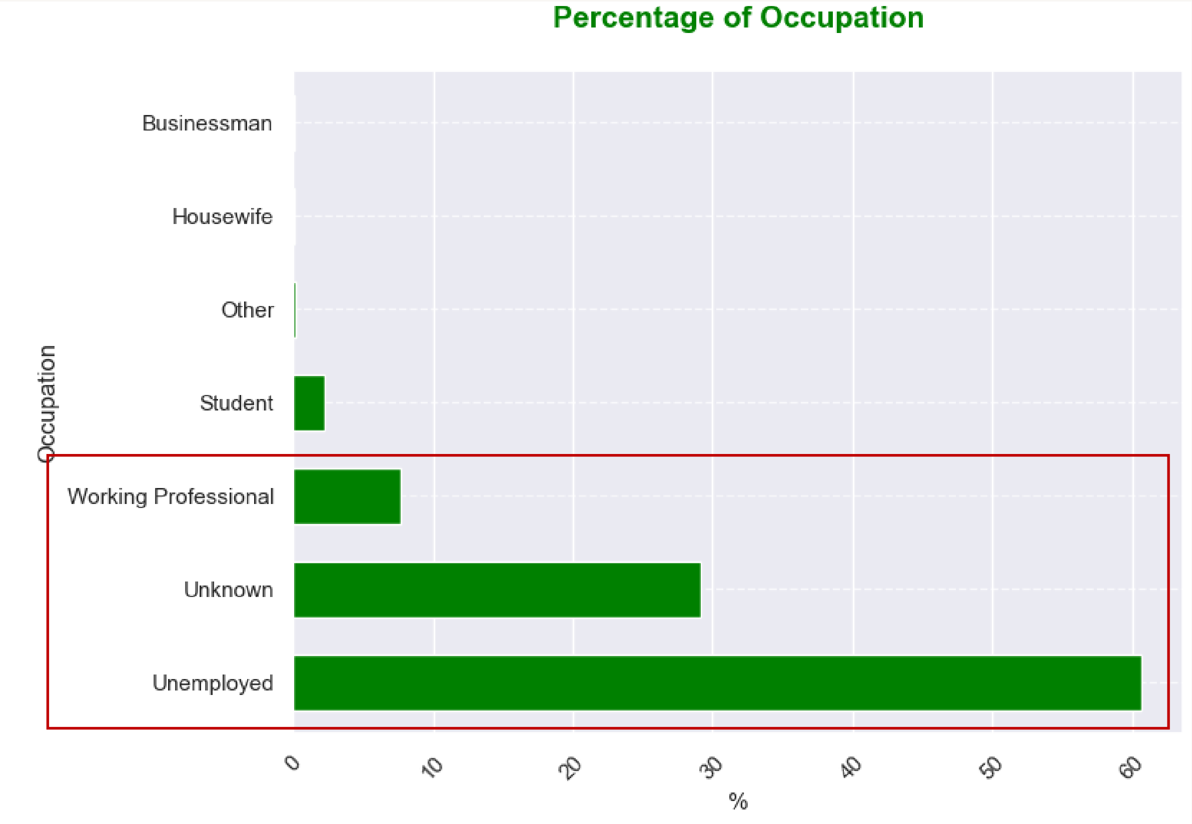


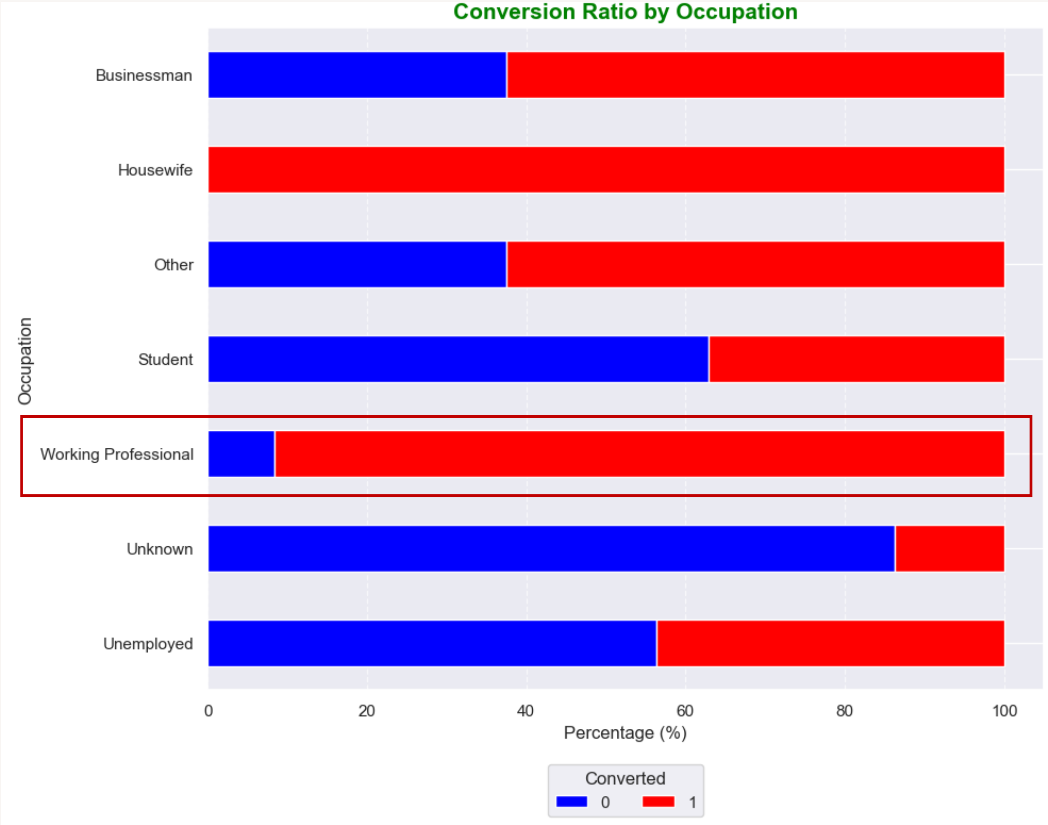
Source of Awareness: 78% of students don't provide information about the source of awareness. Among provided data, online search and word of mouth are another top 2 common sources of awareness which have quite similar conversion rate.



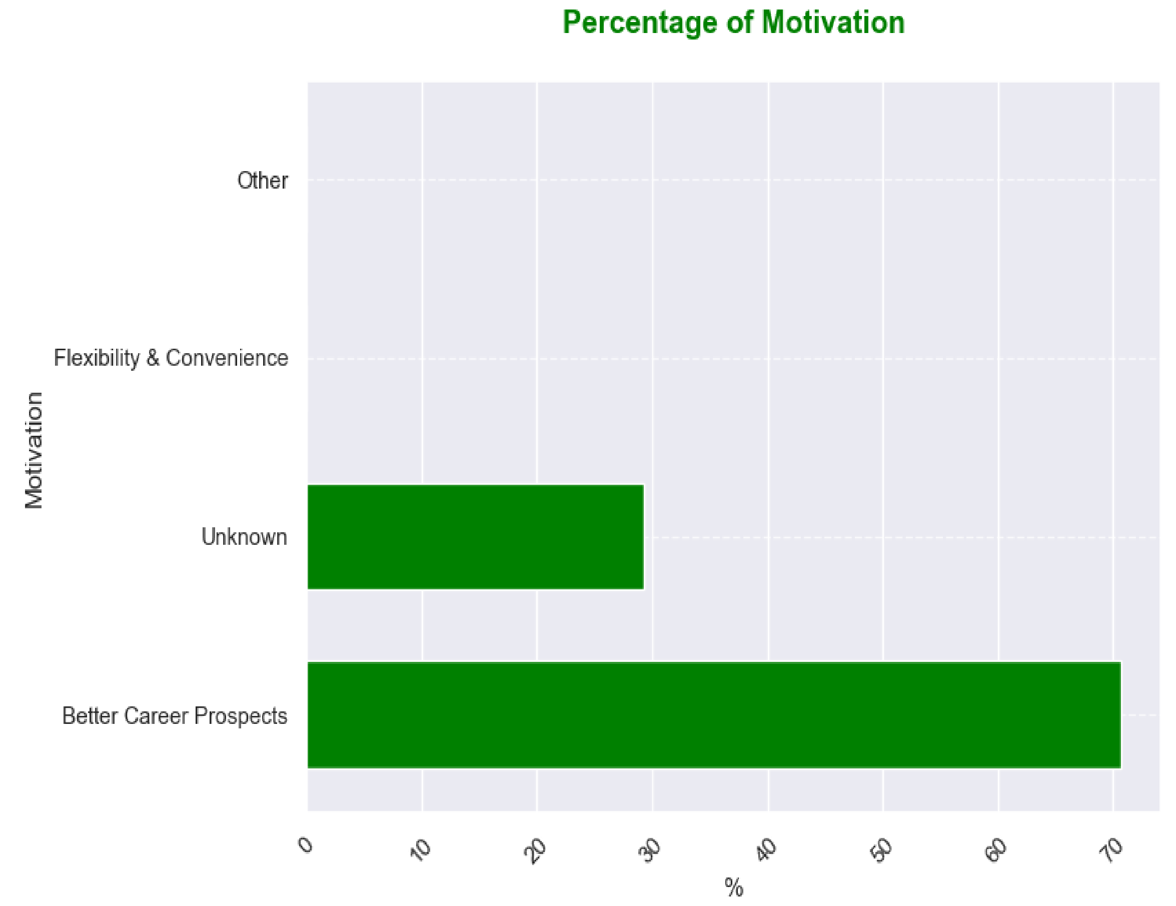


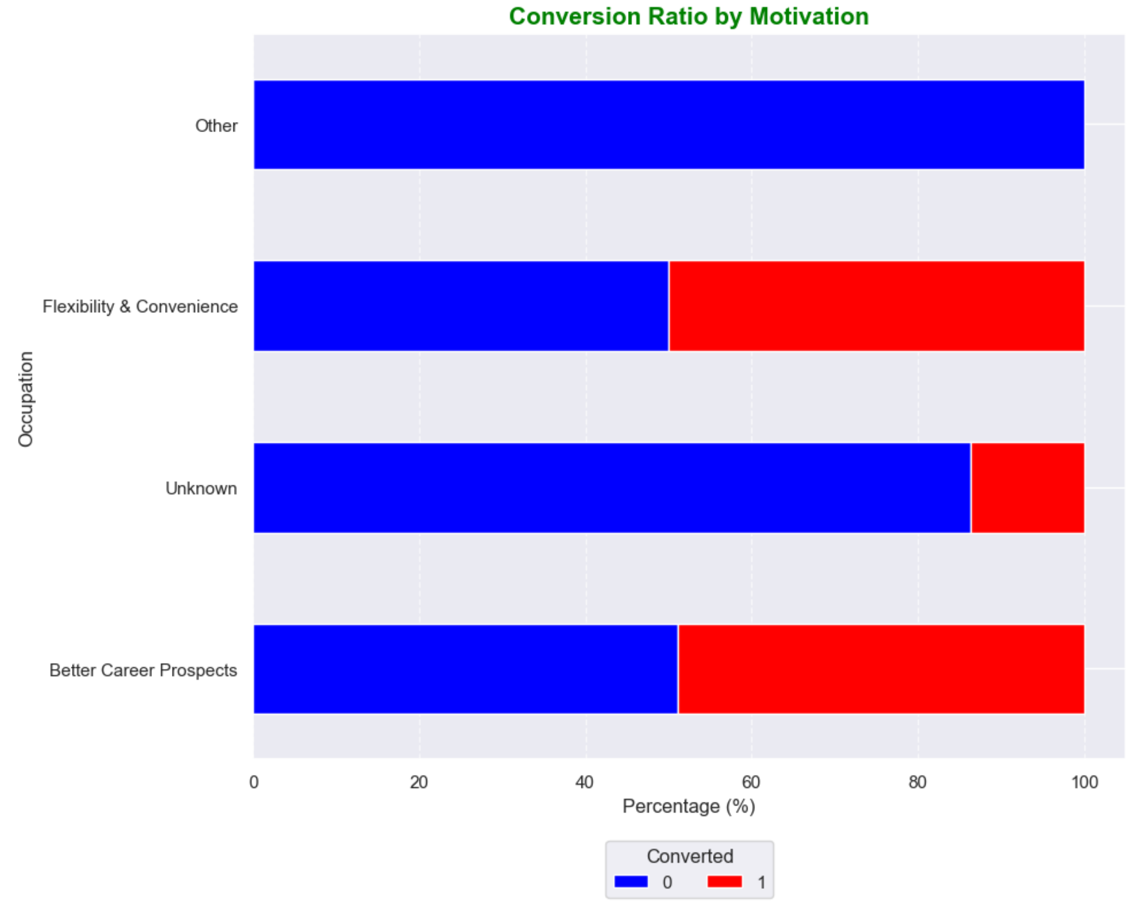
Occupations: 61% of students are unemployed, 29% of them dont provide information about current occupation. Meanwhile 8% are working professionals. The conversion rate among working professionals is very high (91%). Therefore, working professionals are more potential to focus to convert.



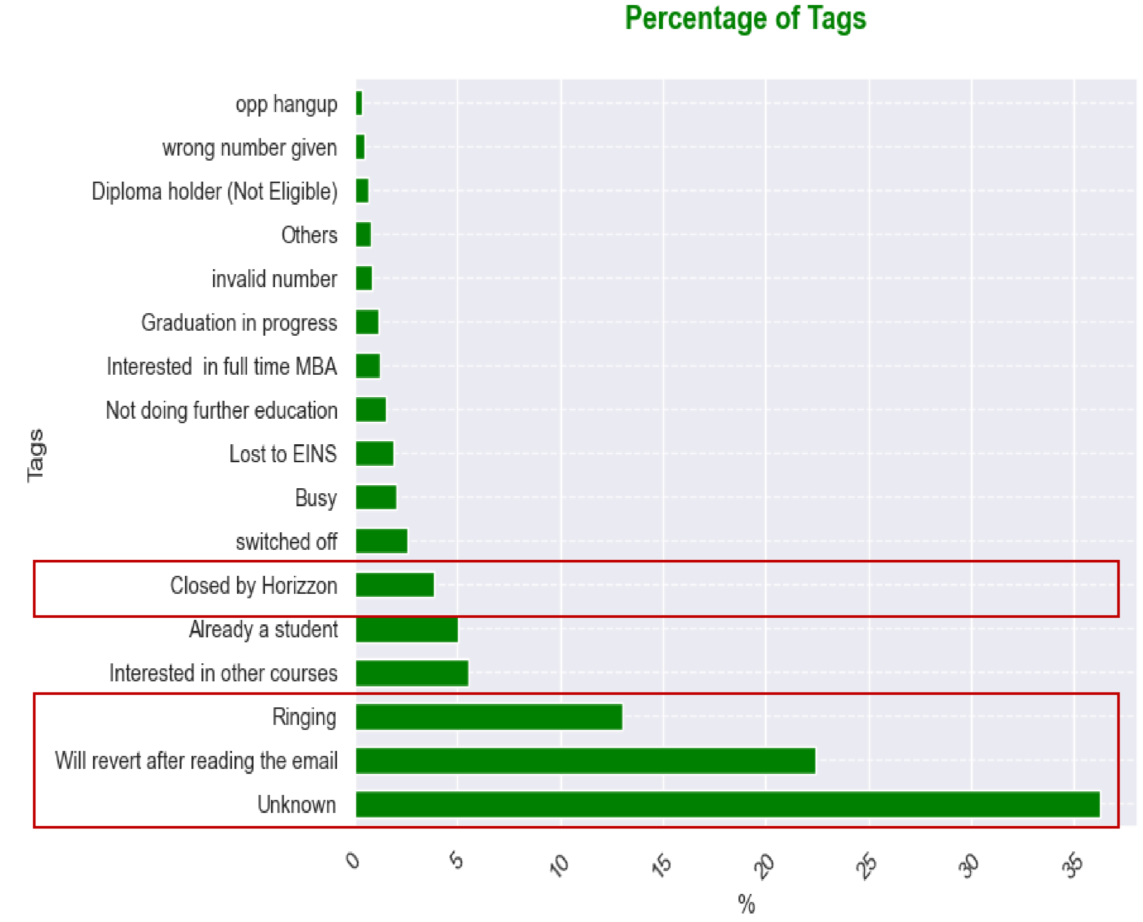


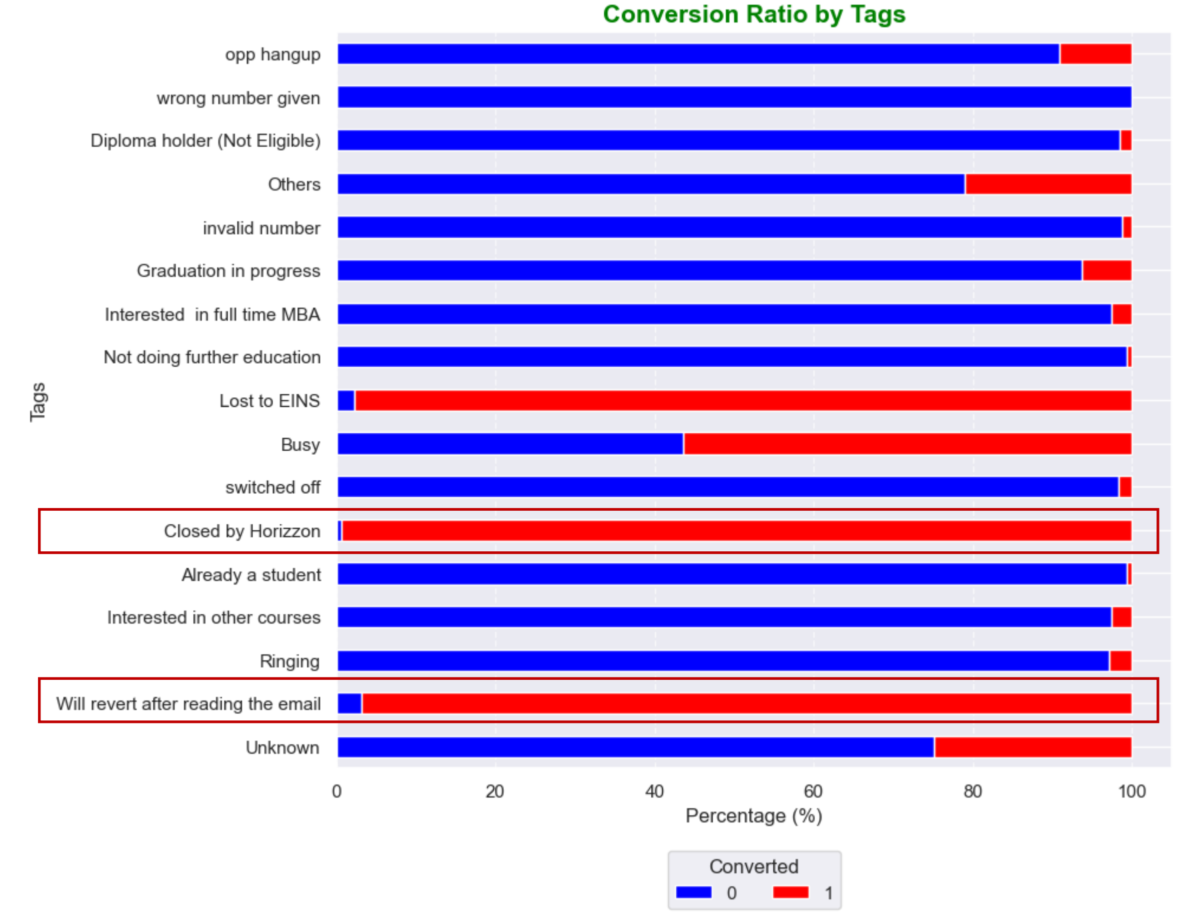
Motivations in choosing a course: 71% of students choose to study for better career prospect in the future while there are 29% not provide this information. The conversion ratio among those who disclosed their study motivation is much higher than among unknown group. Therefore, it can be seen that those who disclose this information show more serious intention to study rather than those who left this information as unknown.



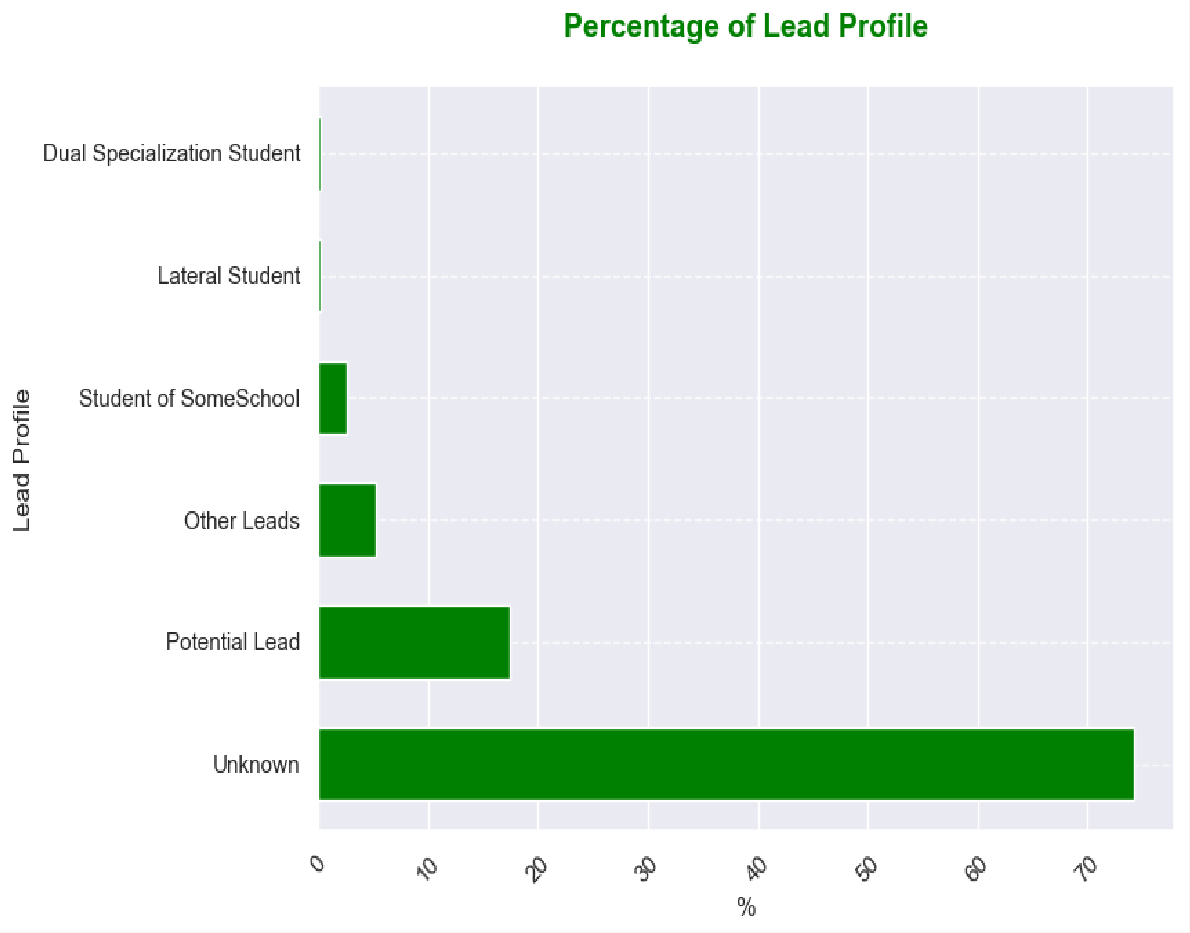


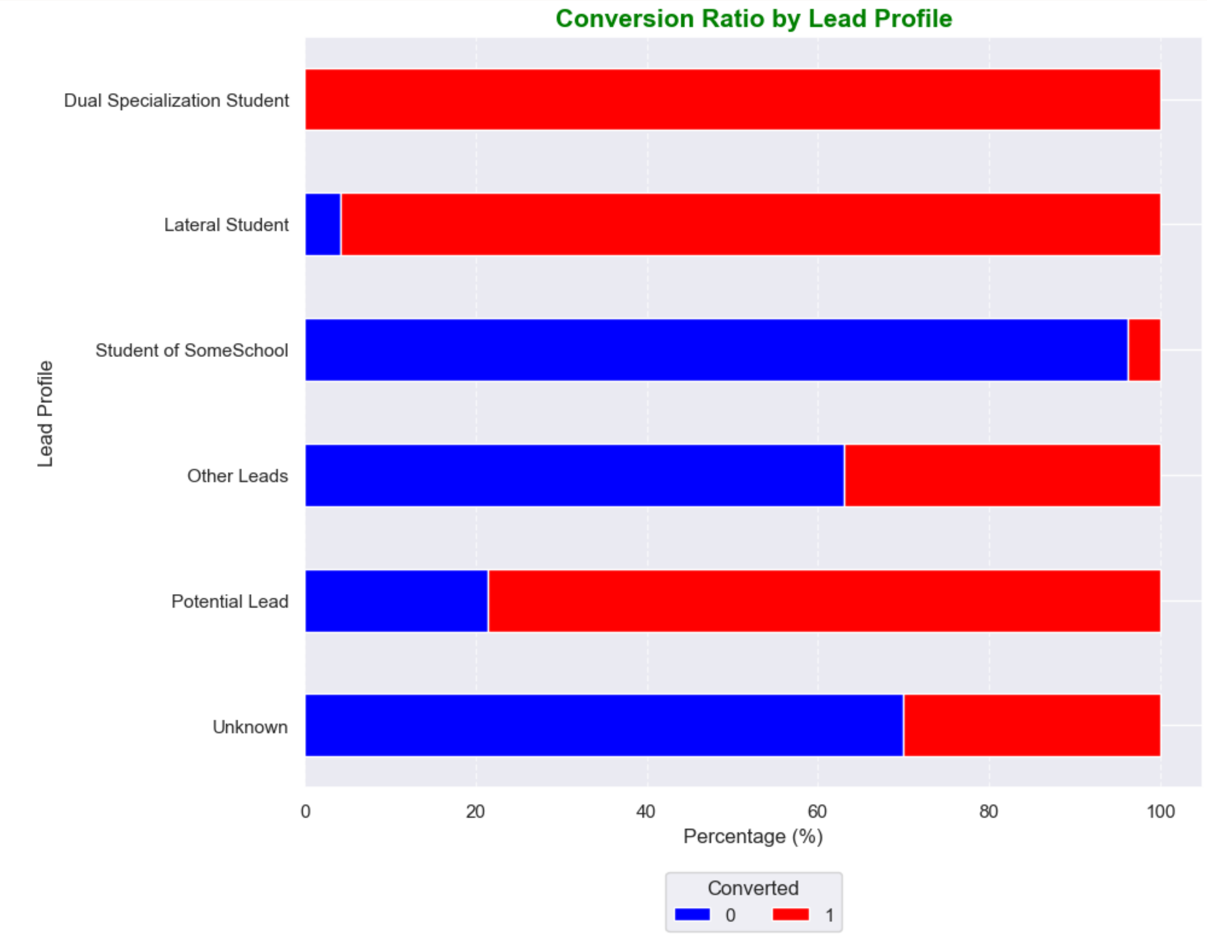
Tags: 36% of students refused to share this information while 22% are stated as "Will revert after reading the email". Also, the conversion rate among those who are at stage of "Will revert after reading the email" is really high (96%). This implies that "Will revert after reading the email" status is also very potential to focus while scanning the customer profile.



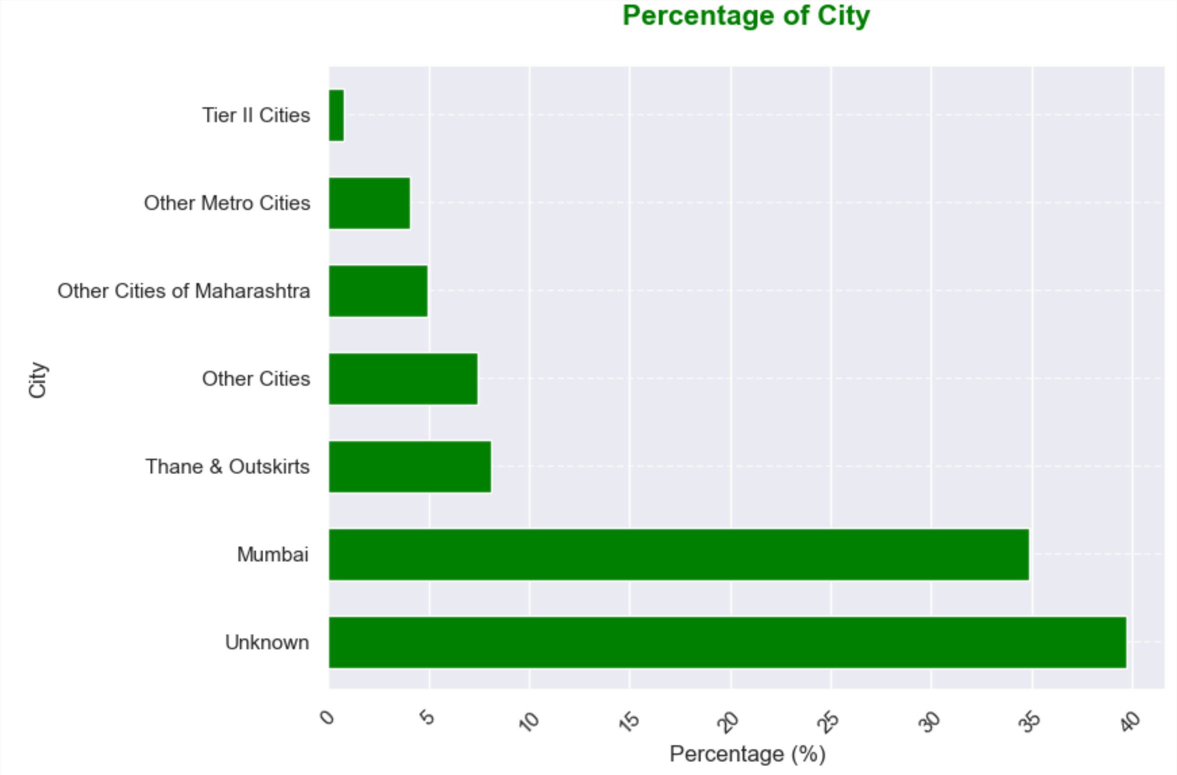


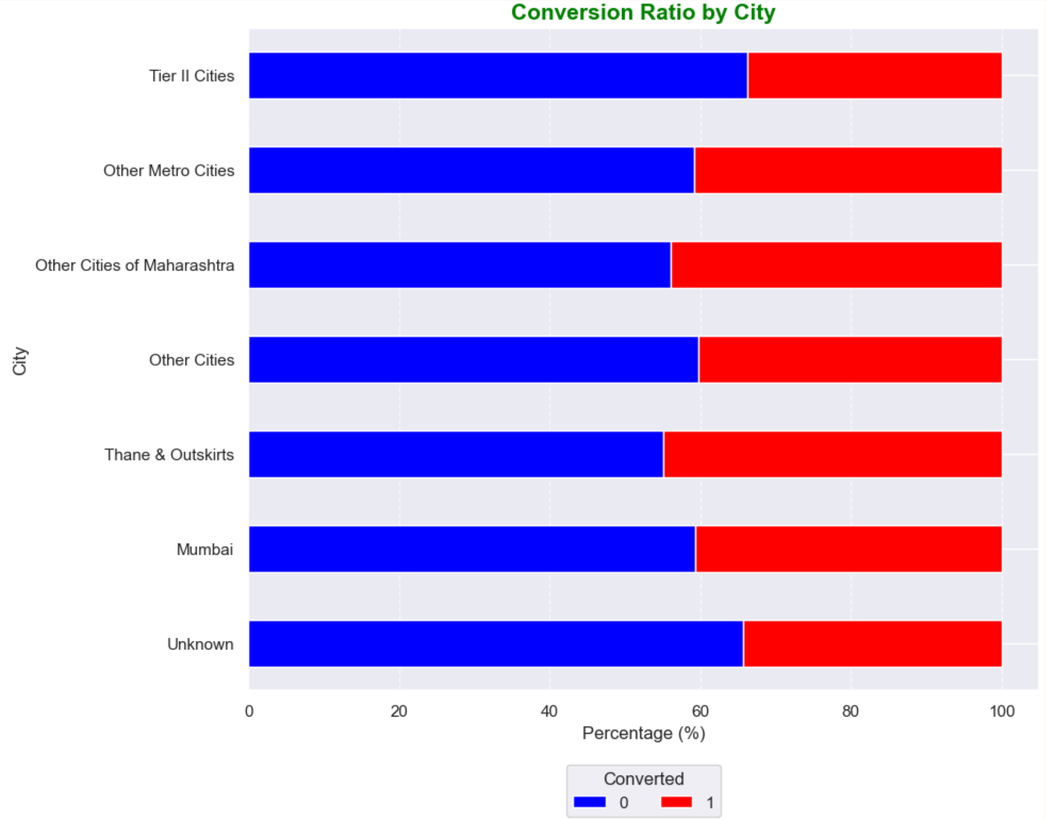
Lead Profile: 74% students have no information for this column, and among that only 30% being converted. 17% are potential lead with 78% are converted.





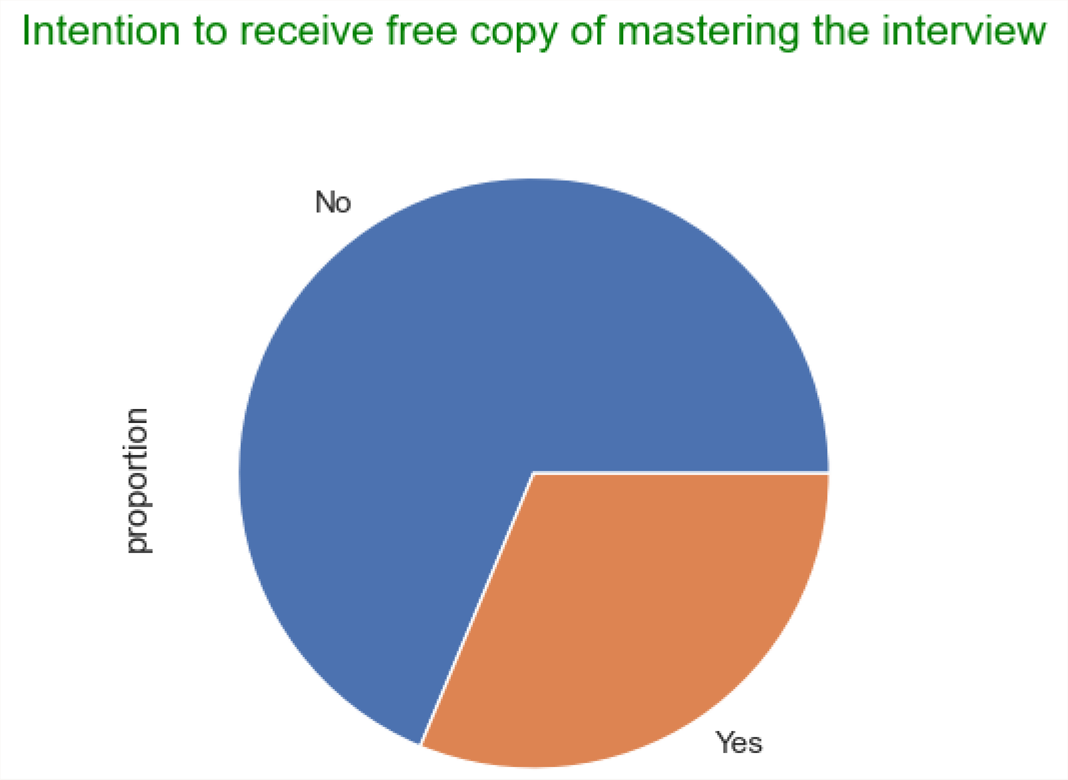
City: 40% of students did not provide information about city. 35% are from Mumbai. There is no significant difference in terms of conversion rate among different cities.

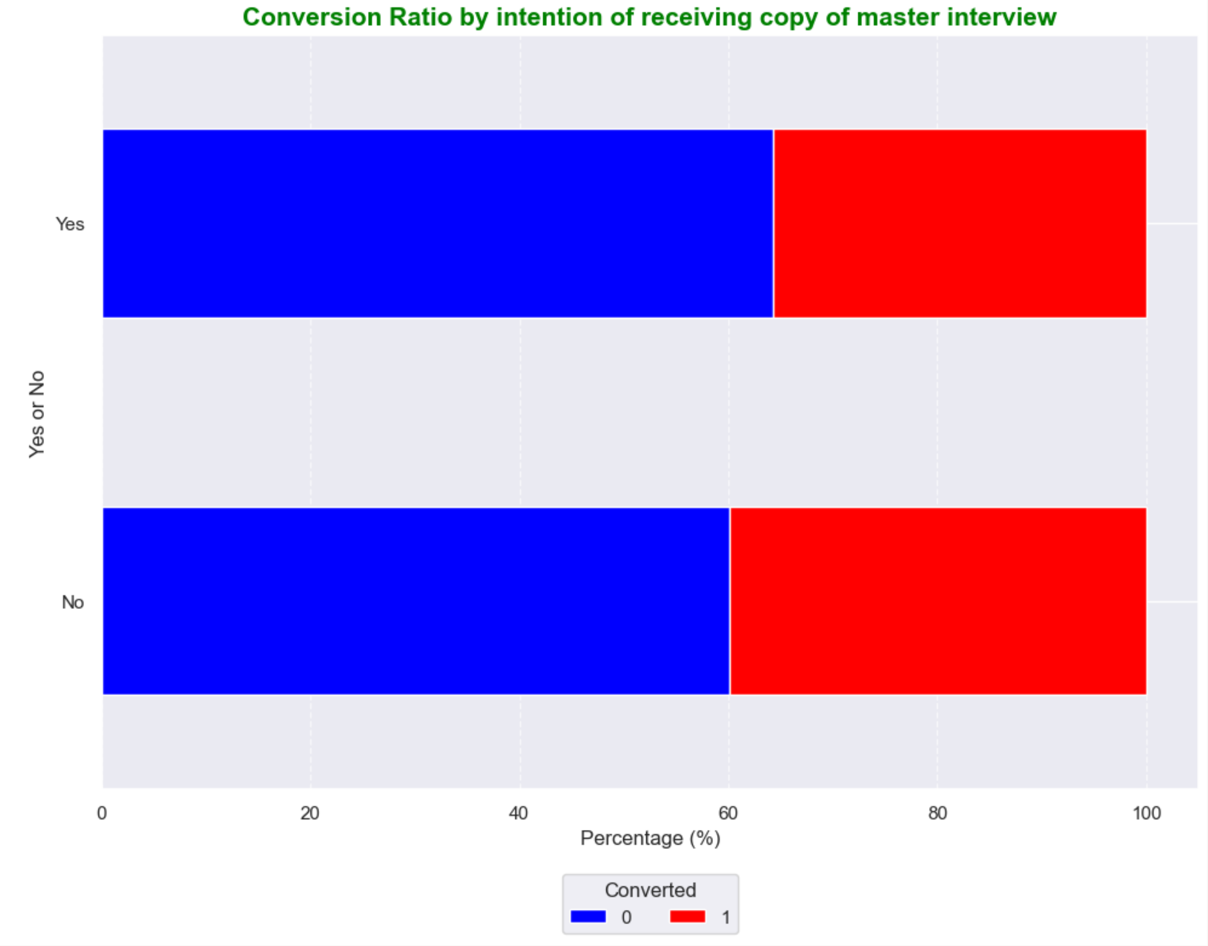




Free copy of mastering the interview

Only 31% of customers would like to receive a free copy of "Mastering the Interview".
There is no significant difference in terms of conversion rate between those 2 groups.

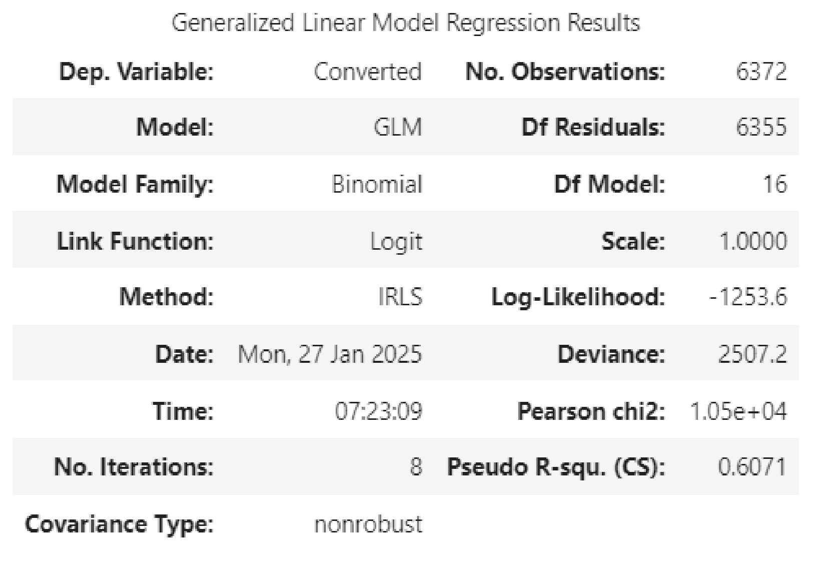




3. Model Building & Evaluation

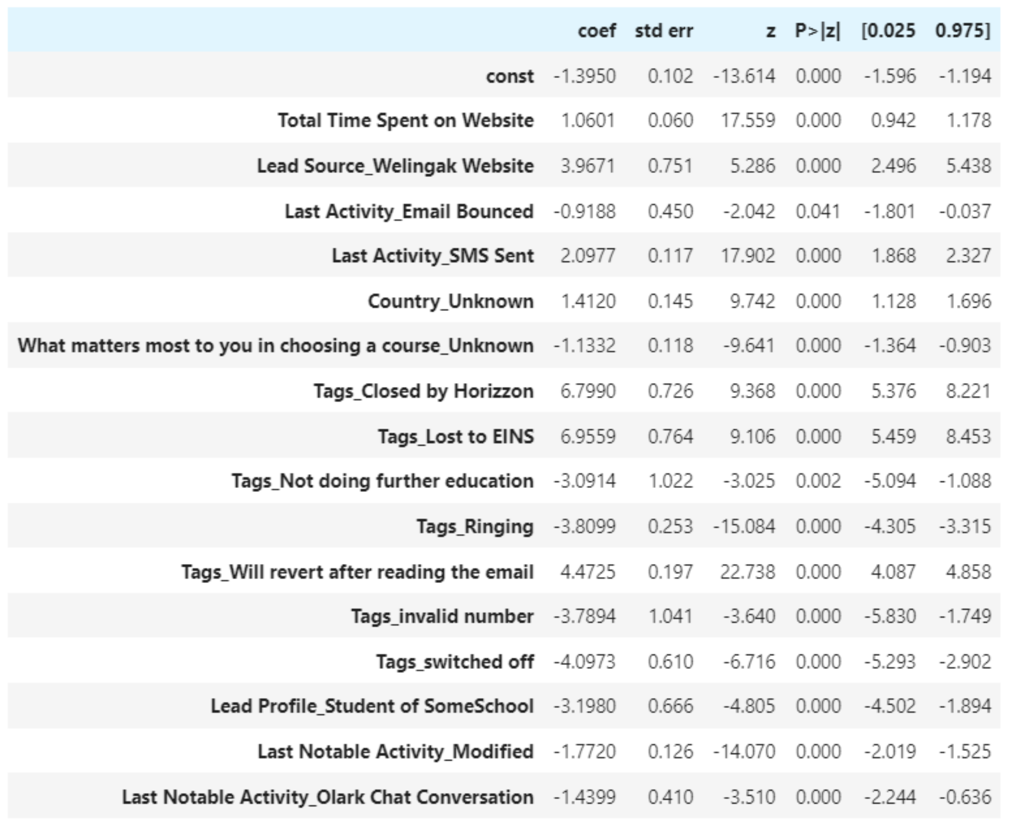
Model Building on Train Set

After multiple steps of RFE as well as checking p-values and VIF scores, a model is generated with 16 variables as below:

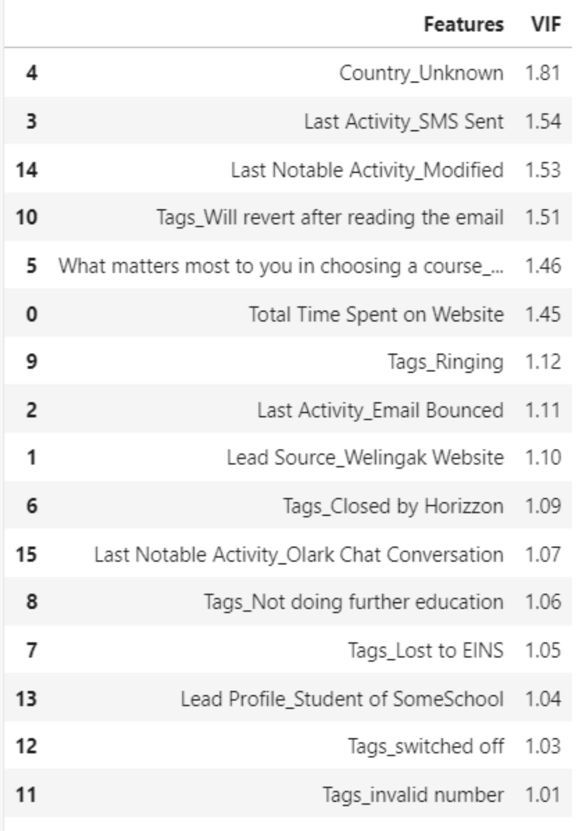


Generalized Linear Model Regression Results

Dep. Variable:	Converted	**No. Observations:**	6372
Model:	GLM	**Df Residuals:**	6355
Model Family:	Binomial	**Df Model:**	16
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-1253.6
Date:	Mon, 27 Jan 2025	**Deviance:**	2507.2
Time:	07:23:09	**Pearson chi2:**	1.05e+04
No. Iterations:	8	**Pseudo R-squ. (CS):**	0.6071
Covariance Type:	nonrobust		



	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3950	0.102	-13.614	0.000	-1.596	-1.194
Total Time Spent on Website	1.0601	0.060	17.559	0.000	0.942	1.178
Lead Source_Welingak Website	3.9671	0.751	5.286	0.000	2.496	5.438
Last Activity_Email Bounced	-0.9188	0.450	-2.042	0.041	-1.801	-0.037
Last Activity_SMS Sent	2.0977	0.117	17.902	0.000	1.868	2.327
Country_Unknown	1.4120	0.145	9.742	0.000	1.128	1.696
What matters most to you in choosing a course_Unknown	-1.1332	0.118	-9.641	0.000	-1.364	-0.903
Tags_Closed by Horizzon	6.7990	0.726	9.368	0.000	5.376	8.221
Tags_Lost to EINS	6.9559	0.764	9.106	0.000	5.459	8.453
Tags_Not doing further education	-3.0914	1.022	-3.025	0.002	-5.094	-1.088
Tags_Ringing	-3.8099	0.253	-15.084	0.000	-4.305	-3.315
Tags_Will revert after reading the email	4.4725	0.197	22.738	0.000	4.087	4.858
Tags_invalid number	-3.7894	1.041	-3.640	0.000	-5.830	-1.749
Tags_switched off	-4.0973	0.610	-6.716	0.000	-5.293	-2.902
Lead Profile_Student of SomeSchool	-3.1980	0.666	-4.805	0.000	-4.502	-1.894
Last Notable Activity_Modified	-1.7720	0.126	-14.070	0.000	-2.019	-1.525
Last Notable Activity_Olark Chat Conversation	-1.4399	0.410	-3.510	0.000	-2.244	-0.636



	Features	VIF
4	Country_Unknown	1.81
3	Last Activity_SMS Sent	1.54
14	Last Notable Activity_Modified	1.53
10	Tags_Will revert after reading the email	1.51
5	What matters most to you in choosing a course_...	1.46
0	Total Time Spent on Website	1.45
9	Tags_Ringing	1.12
2	Last Activity_Email Bounced	1.11
1	Lead Source_Welingak Website	1.10
6	Tags_Closed by Horizzon	1.09
15	Last Notable Activity_Olark Chat Conversation	1.07
8	Tags_Not doing further education	1.06
7	Tags_Lost to EINS	1.05
13	Lead Profile_Student of SomeSchool	1.04
12	Tags_switched off	1.03
11	Tags_invalid number	1.01

Model Evaluation on Train Set (1) - Cut off at 0.5

At the cut-off point for converted probability at 0.5, the confusion matrix is not so balance among true/false or negatives/positives. Moreover, though this model has high accuracy score, the gap between sensitivity and specificity scores are quite concerning. Therefore, we need to adjust the cut off point.

Confusion Matrix

Actual\ Predicted	Not_converted	Converted
Not_Converted	3781	172
Converted	308	2111

- True positives: 2111
- True negatives: 3781
- False positives: 172
- False negatives: 308

Accuracy Score: 92%

Sensitivity Score: 87%

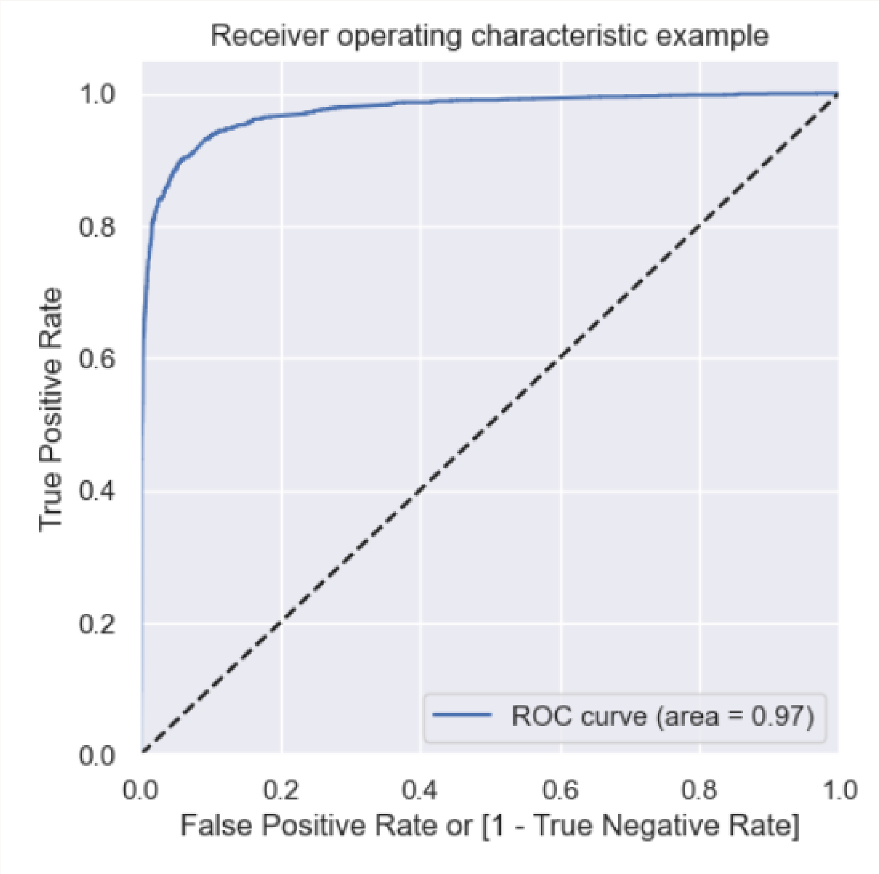
Specificity Score: 96%

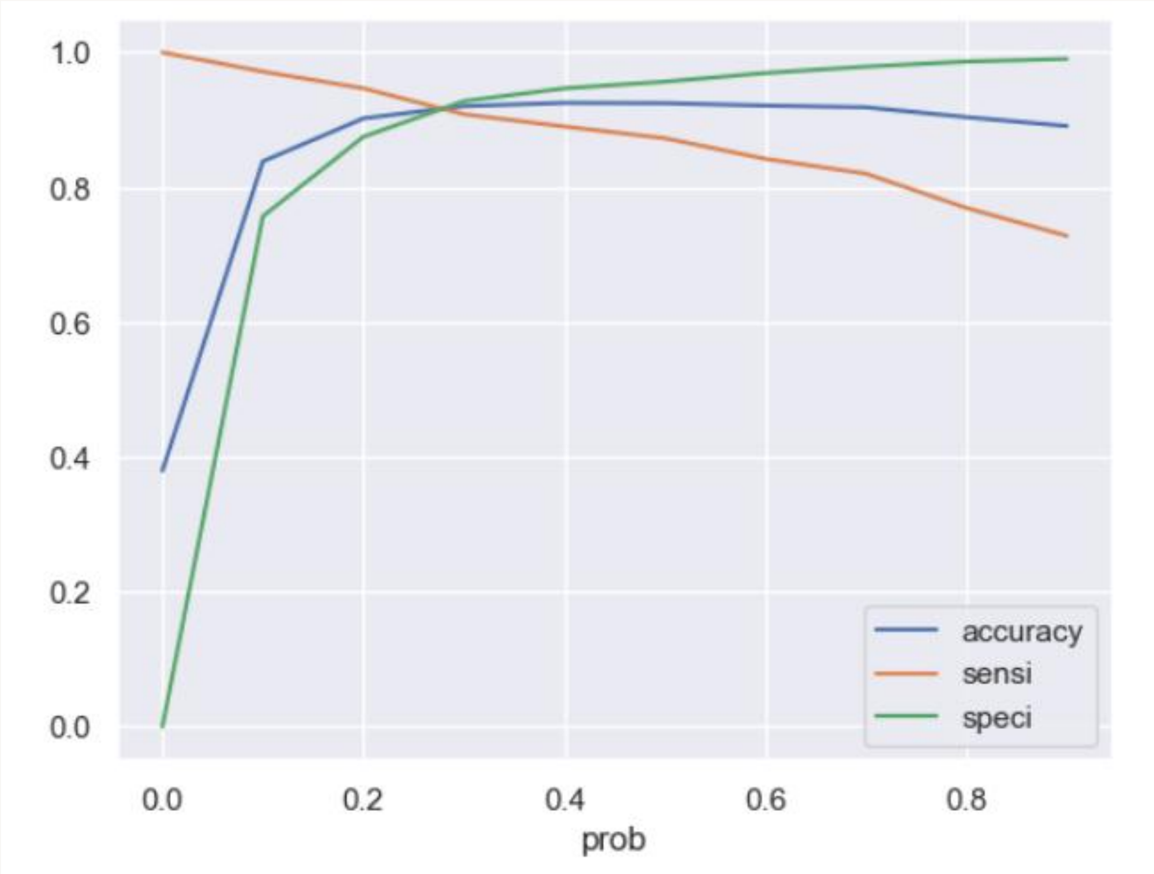
True Positive Rate: 87%

False Positive Rate: 4%

Model Evaluation on Train Set (1)

With the help of ROC curve as well as trade off chart of accuracy, sensitivity and specificity, optimal cut off point should be around 0.3.





Optimal cut off point: 0.3

Model Evaluation on Train Set (2) - Cut off at 0.3

At the cut-off point for converted probability at 0.3, the confusion matrix is more balancing among true/false or negatives/positives. Moreover, the accuracy, sensitivity and specificity scores are quite on similar. True Positive Rate has increased from 87% to 91%. Thus, this model is a good one to proceed further to testing phase.

Confusion Matrix

Actual\ Predicted	Not_converted	Converted
Not_Converted	3667	286
Converted	222	2197

- True positives: 2197
- True negatives: 3667
- False positives: 286
- False negatives: 222

Accuracy Score: 92%

Sensitivity Score: 91%

Specificity Score: 93%

True Positive Rate: 91%

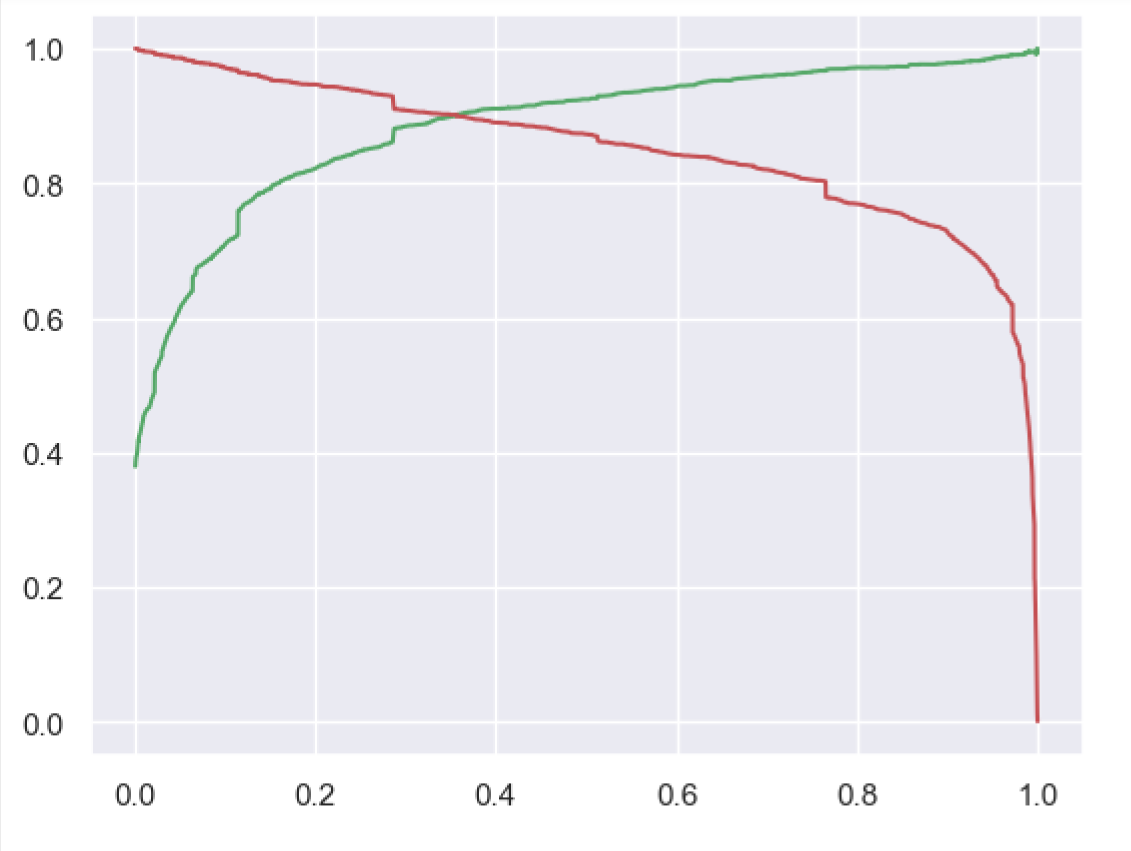
False Positive Rate: 7%

Model Evaluation on Train Set (2) - Cut off at 0.3

Precision:
92%

Recall:
87%

Precision & Recall Trade Off



Prediction on Test Set - Cut off at 0.3

At the cut-off point for converted probability at 0.3, the confusion matrix of test set is quite balancing among true/false or negatives/positives. Moreover, the accuracy, sensitivity and specificity scores are also close to one another at high level. True Positive Rate is still maintained 91%. Thus, this model is a good one to recommend to X Education.

Confusion Matrix

Actual\ Predicted	Not_converted	Converted
Not_Converted	1582	107
Converted	94	948

- True positives: 948
- True negatives: 1582
- False positives: 107
- False negatives: 94

Accuracy Score: 92%

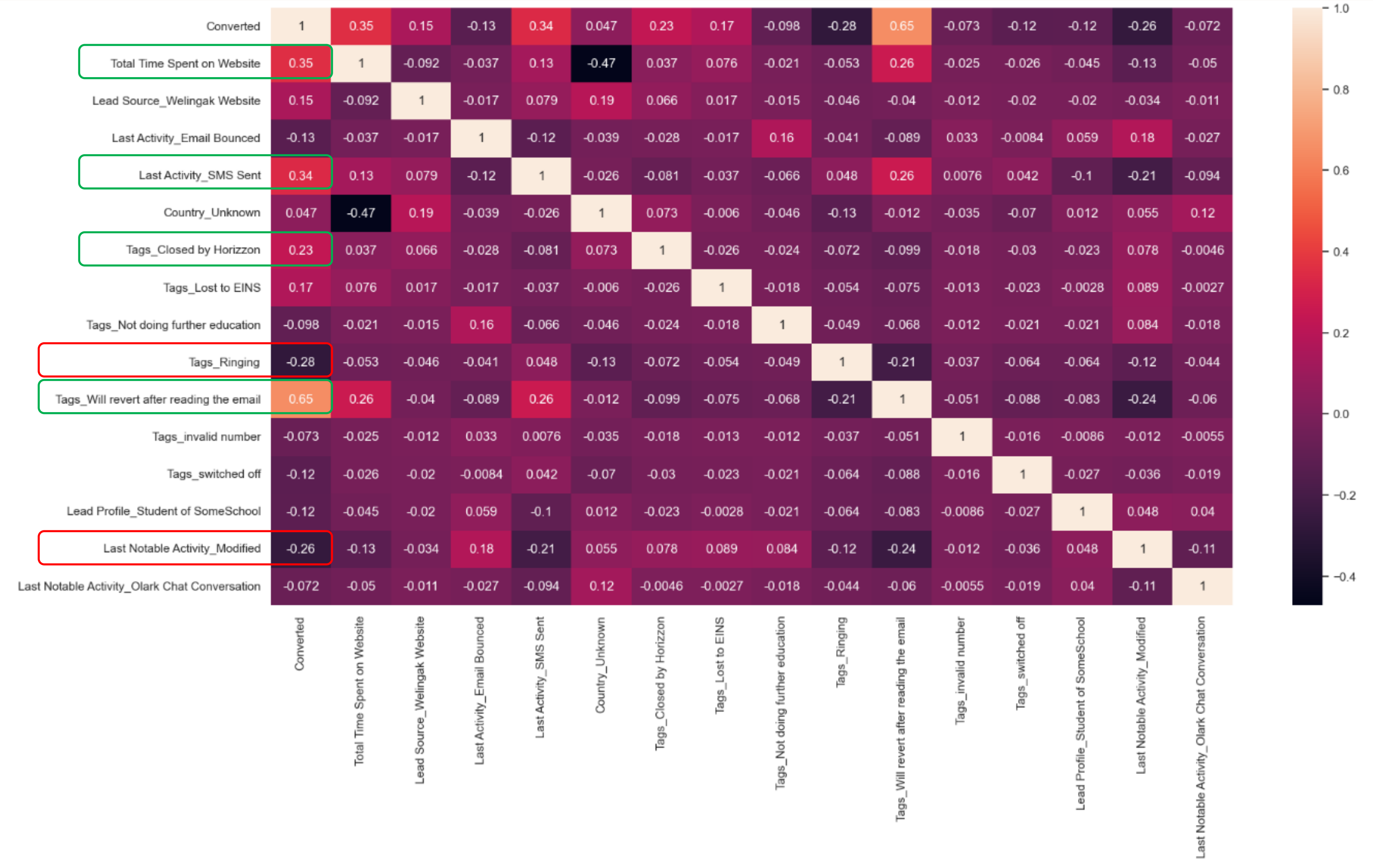
Sensitivity Score: 91%

Specificity Score: 94%

True Positive Rate: 91%

False Positive Rate: 6%

Correlation matrix among key variables with Converted



	Converted	Total Time Spent on Website	Lead Source_Welingak Website	Last Activity_Email Bounced	Last Activity_SMS Sent	Country_Unknown	Tags_Closed by Horizzon	Tags_Lost to EINS	Tags_Not doing further education	Tags_Ringing	Tags_Will revert after reading the email	Tags_invalid number	Tags_switched off	Lead Profile_Student of SomeSchool	Last Notable Activity_Modified	Last Notable Activity_Olark Chat Conversation
Converted	1	0.35	0.15	-0.13	0.34	0.047	0.23	0.17	-0.098	-0.28	0.65	-0.073	-0.12	-0.12	-0.26	-0.072
Total Time Spent on Website	0.35	1	-0.092	-0.037	0.13	-0.47	0.037	0.076	-0.021	-0.053	0.26	-0.025	-0.026	-0.045	-0.13	-0.05
Lead Source_Welingak Website	0.15	-0.092	1	-0.017	0.079	0.19	0.066	0.017	-0.015	-0.046	-0.04	-0.012	-0.02	-0.02	-0.034	-0.011
Last Activity_Email Bounced	-0.13	-0.037	-0.017	1	-0.12	-0.039	-0.028	-0.017	0.16	-0.041	-0.089	0.033	-0.0084	0.059	0.18	-0.027
Last Activity_SMS Sent	0.34	0.13	0.079	-0.12	1	-0.026	-0.081	-0.037	-0.066	0.048	0.26	0.0076	0.042	-0.1	-0.21	-0.094
Country_Unknown	0.047	-0.47	0.19	-0.039	-0.026	1	0.073	-0.006	-0.046	-0.13	-0.012	-0.035	-0.07	0.012	0.055	0.12
Tags_Closed by Horizzon	0.23	0.037	0.066	-0.028	-0.081	0.073	1	-0.026	-0.024	-0.072	-0.099	-0.018	-0.03	-0.023	0.078	-0.0046
Tags_Lost to EINS	0.17	0.076	0.017	-0.017	-0.037	-0.006	-0.026	1	-0.018	-0.054	-0.075	-0.013	-0.023	-0.0028	0.089	-0.0027
Tags_Not doing further education	-0.098	-0.021	-0.015	0.16	-0.066	-0.046	-0.024	-0.018	1	-0.049	-0.068	-0.012	-0.021	-0.021	0.084	-0.018
Tags_Ringing	-0.28	-0.053	-0.046	-0.041	0.048	-0.13	-0.072	-0.054	-0.049	1	-0.21	-0.037	-0.064	-0.064	-0.12	-0.044
Tags_Will revert after reading the email	0.65	0.26	-0.04	-0.089	0.26	-0.012	-0.099	-0.075	-0.068	-0.21	1	-0.051	-0.088	-0.083	-0.24	-0.06
Tags_invalid number	-0.073	-0.025	-0.012	0.033	0.0076	-0.035	-0.018	-0.013	-0.012	-0.037	-0.051	1	-0.016	-0.0086	-0.012	-0.0055
Tags_switched off	-0.12	-0.026	-0.02	-0.0084	0.042	-0.07	-0.03	-0.023	-0.021	-0.064	-0.088	-0.016	1	-0.027	-0.036	-0.019
Lead Profile_Student of SomeSchool	-0.12	-0.045	-0.02	0.059	-0.1	0.012	-0.023	-0.0028	-0.021	-0.064	-0.083	-0.0086	-0.027	1	0.048	0.04
Last Notable Activity_Modified	-0.26	-0.13	-0.034	0.18	-0.21	0.055	0.078	0.089	0.084	-0.12	-0.24	-0.012	-0.036	0.048	1	-0.11
Last Notable Activity_Olark Chat Conversation	-0.072	-0.05	-0.011	-0.027	-0.094	0.12	-0.0046	-0.0027	-0.018	-0.044	-0.06	-0.0055	-0.019	0.04	-0.11	1

Top variables impact **positively** on conversion rate:

- Tags_Will revert after reading the email: 0.65
- Total time spent on website: 0.35
- Last Activity_SMS Sent: 0.34
- Tags_Closed by Horizzon: 0.23

Top variables impact **negatively** on conversion rate:

- Tags_Ringging: -0.28
- Last Activity_Modified: -0.26

4. Conclusion and Recommendation

Conclusions & Recommendation (1)

Top most important variables that X Education company should pay more attention are:

- Tags_Will revert after reading the email: 0.65
- Total time spent on website: 0.35
- Last Activity_SMS Sent: 0.34
- Tags_Closed by Horizzon: 0.23

The first 3 factors impact the most positively to drive for the successful conversion which are from , “Total Time Spent On Website”, “Tags” and “Last Activity” columns. Therefore, when collecting the data, we need to pay more attention to improve the quality of inputs for these 3 columns by:

1. Tags: There are 35% of tags data is unknown, meaning missing values, we need to minimize the missing values of this column by set a compulsory rule for this filter.
2. Last Activity_SMS Sent: similarly to the case of ‘Tags_Close by Horizzon’, we need to explore further to trigger customers to SMS Sent Step.
3. Total Time Spent On Website: is moderately important, therefore, we need to improve our information quality on our website to trigger customers browsing information more often, the more time they spent on website, the higher chance they will be converted. Based on exploratory data from this column, the group of customers that spend more than 3 hours on the website have significantly higher chance to be converted.

‘Close by Horizzon’ though being moderately important, the percentage of this group is very small (~5%) with nearly 100% is converted, therefore, we need to explore further into this group to expand it further.

Conclusions & Recommendation (2)

X Education has a period of 2 months every year during which they hire some interns. The sales team, in particular, has around 10 interns allotted to them. So during this phase, they wish to make the lead conversion more aggressive. So they want almost all of the potential leads (i.e. the customers who have been predicted as 1 by the model) to be converted and hence, want to make phone calls to as much of such people as possible. After running logistic regression model, we have found a good model with high accuracy rate (92%), sensitivity (93%) and specificity (91%) to classify whether a lead is a potential one or not. Based on this model, we can generate a good list of potential leads for sales team to maximize the conversion rate with the cut-off point of converted probability at 0.3.

Similarly, at times, the company reaches its target for a quarter before the deadline. During this time, the company wants the sales team to focus on some new work as well. So during this time, the company's aim is to not make phone calls unless it's extremely necessary, i.e. they want to minimize the rate of useless phone calls.

Thus, in order to minimize the rate of useless phone calls, sales team also should not waste time to those customers who have status marked as below variables because these variables impact negatively to the conversion which includes:

1. Tags_Ringging: -0.28
2. Last Activity_Modified: -0.26

Thank you

Truong Tuyet Lam

Lam.truongtuyet@gmail.com