# LEAD SCORING CASE STUDY

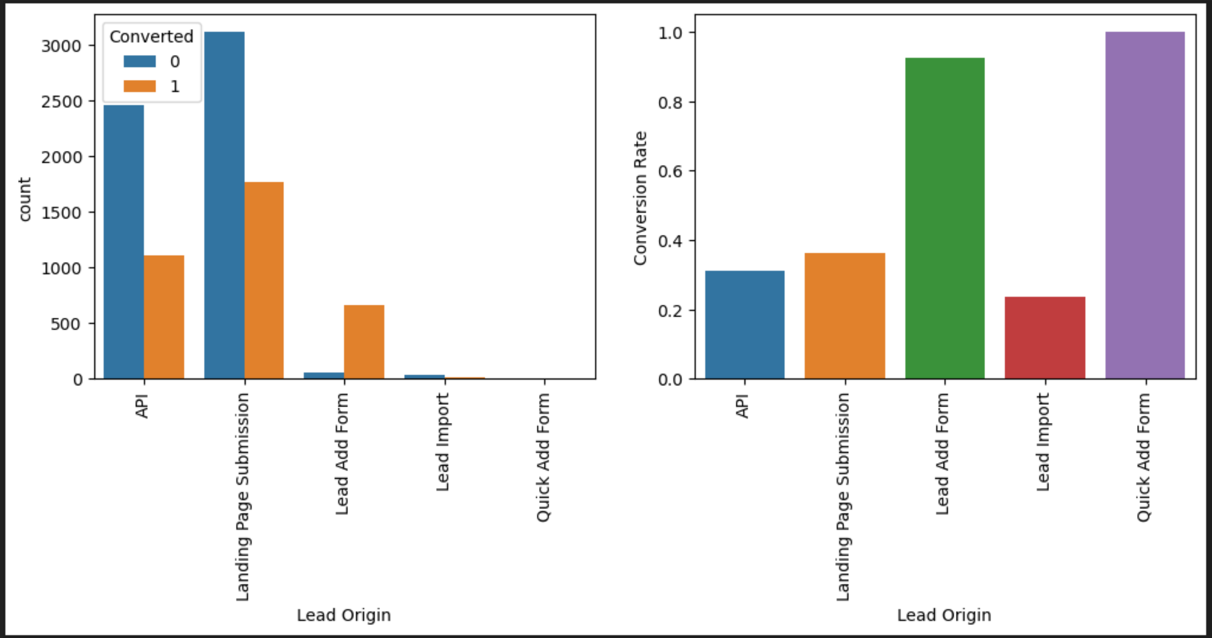
PHAM LAM PHONG

# 1. Problem Statement

- Problem Statement: Although X Education gets a lot of leads, its lead conversion rate is very poor. To make the process more efficient, 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.
- Data science problem: 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.

# Categorical Attributes Analysis

## Lead Source

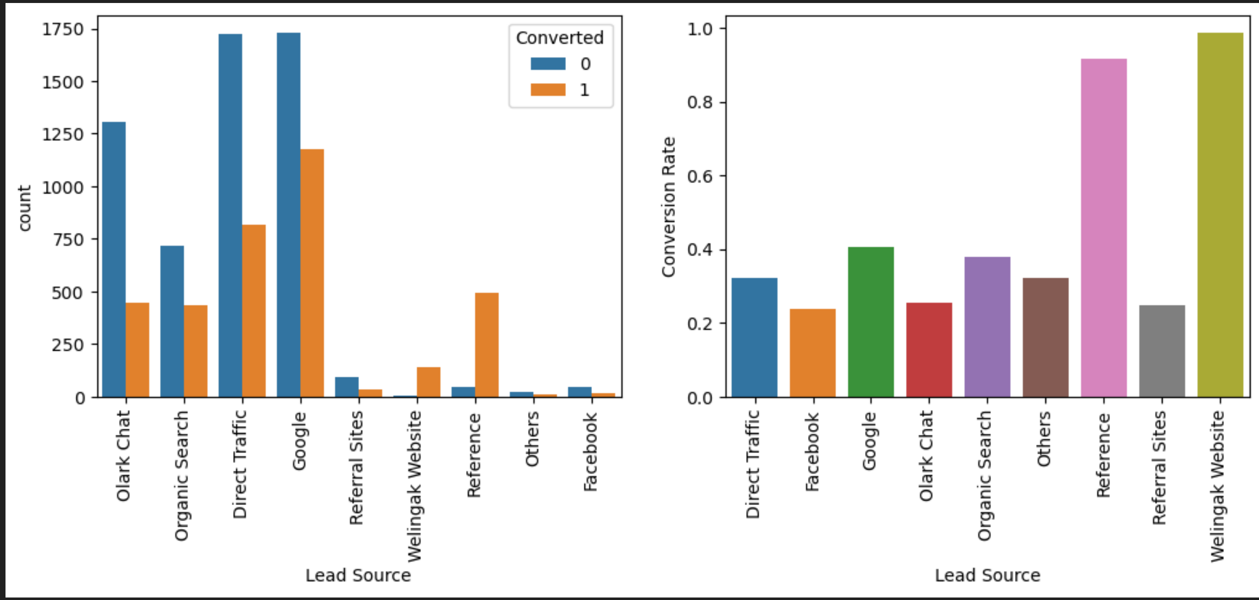



Inference:

- Although the main source of customers come from API and Landing Page Submissions, the conversion rate of 'Lead Add Form' is the highest over 90%.
- Quick add form' has only 1 observation, therefore it has no statistical significance.

# Categorical Attributes Analysis

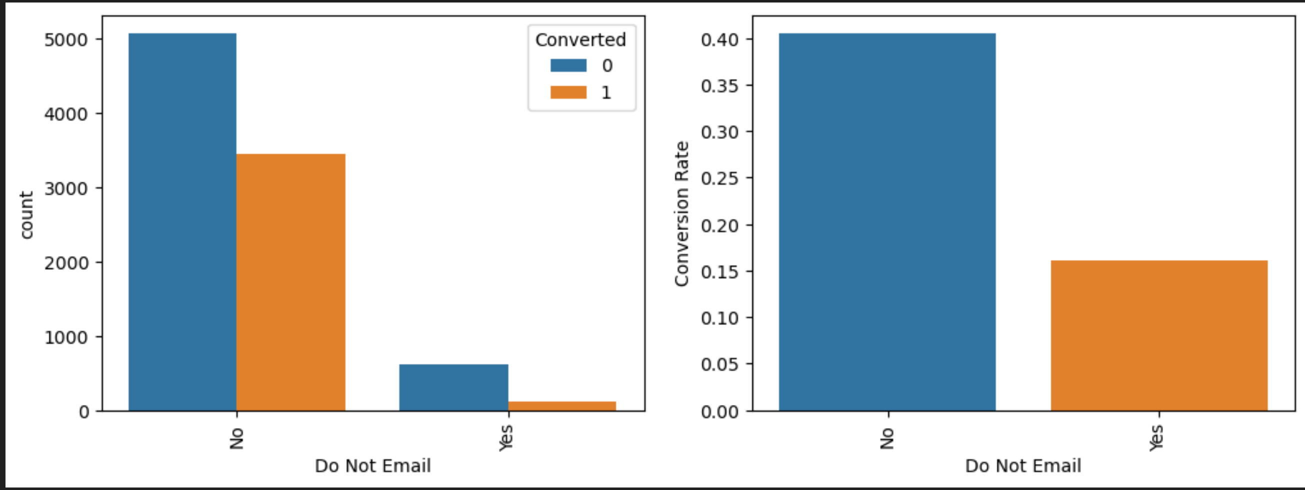
Lead Origin




Inference:
- Welingak website' and 'Reference' are two lead sources that have significantly higher conversion rates than the other lead sources

# Categorical Attributes Analysis

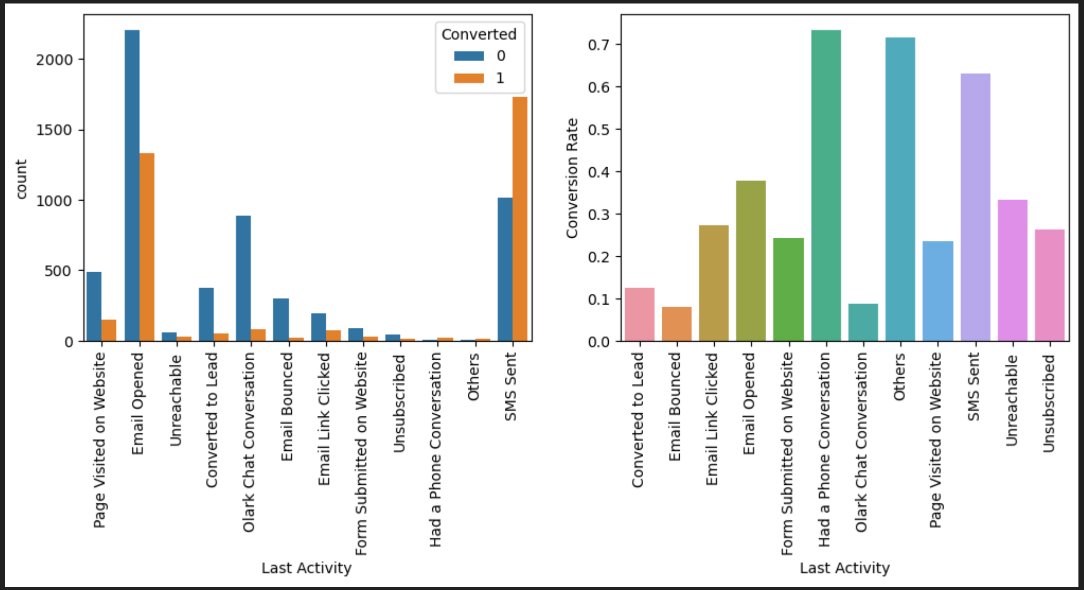
Do not email




Inference:
- We can easily observe that customers who don't send email ad have a significantly lower conversion rate

# Categorical Attributes Analysis

Last Activity

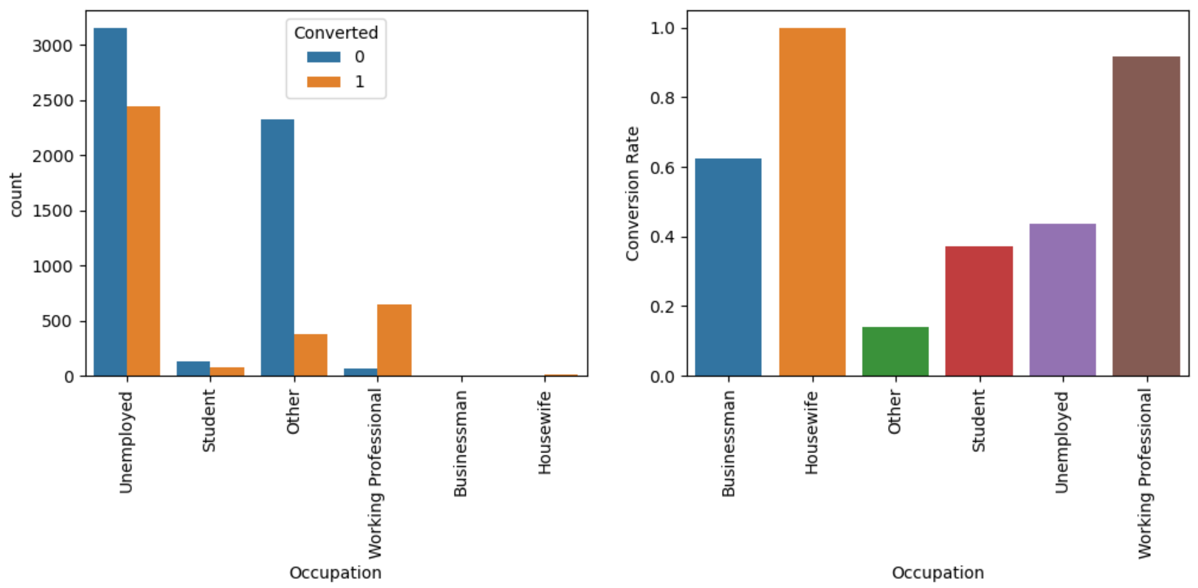



Inference:

- Customers who were last contacted via phone consultation or text message tend to have a higher conversion rate than other types of activities.
- Although 'others' activities also have high conversion rates, they have a small number of observations and many activities are grouped together, so they are not statistically significant

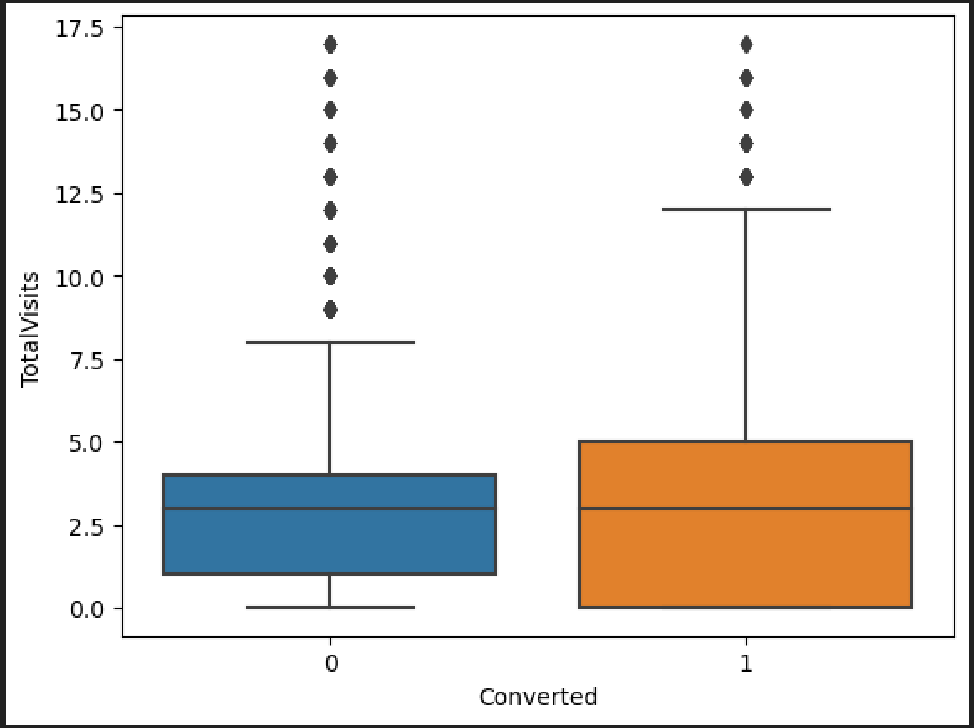
# Categorical Attributes Analysis

Occupation



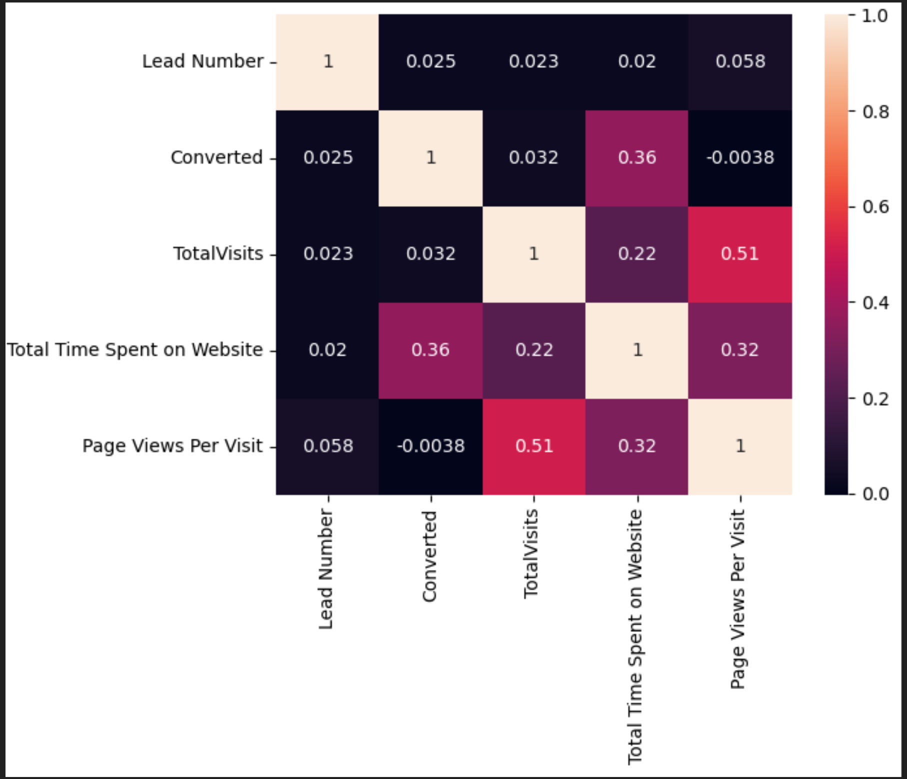

Inference:
- Professional workers typically have a significantly higher course enrollment rate than other types of occupations

# Numerical Attributes Analysis




Inference:
- Based on the graph, we can only predict that customers who are converted to lead tend to have higher total visit numbers

# Numerical Attributes Analysis




Inference:

- We can see that Total Time spend on website seems to have high correctlation than others features

# Model Building

Generalized Linear Model Regression Results

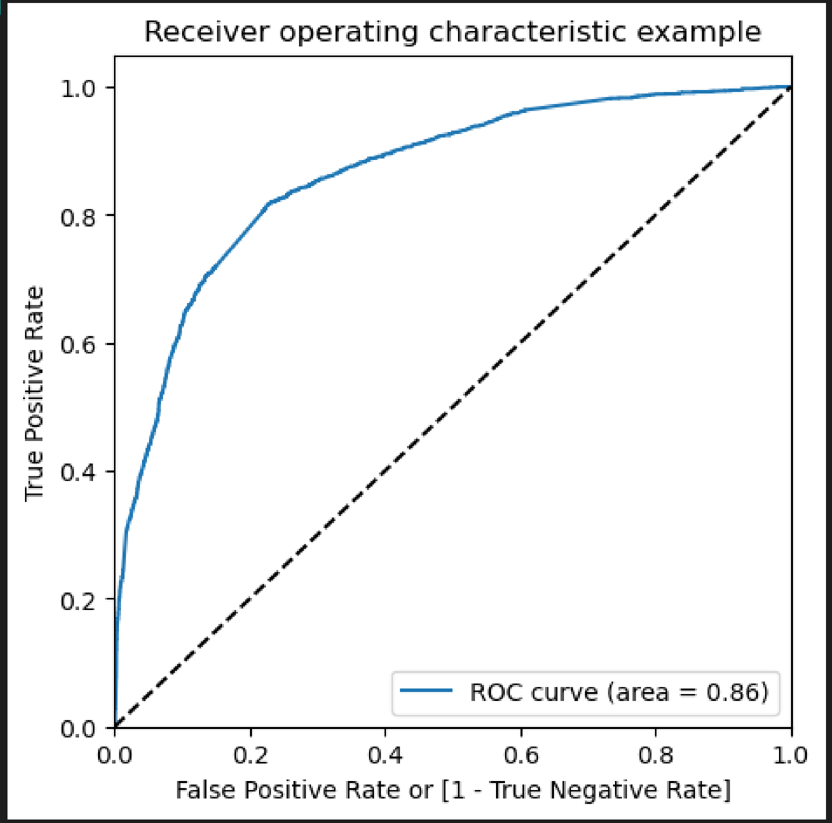
| | | | |
|---|---|---|---|
| Dep. Variable: | Converted | No. Observations: | 6409 |
| Model: | GLM | Df Residuals: | 6398 |
| Model Family: | Binomial | Df Model: | 10 |
| Link Function: | Logit | Scale: | 1.0000 |
| Method: | IRLS | Log-Likelihood: | -2877.1 |
| Date: | Tue, 07 Mar 2023 | Deviance: | 5754.2 |
| Time: | 23:30:29 | Pearson chi2: | 7.37e+03 |
| No. Iterations: | 7 | Pseudo R-squ. (CS): | 0.3540 |
| Covariance Type: | nonrobust | | |

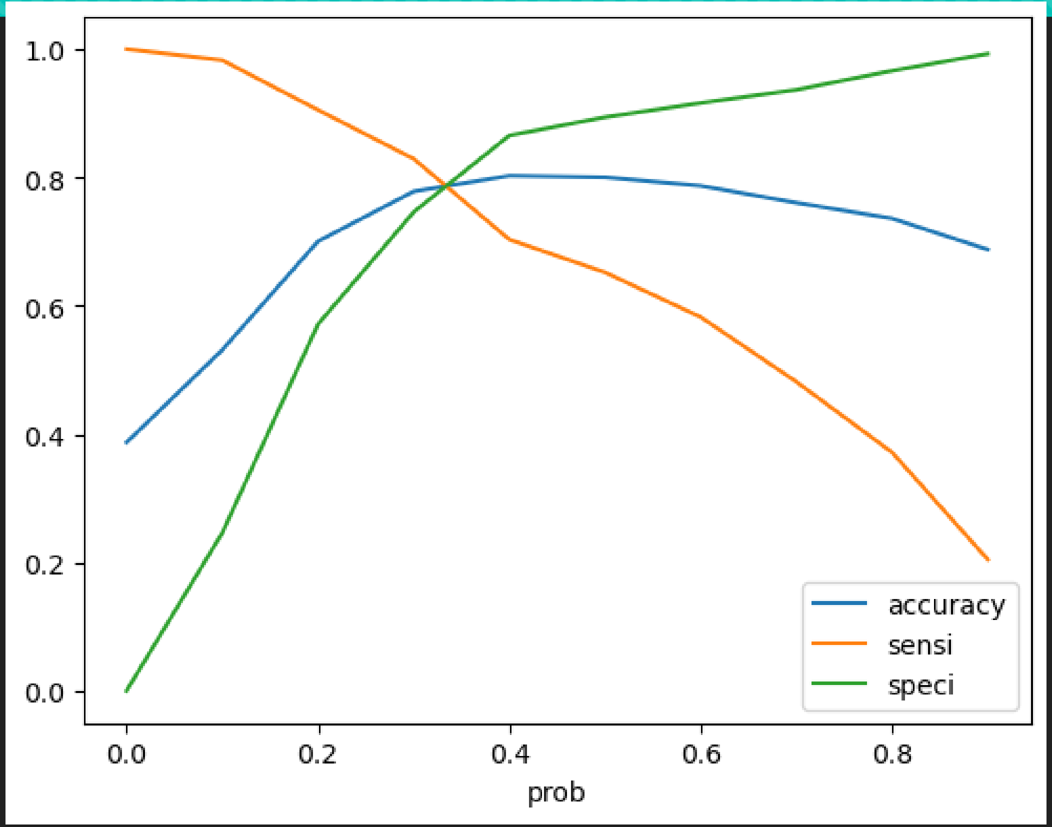
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.5052 | 0.890 | 0.567 | 0.570 | -1.240 | 2.250 |
| Do Not Email | -0.3252 | 0.042 | -7.763 | 0.000 | -0.407 | -0.243 |
| Total Time Spent on Website | 1.1343 | 0.039 | 29.079 | 0.000 | 1.058 | 1.211 |
| Lead Origin_Lead Add Form | 3.6918 | 0.198 | 18.650 | 0.000 | 3.304 | 4.080 |
| Lead Source_Google | 0.3091 | 0.076 | 4.079 | 0.000 | 0.161 | 0.458 |
| Lead Source_Olark Chat | 1.2831 | 0.101 | 12.727 | 0.000 | 1.086 | 1.481 |
| Lead Source_Welingak Website | 2.0024 | 0.743 | 2.694 | 0.007 | 0.546 | 3.459 |
| Occupation_Other | -2.8474 | 0.892 | -3.191 | 0.001 | -4.597 | -1.098 |
| Occupation_Student | -1.8400 | 0.912 | -2.018 | 0.044 | -3.627 | -0.053 |
| Occupation_Unemployed | -1.4604 | 0.890 | -1.641 | 0.101 | -3.205 | 0.284 |
| Occupation_Working Professional | 0.8887 | 0.905 | 0.982 | 0.326 | -0.885 | 2.663 |

Using RFE technique to remove attributes and build a model for the remaining attributes.

Making the model stable by using StatsModels library

# Model Building




- Performing is quite well with above model. The ROC curve has a value of 0,86 which is good.
- From the curve above, 0.2 is the optimum point to take it as a cutoff probability.

○ Accuracy : 76.38% ○ Sensitivity :83.06% ○ Specificity : 72.28%