Lead Scoring Case Study

Aparna Madireddy
&
Apurva Singh

The Problem

Problem Context

- ➔ X Education sells online courses to industry professionals.
- ➔ The company markets its courses by various methods. It gets a lot of leads but its lead conversion rate is very poor (around 30%). For example, for every 100 leads acquired, only 30 of them convert.
- ➔ The company wants to make the process more efficient and improve the lead conversion rate. To do so, the company wants to identify the most potential leads, also known as ‘Hot Leads’.
- ➔ Upon doing so, the company can focus more on acquiring leads classified as ‘Hot Leads’. This will make the process more resource-efficient and business more profitable.

Business Objective

- ➔ Identify the most promising leads, i.e, leads that are most likely to convert into paying customers.
- ➔ Identify the variables that have significant effect on the outcome/ variables that are indicative of converting.
- ➔ Build a logistic regression model to assign a lead score (0 - 100) to every lead such that a lead with higher lead score will have a higher chance of conversion.
- ➔ Achieve a target lead conversion rate of around 80%.

Solution Methodology

Data Cleaning and Preparation

1. Checking and handling duplicates in data.
2. Checking and handling missing values and other follies in data.
3. Identifying outliers and treating them accordingly.
4. Dropping redundant columns that are not useful for analysis.

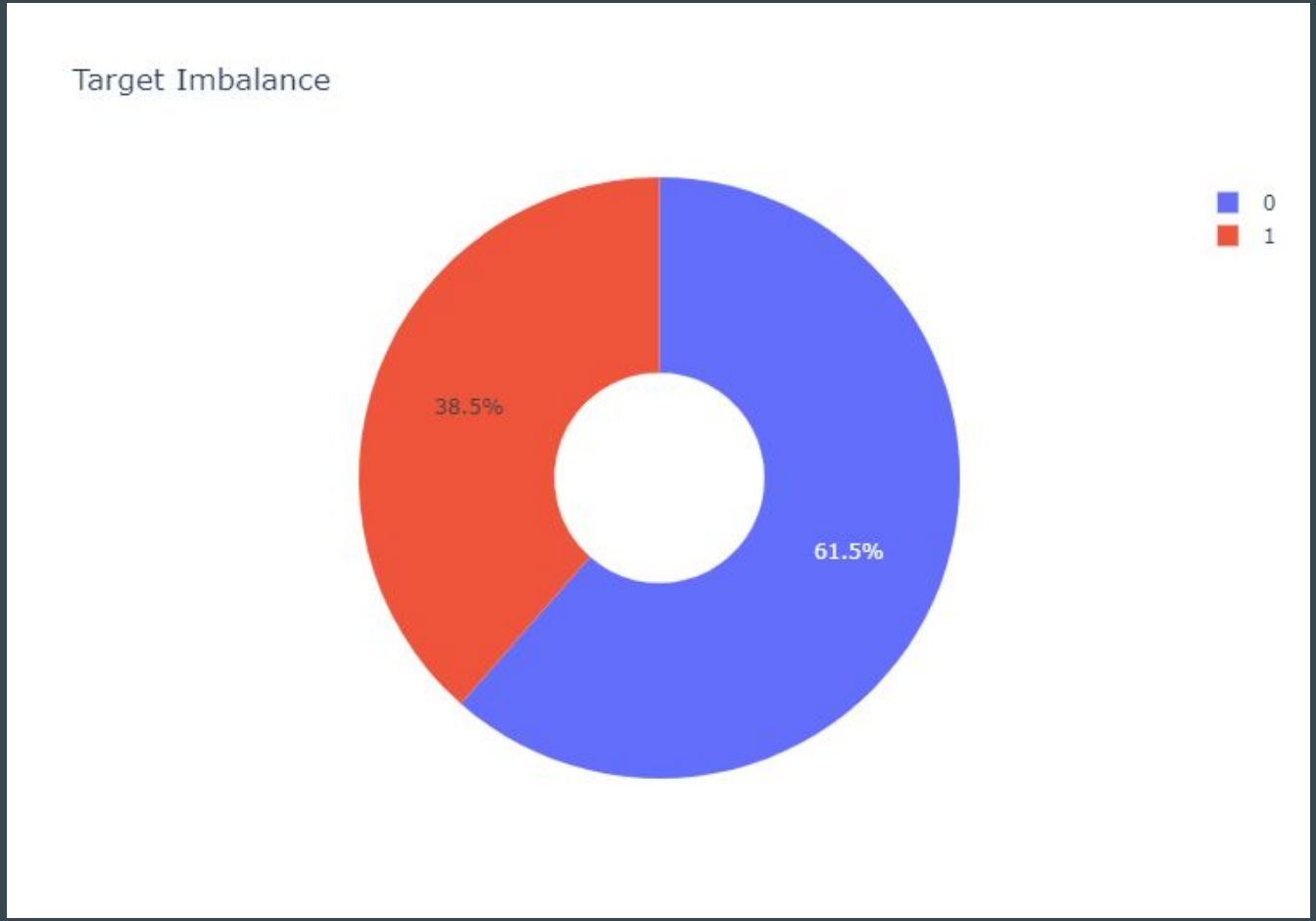
EDA and Data Preprocessing

1. Univariate and bivariate analysis on categorical and numerical features with respect to target, to identify patterns in data.
2. One-hot encoding categorical features.
3. Scaling numerical variables.
4. Model building: Logistic Regression.
5. Probability Threshold tuning.

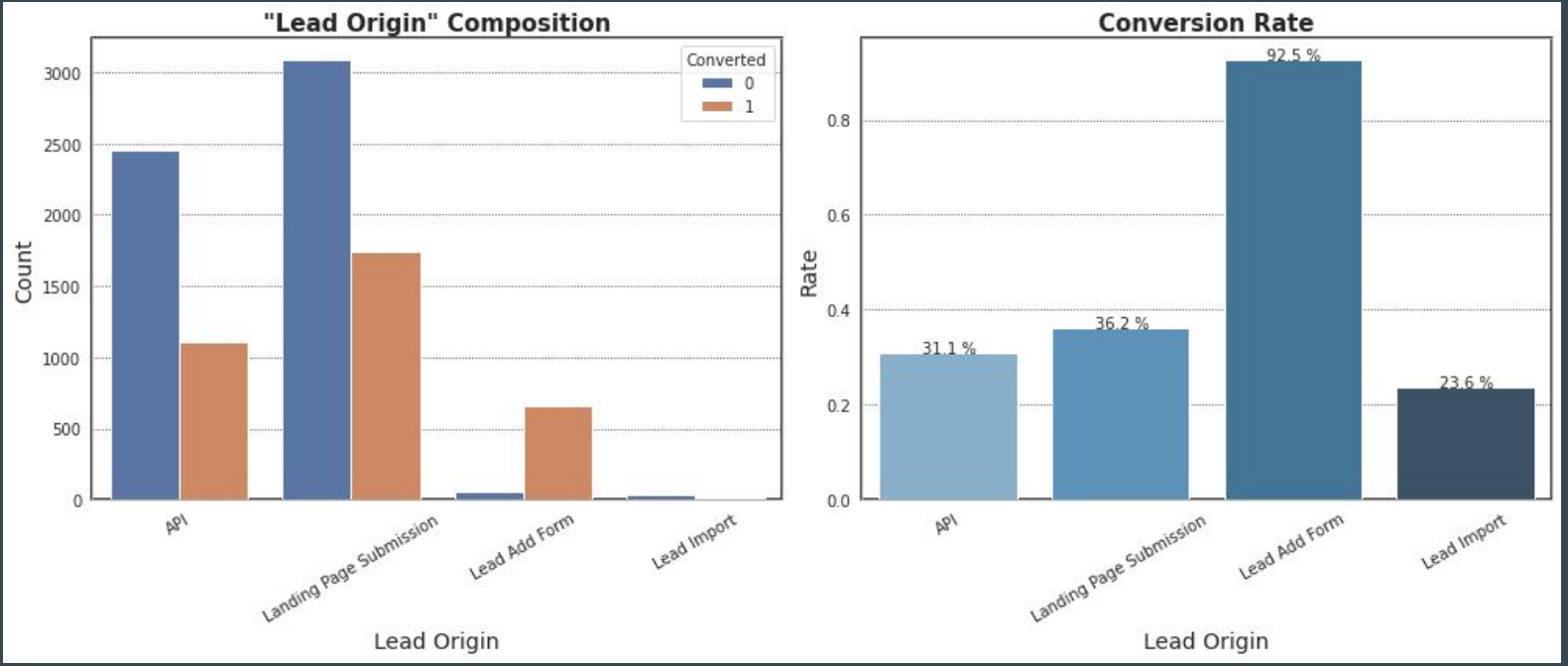
Model building and Model Evaluation

1. Using sensitivity, specificity, ROC curve, precision recall curve, AUC, and other advanced evaluation metrics for model evaluation.
2. Cross Validation by making prediction on test set.
3. Interpreting the model by accessing the results and coefficients of variables.
4. Conclusions and final recommendation.

Initial Overall Conversion Rate is 38.5%.

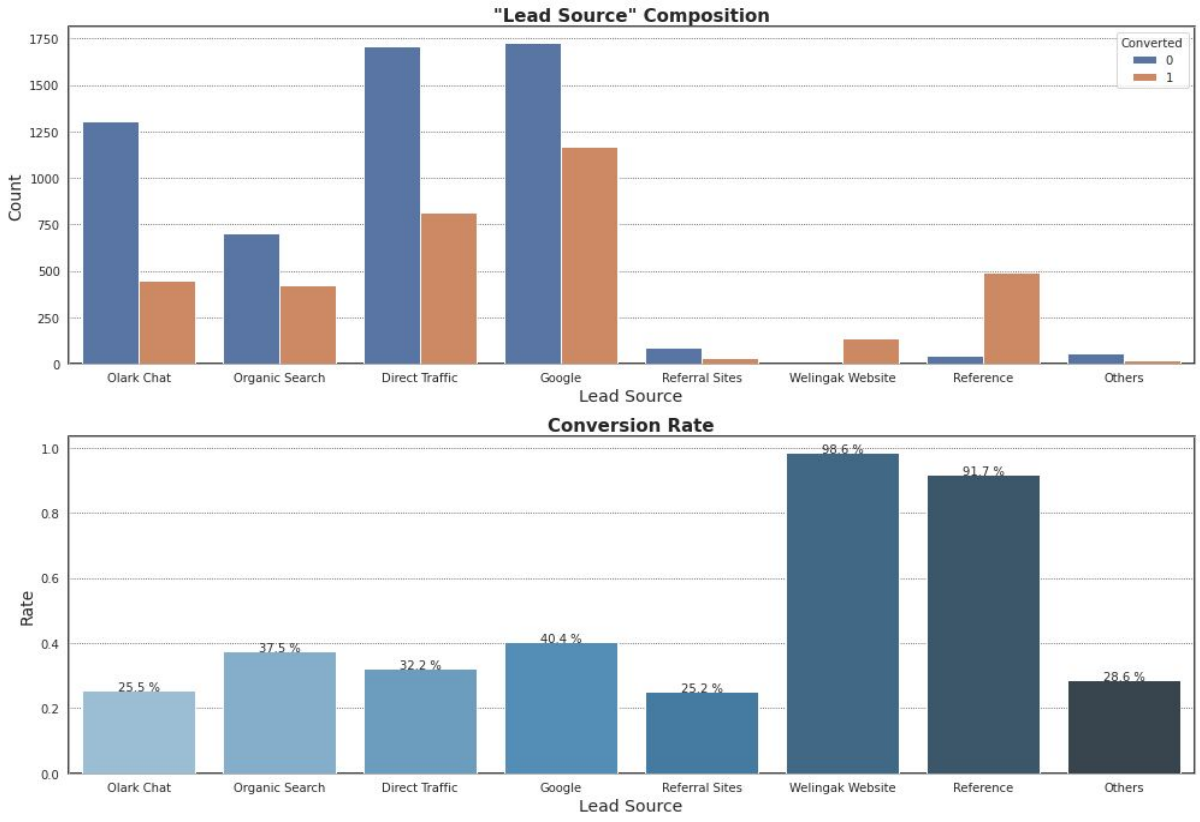


Lead Origin



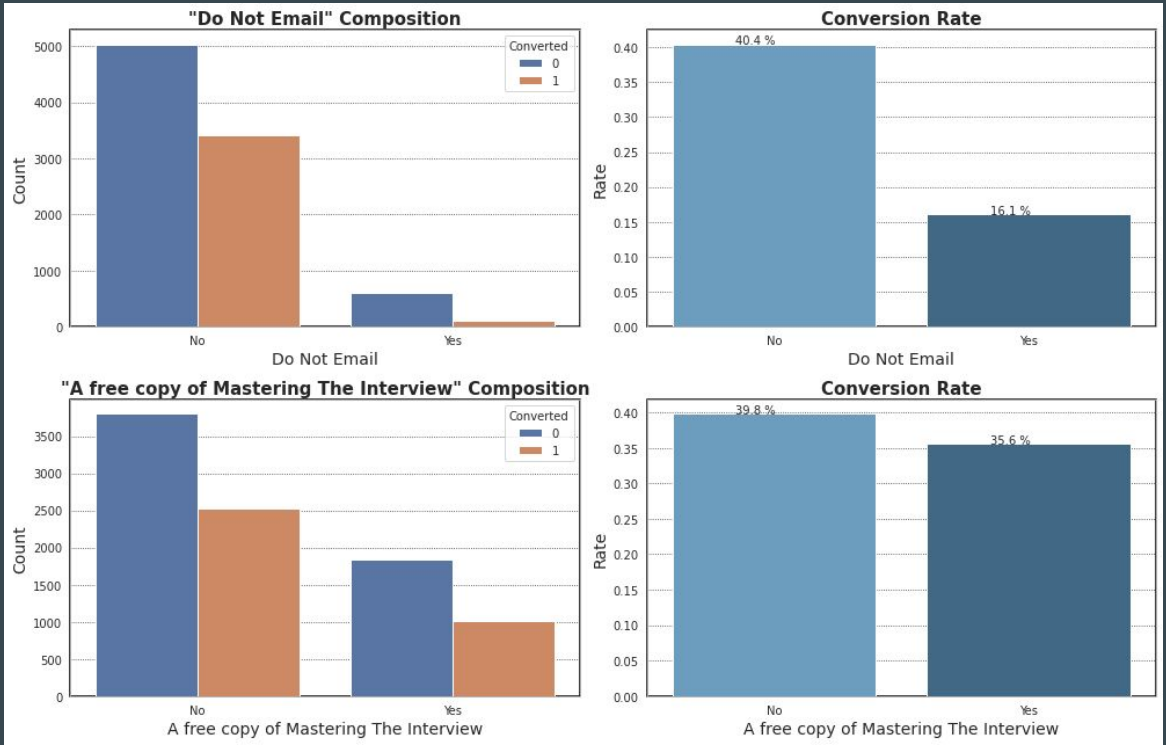
- The number of customers that were identified as leads by API and Landing Page Submission are highest, but their conversion rate is less than the average overall conversion rate, i.e, 38.5 %.
- Eventhough, Lead Add Form brings in less leads but the conversion rate of the leads identified by the it is very high.

Lead Source

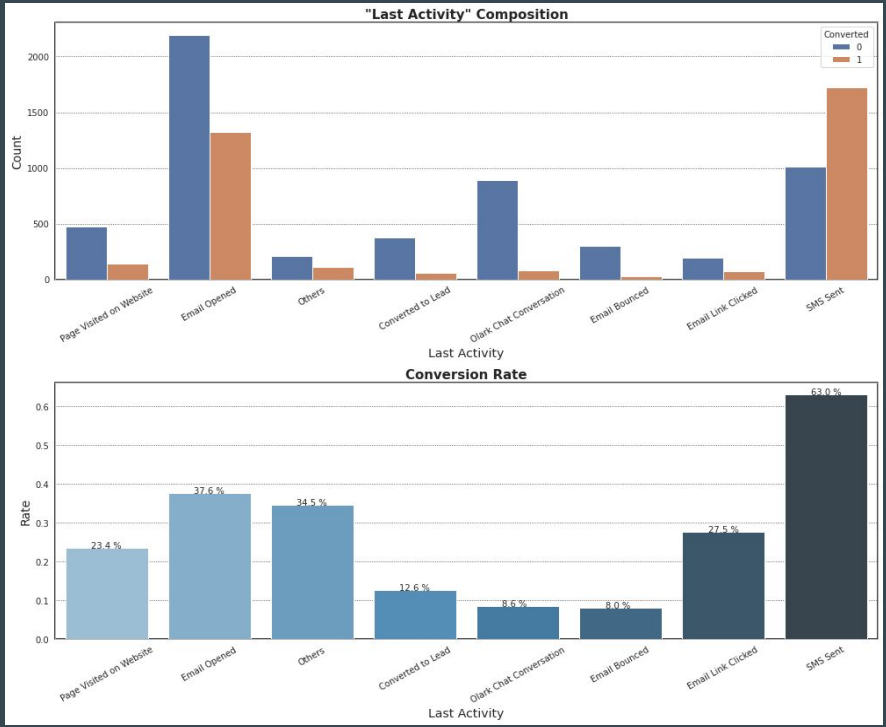


- Most number of leads come from Google and Direct Traffic. Conversion rate of leads from direct traffic is less than overall conversion rate and the same for Google is slightly more than overall average.
- A very high percentage of leads from welingak website and References have converted.

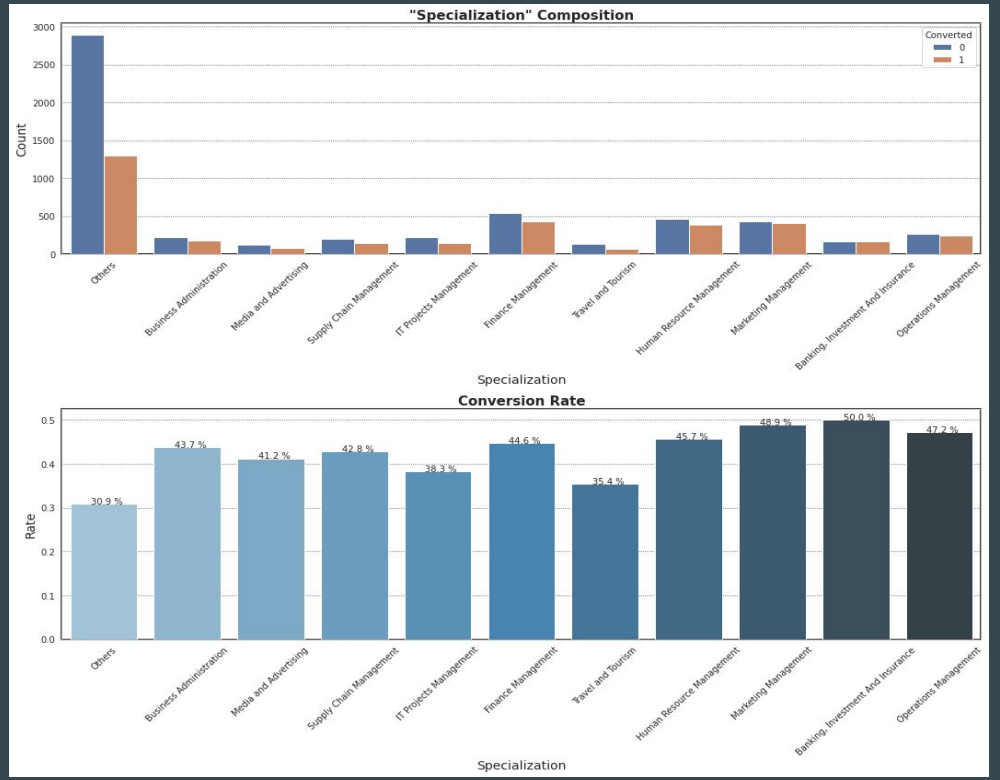
Do not Call & Do not email



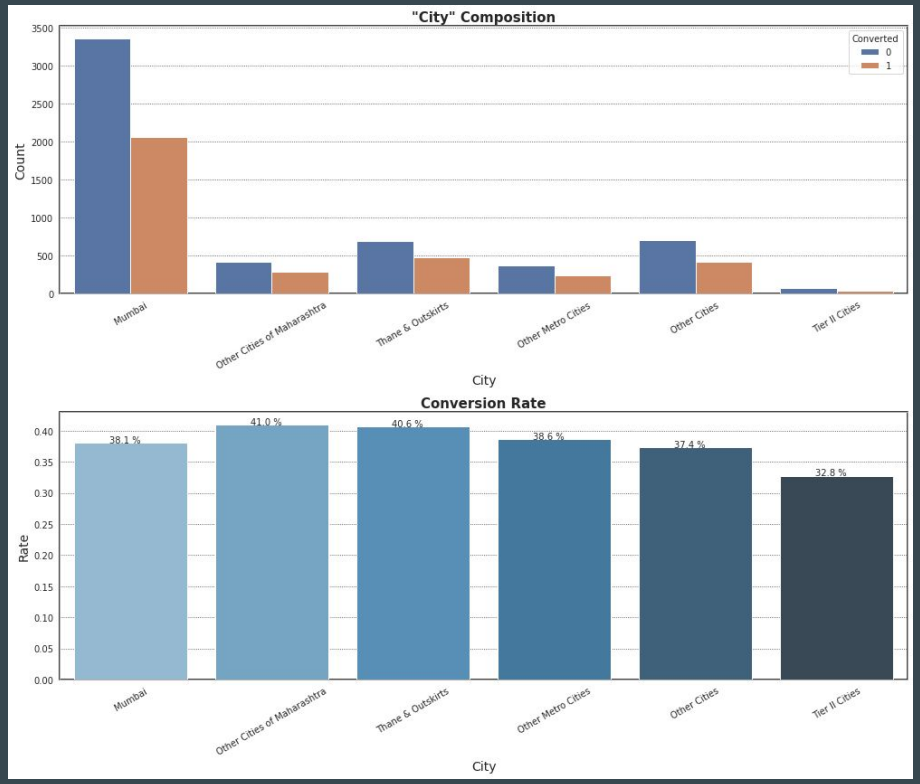
- Customers who selected to not get emailed about course converted significantly more than those who selected to get emailed about course. This group can be targeted more.
- The customers who did not want a copy of 'Mastering The Interview' had slightly more conversion rate.



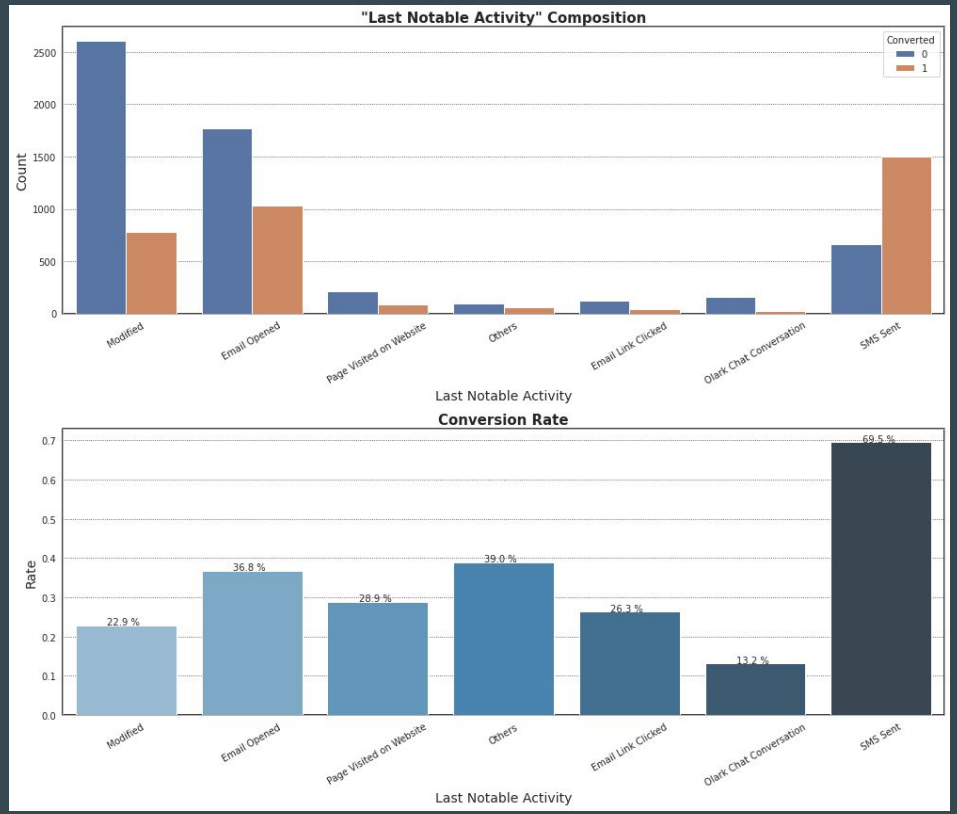
➢ Last Activity ‘SMS sent’ had the highest conversion rate.



➢ Leads from Management sector, like HR and Marketing Management, and Banking, Investment and Insurance specialization are relatively more likely to convert. Their average conversion rate is higher than the overall average. These groups can be targeted more.



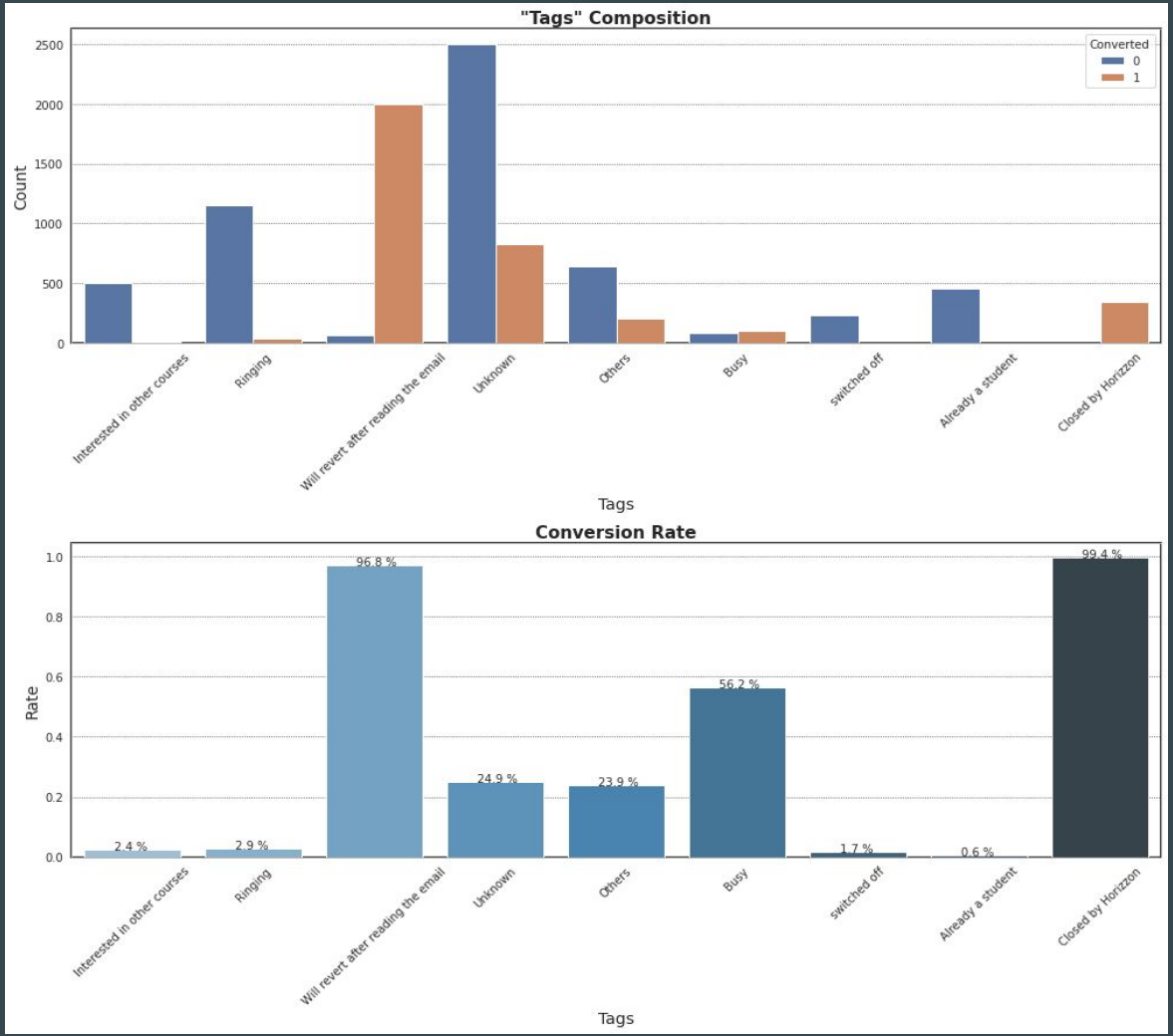
➢ A huge proportion of leads acquired are from Mumbai. Conversion rates for all the cities is close to the overall average, 38.5 %.



➢ Last Notable Activity ‘SMS sent’ had the highest conversion rate.

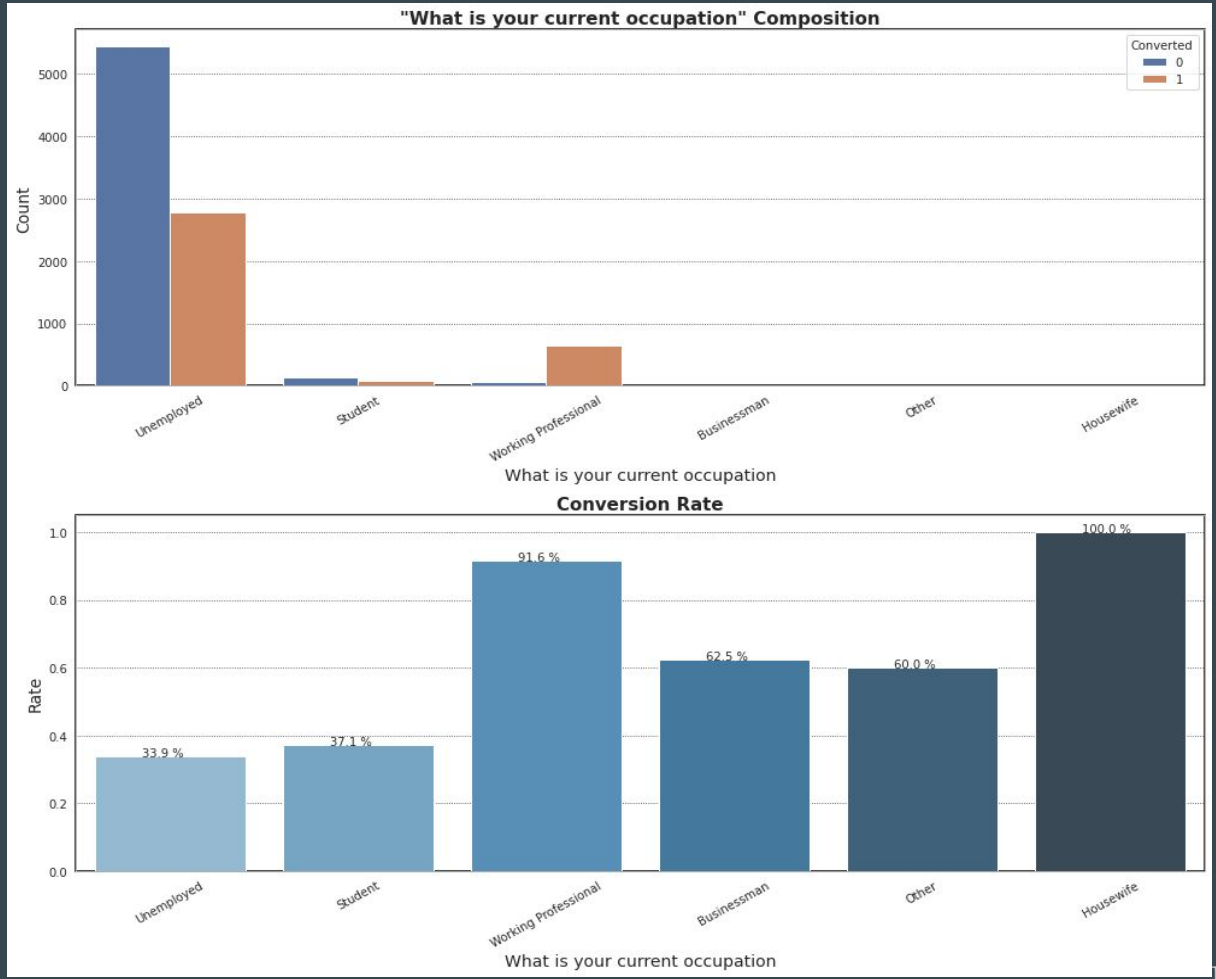
Tags

- Leads with tags/current status, 'Will revert after reading the email' have a very high likelihood of converting. This group has high potential leads.
- People with tags, 'Already a Student', 'Interested in other courses', 'Ringing' have very low conversion rate. The company should spend less resources on people in this group.

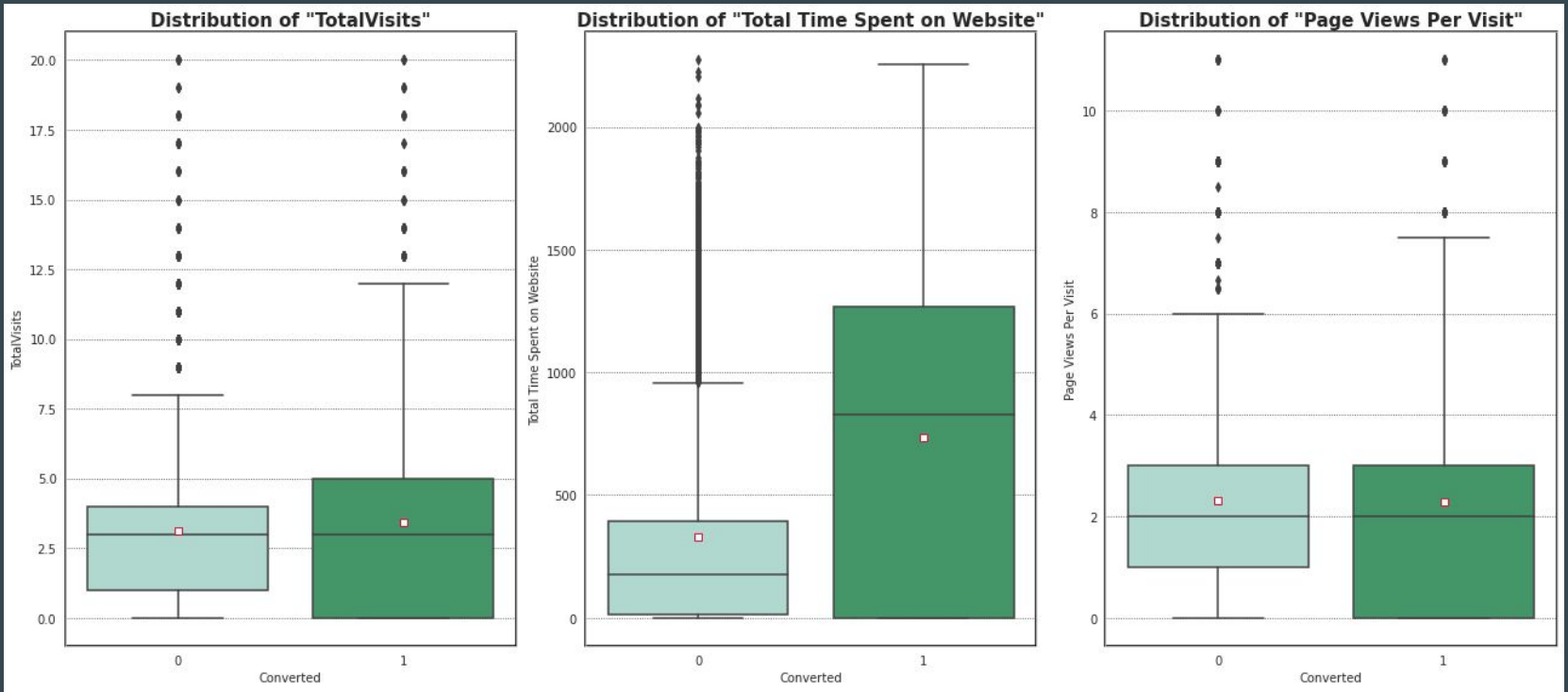


Current Occupation

- Unemployed people are least likely to convert. Working professionals have a very high conversion rate.
- Housewives have a 100% conversion rate but the data for housewives is too small to make a confident inference.



Numerical Attributes

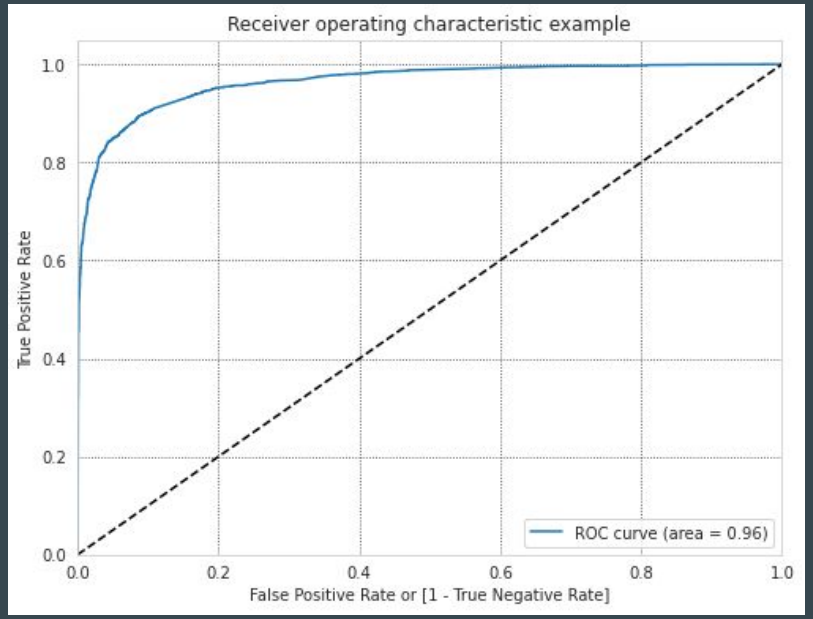


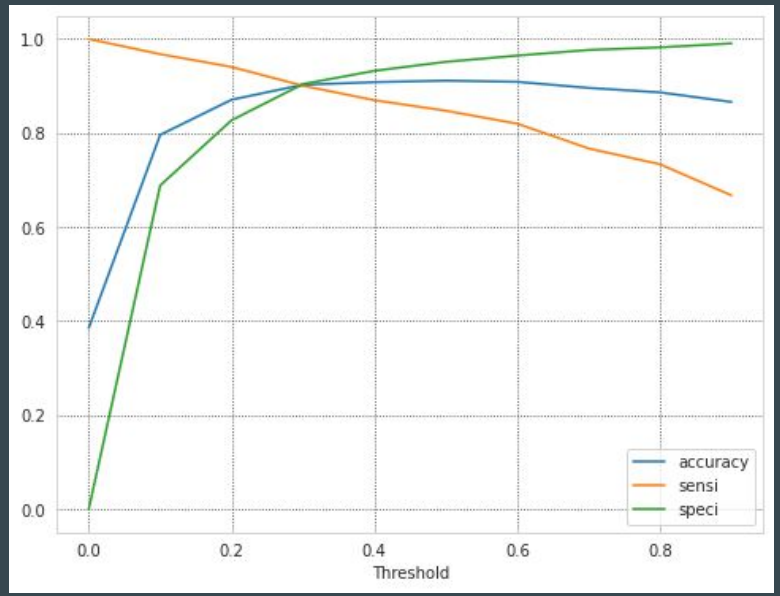
➢ Leads that spend more time on the website are more likely to convert. These people should be pursued more. Also, websites can be made more engaging and user-friendly to improve the numbers.

Model Building : Logistic Regression
Cut-off Tuning

ROC Curve

Model Performed very well considering in-sample AUC score. A trade-off between TPR and FPR is evident here.

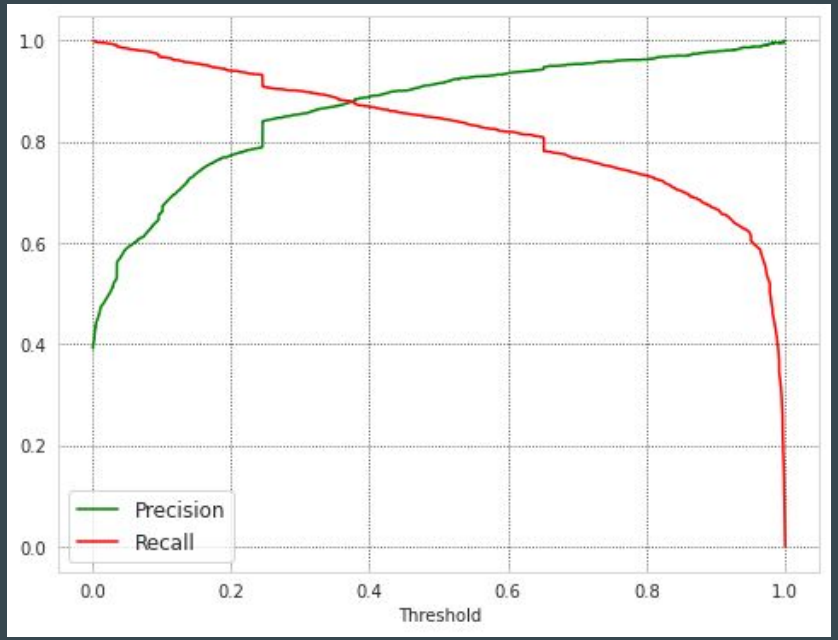




Sensitivity, Specificity and Accuracy

- We would like both sensitivity and specificity to be high. For the given context, we would like to detect as many positives(leads that convert) as possible accurately and therefore more focus should be on sensitivity.
- Eventhough it's good to improve specificity, the problem context makes positive class more important than negative class.

Precision and Recall



If we were to classify only leads predicted as positives(leads that will convert) as hot leads, the precision score has to be above 0.8 to achieve the target.

- **Hot Leads** = predicted positives(leads that will convert),
- **New Conversion Rate** = Precision Score

So, we pushed the threshold a little lower to 0.27. Following is the confusion matrix with this threshold.

4048	**460**
271	**2554**

- Overall Accuracy : 90.03%
- Precision : 84.74%
- Recall : 90.41%
- F-measure : 87.48%

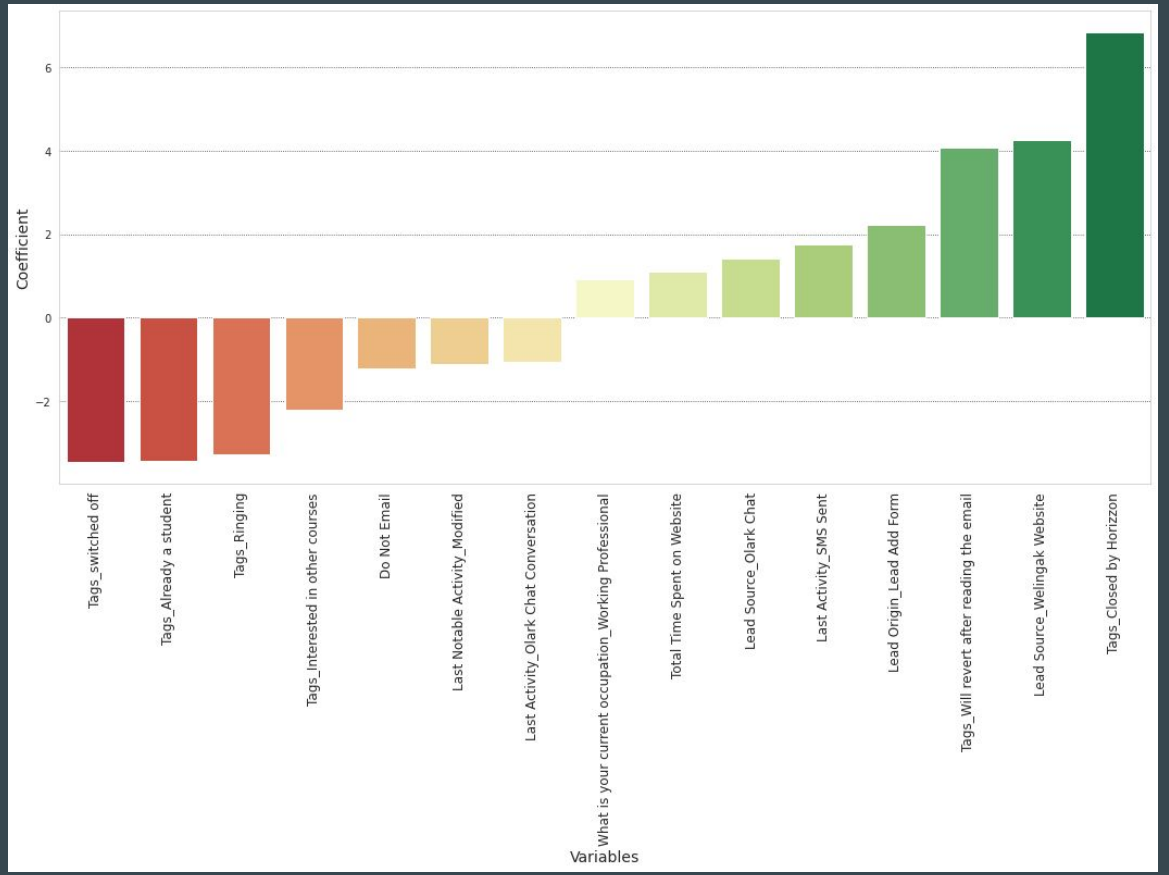
Cross Validation (making prediction on test set)

1012	**116**
65	**641**

- Overall Accuracy : 90.13%
- Precision : 84.68%
- Recall : 90.79%
- F-measure : 87.63%

Variables that impact the outcome

- 'Do Not Email'
- 'Total Time Spent on Website',
- 'Lead Origin_Lead Add Form',
- 'Lead Source_Olark Chat',
- 'Lead Source_Welingak Website',
- 'Last Activity_Olark Chat Conversation',
- 'Last Activity_SMS Sent',
- 'What is your current occupation_Working Professional',
- 'Tags_Already a student',
- 'Tags_Closed by Horizzon',
- 'Tags_Interested in other courses',
- 'Tags_Ringing',
- 'Tags_Will revert after reading the email',
- 'Tags_switched off',
- 'Last Notable Activity_Modified'



Conclusion and Recommendations

- ➔ Lead Origin: Company should try to bring in more leads by Lead Add Form.
- ➔ Lead Source: The company should invest more resources into acquiring leads from the following sources-
  - ◆ welingak website
  - ◆ References
- ➔ The company should try to acquire people who happen to be Working professionals.
- ➔ Lead who spent more time on the website should be pursued more. Also, websites can be made more engaging and user-friendly to improve the numbers.

Most important variables to consider:

- 'Tags_Closed by Horizzon'
- 'Lead Source_Welingak Website'
- 'Tags_Will revert after reading the email'
- 'Tags_switched off'
- 'Tags_Already a student'
- 'Tags_Ringing'

Improve Lead Score

Worsen Lead Score

Finally, the model has performed good and is generalising enough. It was able to identify around 90% of the leads that convert and could also deliver a precision score of about 85% in the validation set, making the new conversion rate well above 80%.

END