Lead Scoring Case Study

Submitted by
Jayaram Balakrishnan
Neha Manjrekar
Mohankumar Selvaraj

Problem Statement and Business Goals

- 'X' Education sells online courses to industry professionals. 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.
- X Education needs help in selecting 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 between 0 and 100 to each of the leads, such that the customers with a higher lead score ('Hot Leads') 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%.

Methodology

Data Sourcing, Cleaning, and Preparation

- Read the Data from the Source
- Convert data into a clean format suitable for analysis
- Remove duplicate data
- Outlier Treatment
- Exploratory Data Analysis
- Feature Standardization.

Feature Scaling and Splitting

- Creating dummy variables
- Feature Scaling of Numeric data
- Splitting data into train and test set

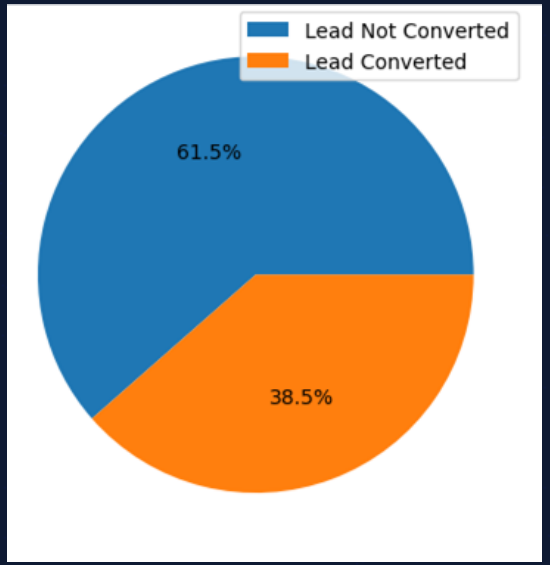
Model Building

- Initial Feature Selection using RFE
- Determine the optimal model using Logistic Regression
- Calculate various metrics like accuracy, sensitivity, specificity, precision, and recall and evaluate the model.

Result

- Determine the lead scores and check if the target final predictions amount to an 80% conversion rate.
- Evaluate the final prediction on the test set using cut off threshold from sensitivity and specificity metrics
- Make actionable recommendations based on the results.

Dataset Properties



1. Dataset contains 9240 rows and 37 columns.

2. High null count (~45%) in 'Asymmetric' columns and 51% in 'Lead Quality'; these columns are to be dropped.

3. No duplicate rows are present in the dataset.

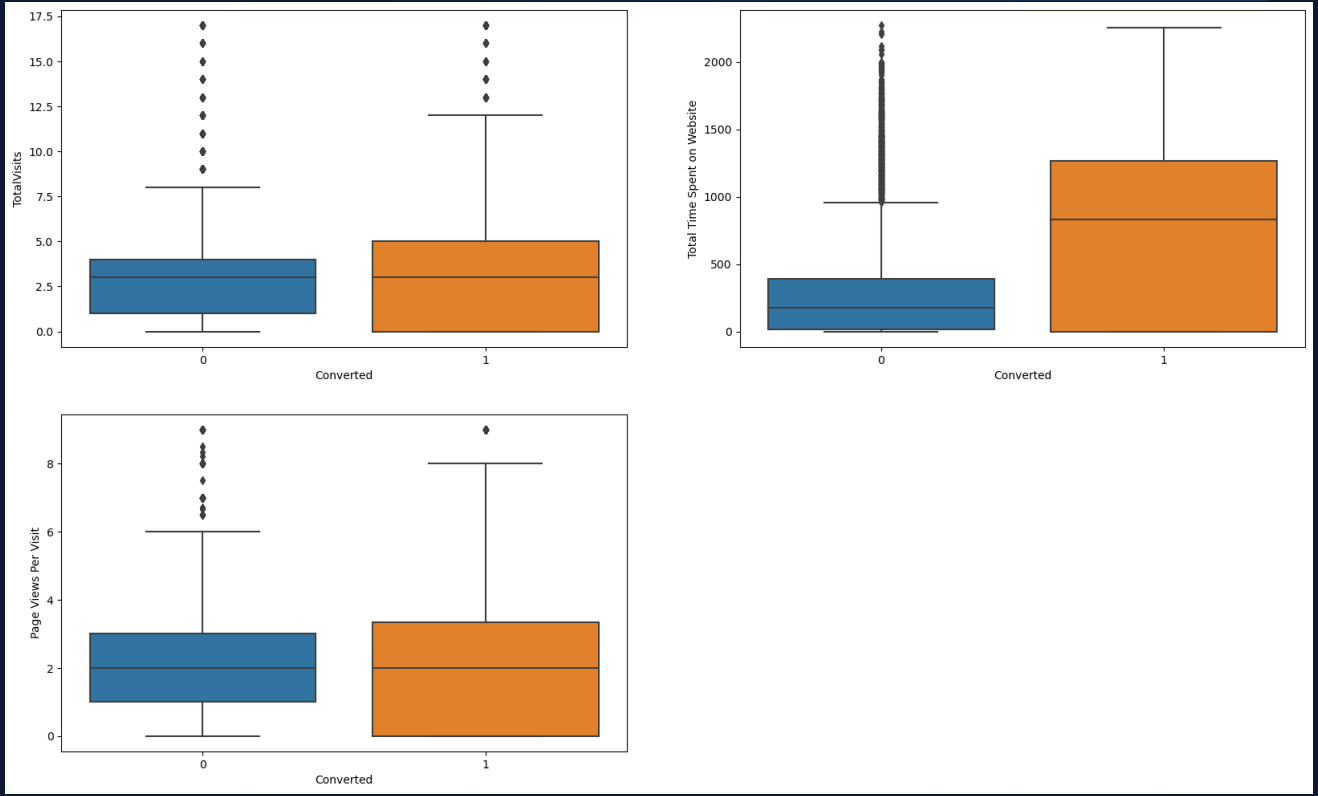
4. Data imbalance: 61.5% skewed towards the 'Lead Not Converted' category. 'Yes/No' categorical variables need binary conversion for modelling.

6. Columns like 'Receive More Updates About Our Courses,' 'Update me on Supply Chain Content,' etc., with only 'No' values offer no contribution, suggested for removal.

7. Summary emphasizes key dataset attributes, nulls, imbalance, and categorical conversions for effective preprocessing and analysis.

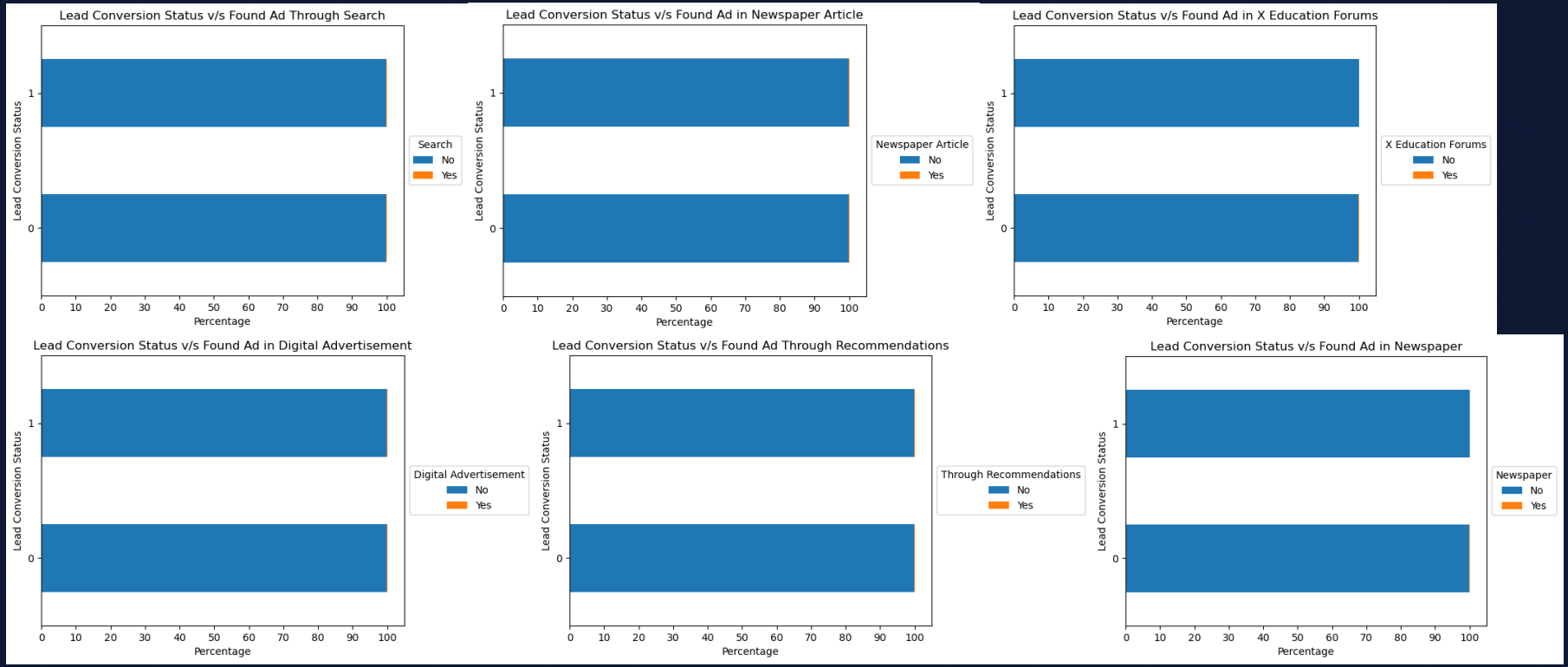
Exploratory Data Analysis

Analysing the Numerical Variables



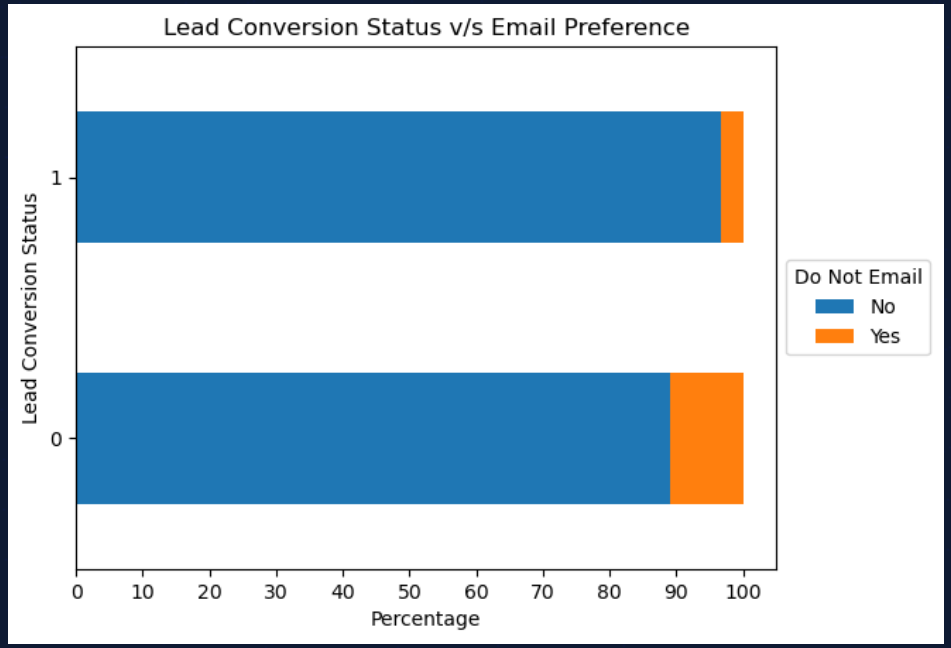
1. Median total visits and median page views per visit are more or less the same for both converted and non-converted leads.
2. Total time spent on site is significantly higher for the converted leads.

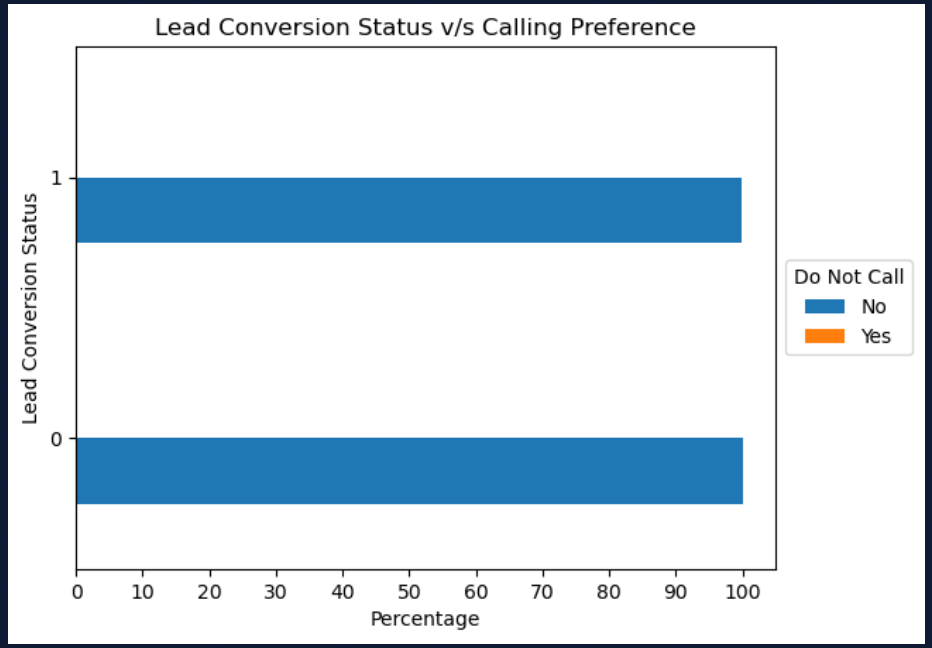
Analysing Various Sources of Lead Generation



A vast majority (~99%) did not receive the marketing information from these sources for both lead categories. 'Yes' values are miniscule in number.

Call and E-Mail Preferences

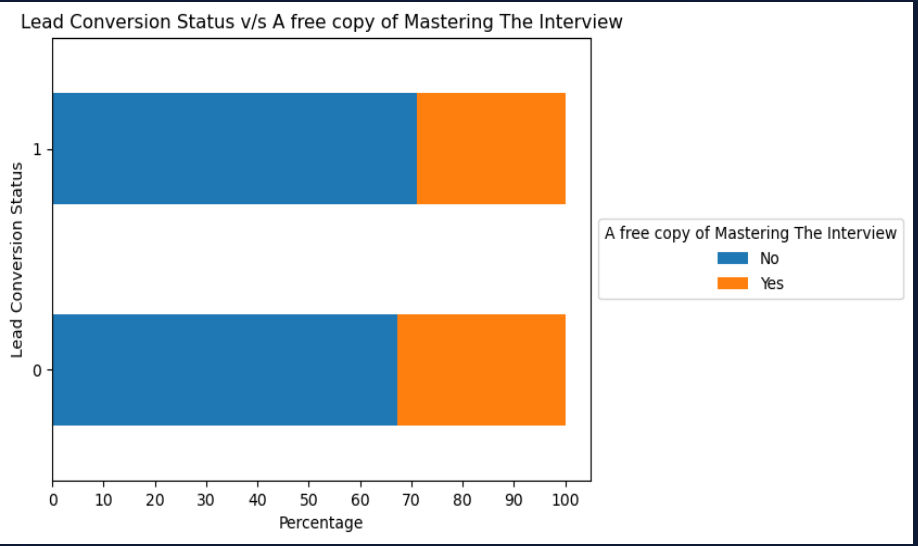




1. Barring a small percentage, majority of the leads have opted to not receive mails. This number is higher among the converted leads (~96%) than the non-converted leads (89%).

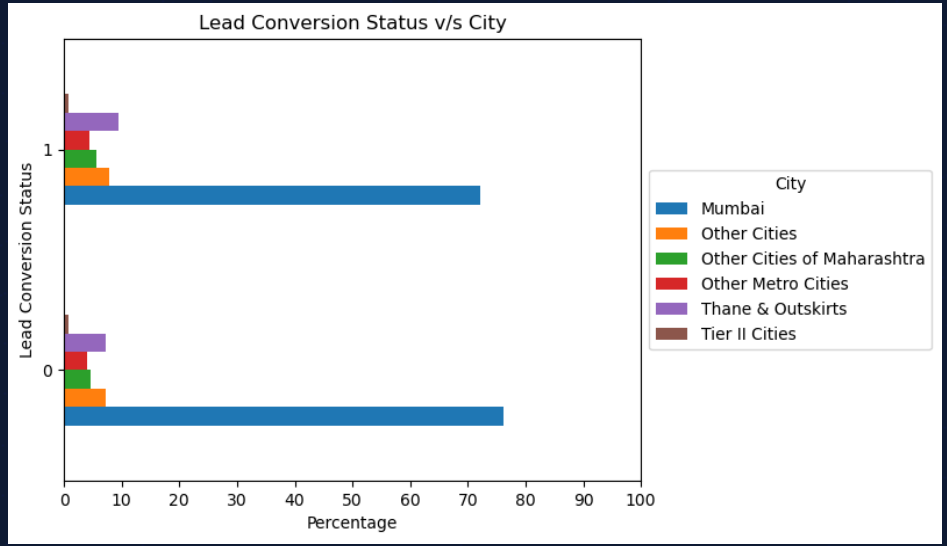
2. 2. Similarly, barring a miniscule percentage (~0.06%), majority of the leads prefer not to be called. None of the converted leads have opted to receive calls.

A free copy of Mastering the Interview



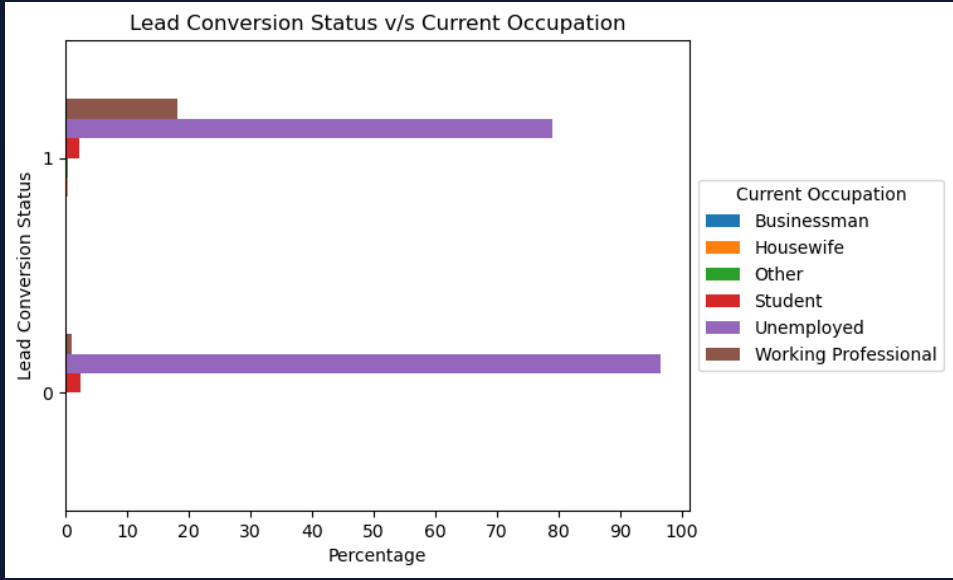
Majority (around 70% in both lead categories) have opted 'No'. There is no significant difference in the value distribution between both lead categories, which suggests that this is not an influencing parameter.

City



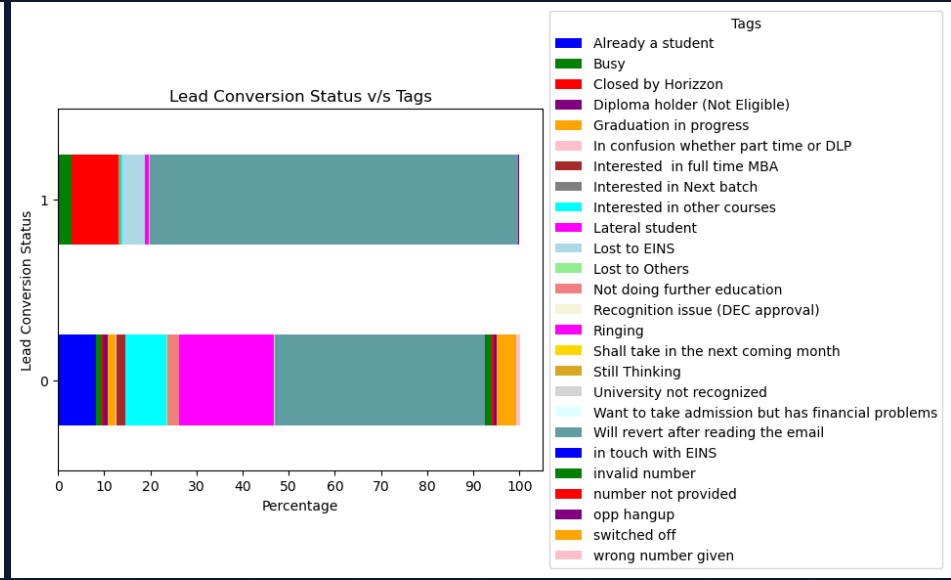
Mumbai is the dominant city, by a huge margin.

Current Occupation



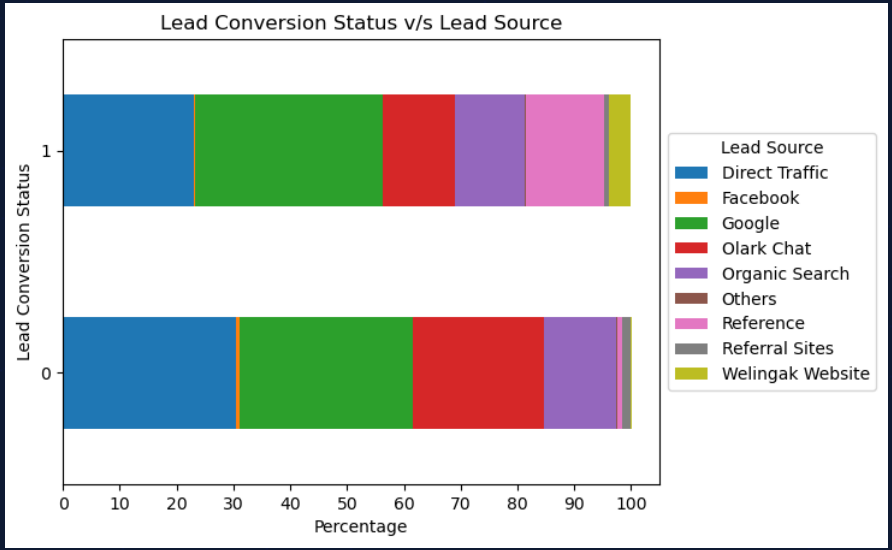
Unemployed is the dominant category, by a huge margin. Working professional is a distant second.

Tags



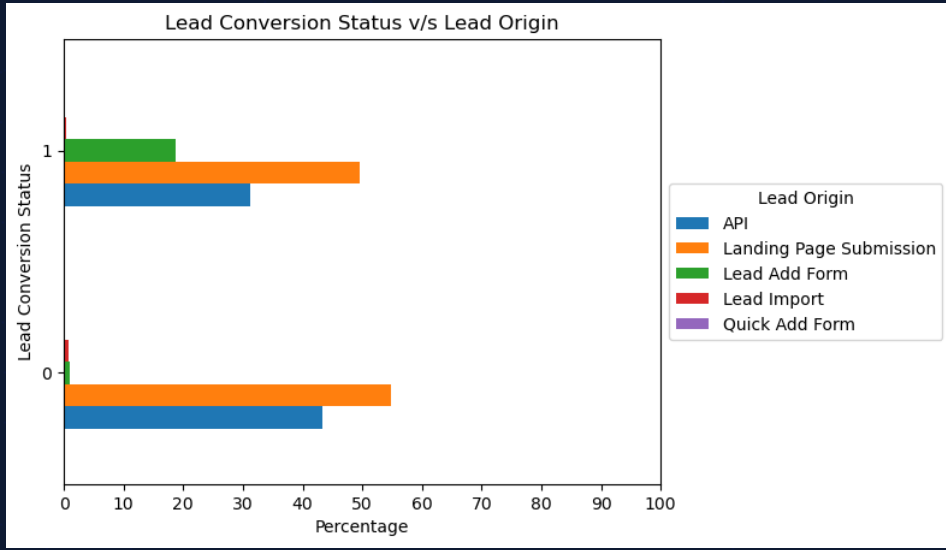
This variable represents the tags assigned to customers indicating the current status of the lead. As such, this is not a variable that will have an influence on the lead conversion status. Hence, this variable can be dropped.

Lead Source



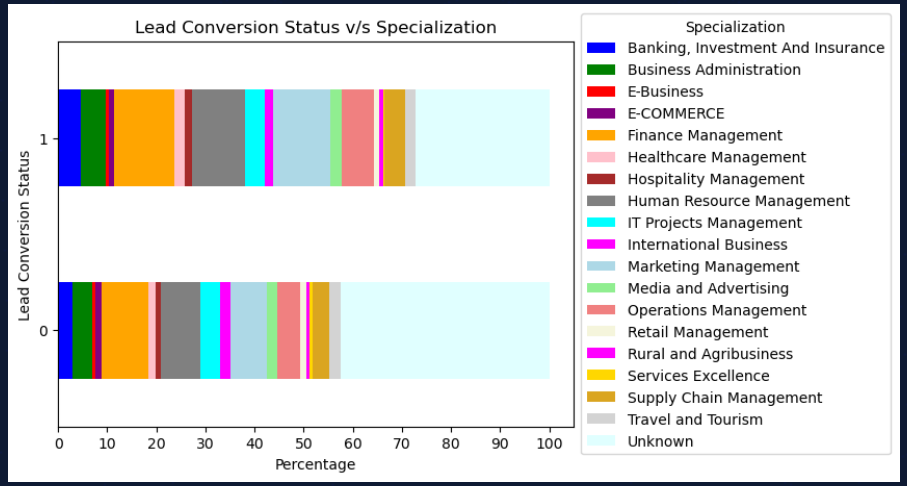
Google and Direct traffic are the top lead sources. For the converted leads, Google is the top lead source, followed by Direct Traffic and Reference.

Lead Origin



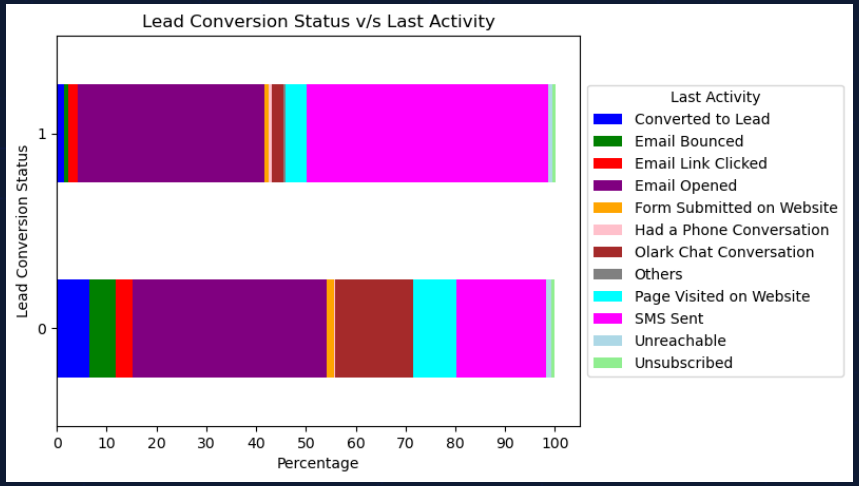
Landing Page Submission is the dominant category, followed by API and Lead Add Form.

Specialization



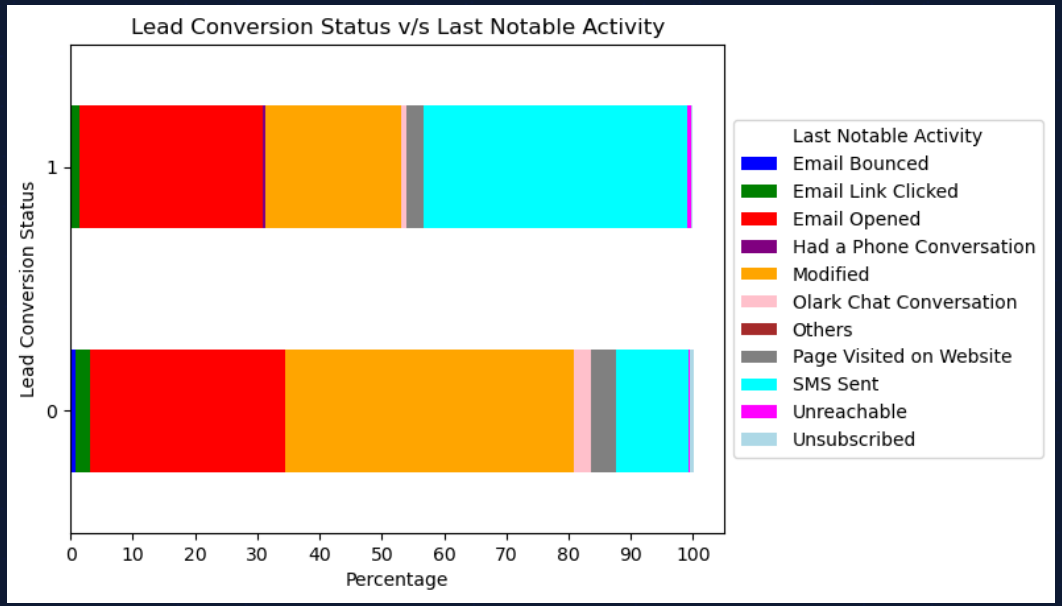
Converted leads have less unknown values for specialization. This suggests that the converted leads provide proper information as they are actually interested in pursuing the course.

Last Activity



'SMS Sent' is the top last activity among the converted leads, followed by 'Email Opened' and 'Page Visited on Website'.

Last Notable Activity



'SMS Sent' is the top last notable activity among the converted leads, followed by 'Email Opened' and 'Modified'.

In this case, we will take a call to drop 'Last Activity' and retain 'Last Notable Activity' due to the below reasons:

1.'Last Notable Activity' had no missing values, and hence we can trust the data in this column more, as the data has more "integrity".

2. For 'Last Activity', there were missing values which were imputed with the mode; although an accepted method, this may not be entirely accurate.

Insights from Exploratory Data Analysis

1. Barring a small percentage, the majority of the leads have opted to not receive mails. This number is higher among the converted leads (~96%) than the non-converted leads (89%).

2. Similarly, barring a minuscule percentage (~0.06%), the majority of the leads prefer not to be called. None of the converted leads have opted to receive calls.

3. Barring a miniscule percentage (~0.1%), the majority of the leads did not find the advertisement through search. This proportion is more or less consistent for both converted and non-converted leads. The situation is very similar for finding the advertisement through newspaper articles, digital advertisements, X Education Forums, and recommendations.

4. 71% of the converted leads opted to receive a free copy of 'Mastering The Interview'. For the non-converted leads, this number stands at 67%. This small difference suggests that this is not really a parameter that is going to influence the lead conversion.

5. Mumbai is the top city when it comes to converted leads. 72% of the converted leads are from Mumbai. However, Mumbai is similarly higher in terms of the non-converted leads as well (76%). This suggests that 'X Education' has primarily conducted its marketing campaign in Mumbai, which is why the number is higher for this city.

6. Among both the converted and non-converted leads, unemployed people have the highest percentage (79% and 96% respectively). This again suggests that the company had mainly targeted its marketing campaign towards unemployed people. Only 18% of the converted leads are Working professionals

7.Google is the top lead source (>30%) for both converted and non-converted leads.

Insights from Exploratory Data Analysis

8.'Landing Page Submission' is the top lead origin (~50%) for both converted and non-converted leads, followed by API.

9.Apart from the leads for which the specialisations are unknown, Finance Management, Human Resource Management and Marketing Management (in that order) are the top specialisations among both the converted and non-converted leads.

10.The most common last activity among the converted leads is 'SMS Sent' (48%), and the most common last activity among the non-converted leads is 'Email Opened' (~37%). As the data in this variable is too similar to that in the variable 'Last Notable Activity', this variable has been dropped to avoid redundant data.

11. The most common last notable activity among the converted leads is 'SMS Sent' (42%), and the most common last notable activity among the non-converted leads is 'Modified' (~46%). This variable seems to be typical (correlated) to the 'Last Activity' variable, and this needs to be investigated further.

12. The most dominating tag for both lead categories is 'Will revert after reading the email'. However, this is the tag for 80% of the converted leads, and for 45% of the non-converted leads. So, this means, most of the converted leads actually reverted expressing their interest after reading the mail. This variable represents the tags assigned to customers indicating the current status of the lead. As such, this is not a variable that will have an influence on the lead conversion status, as this represents the tags assigned to customers indicating the current status of the lead. Hence, we may drop this variable. . As such, this is not a variable that will have an influence on the lead conversion status.

Model Building

Model Building

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.3019	0.149	-8.740	0.000	-1.594	-1.010
Do Not Email	-1.2080	0.167	-7.246	0.000	-1.535	-0.881
TotalVisits	0.7698	0.208	3.692	0.000	0.361	1.178
Total Time Spent on Website	4.4612	0.164	27.165	0.000	4.139	4.783
Lead Origin_Landing Page Submission	-1.1188	0.127	-8.795	0.000	-1.368	-0.869
Lead Origin_Lead Add Form	3.4373	0.210	16.358	0.000	3.025	3.849
Lead Source_Olark Chat	1.0246	0.126	8.101	0.000	0.777	1.272
Lead Source_Welingak Website	2.5380	0.749	3.388	0.001	1.070	4.006
Specialization_Hospitality Management	-1.0558	0.332	-3.182	0.001	-1.706	-0.406
Specialization_Unknown	-1.1571	0.123	-9.440	0.000	-1.397	-0.917
Current Occupation_Working Professional	2.6225	0.192	13.658	0.000	2.246	2.999
Last Notable Activity_Modified	-0.6751	0.084	-8.031	0.000	-0.840	-0.510
Last Notable Activity_Olark Chat Conversation	-1.2319	0.343	-3.590	0.000	-1.904	-0.559
Last Notable Activity_SMS Sent	1.3965	0.085	16.410	0.000	1.230	1.563

	Features	VIF
3	Lead Origin_Landing Page Submission	3.05
1	TotalVisits	2.54
8	Specialization_Unknown	2.52
2	Total Time Spent on Website	2.08
10	Last Notable Activity_Modified	1.94
5	Lead Source_Olark Chat	1.87
12	Last Notable Activity_SMS Sent	1.64
4	Lead Origin_Lead Add Form	1.46
6	Lead Source_Welingak Website	1.27
9	Current Occupation_Working Professional	1.21
0	Do Not Email	1.12
11	Last Notable Activity_Olark Chat Conversation	1.09
7	Specialization_Hospitality Management	1.02

Generalized Linear Model Regression Results

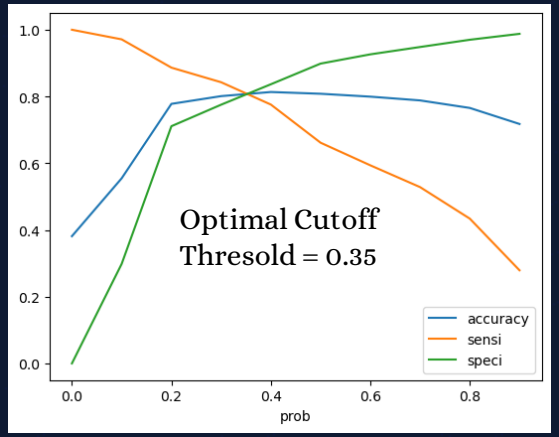
Dep. Variable:	Converted	**No. Observations:**	6468
Model:	GLM	**Df Residuals:**	6454
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-2680.8
Date:	Sun, 13 Aug 2023	**Deviance:**	5361.6
Time:	08:49:16	**Pearson chi2:**	7.29e+03
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3937
Covariance Type:	nonrobust		

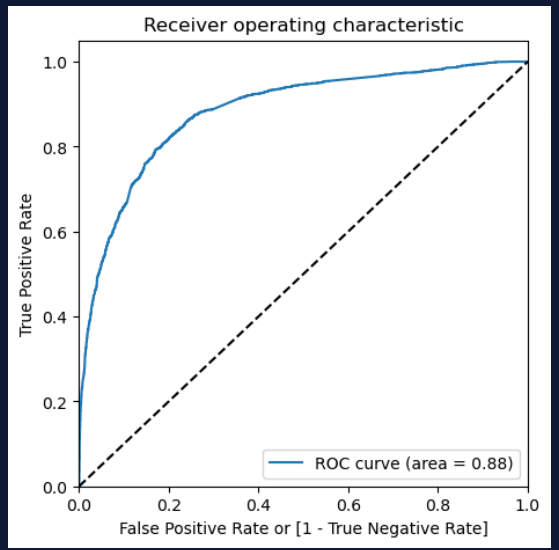
- Initial feature elimination was done by iteratively using RFE, till the variable count came down to 20.
- Subsequently, Manual Feature Elimination was followed.
- The final model has 13 predictor variables.
- All the p-values and VIFs are in the acceptable range.
- This model was used to make the predictions on the training and test data.
- Initial cutoff threshold was arbitrarily chosen as 0.5, and was further optimized to 0.35 by plotting Accuracy, Sensitivity, and Specificity.

Model Evaluation and Observations

Training Data:
- Accuracy = 80.97%
- Sensitivity / Recall = 81.02%
- Specificity = 80.93%
- Precision = 72.36%

Test Data:
- Accuracy = 81.53%
- Sensitivity / Recall = 80.91%
- Specificity = 81.93%
- Precision = 74.51%





1.Precision values for both the train and test sets are lower than the Recall values. This is as per expectations, because Recall focuses on minimizing false negatives, i.e. it is concerned with ensuring that as many true positive cases as possible are captured.

2. In the context of this business case, even if a lead is falsely classified as positive, there is no harm, as there is no loss of business. But if a lead is falsely classified as negative, that is undesirable as it is a loss of business. Hence, Recall needs to be given more importance than Precision in our case, and we have very good Recall values for both train and test sets.

3.Overall Accuracy for both train and test sets is more than 80%, and so the model is overall performing well. This is aligned with the CEO's lead conversion target of 80%.

4.Sensitivity (Recall) and specificity are more than 80% for both sets, which can be considered good.

5. Lead Score = Conversion Probability x 100.

Recommendations

1. The time a lead spends on the website has a significant impact on the lead conversion rate. Leads who spend more time on the website are more likely to convert. Hence, such leads should be contacted on priority.

2. Leads originating from the 'Lead Add Form' have higher chances of conversion; hence, such leads should be attended to on priority.

3. Come up with marketing campaigns targeted towards working professionals. Better, tailor the course structure to suit working professionals, so that they are able to manage the courses along with their work. This can be a big selling point, so the marketing executives should really stress this point during their calls/e-mails with the leads.

4. Leads from the 'Welingak' website have higher chances of conversion; hence, such leads should be attended to on priority.

5. Leads whose last notable activity was 'SMS sent' have higher chances of conversion, and should be attended to on priority.

6. Leads from 'Olark Chat' have higher chances of conversion; hence, such leads should be attended to on priority.

7. Higher the total number of visits made by a lead on the website, higher is the chance of lead conversion. Hence, frequent website visitors should promptly be attended to.

Recommendations

8. Leads whose last notable activity was 'Modified' have lower chances of conversion.

9. Leads who are specialized in Hospitality Management have lower chances of conversion. Hence, such leads can be given lower priority.

10. Leads originating from the 'Landing Page Submission' have lower chances of conversion; Hence, such leads can be given lower priority.

11. Leads whose specialization is not known have lower chances of conversion. Hence, such leads can be given lower priority.

12. If a lead has opted to not receive E-Mails, there are much lower chances of conversion. Hence, such leads are to be given lower priority.

13.Leads whose last notable activity was 'Olark Chat' conversation have lower chances of conversion. Hence, such leads are to be given lower priority.

THANK YOU!