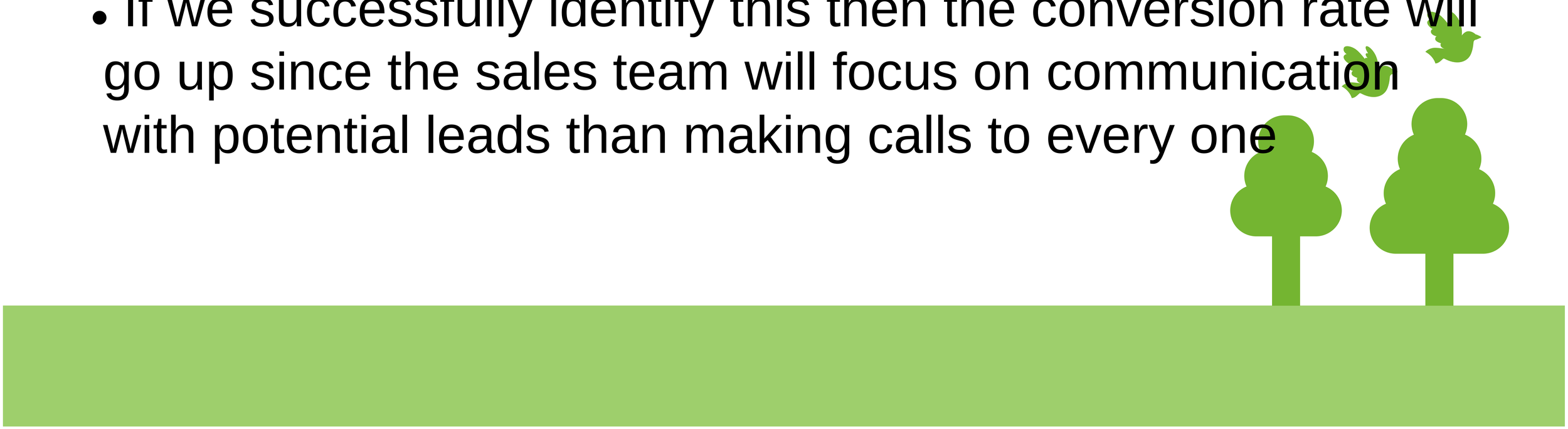
Lead Score Case Study

- Himesh Chopra
- Janakiramana R
- Jaspal Singh Sandhu

Problem Statement

- X Education sells online courses to industry professionals.
- X Education get a lot of leads but the conversion rate is very poor. Its only 30%
- To make this better, We need to identify ‘hot leads’ i.e. the most potential leads
- If we successfully identify this then the conversion rate will go up since the sales team will focus on communication with potential leads than making calls to every one



Business Goals

- X education wants to know most promising leads
- They need a model that identify hot leads
- The model needs to be deployed for future use

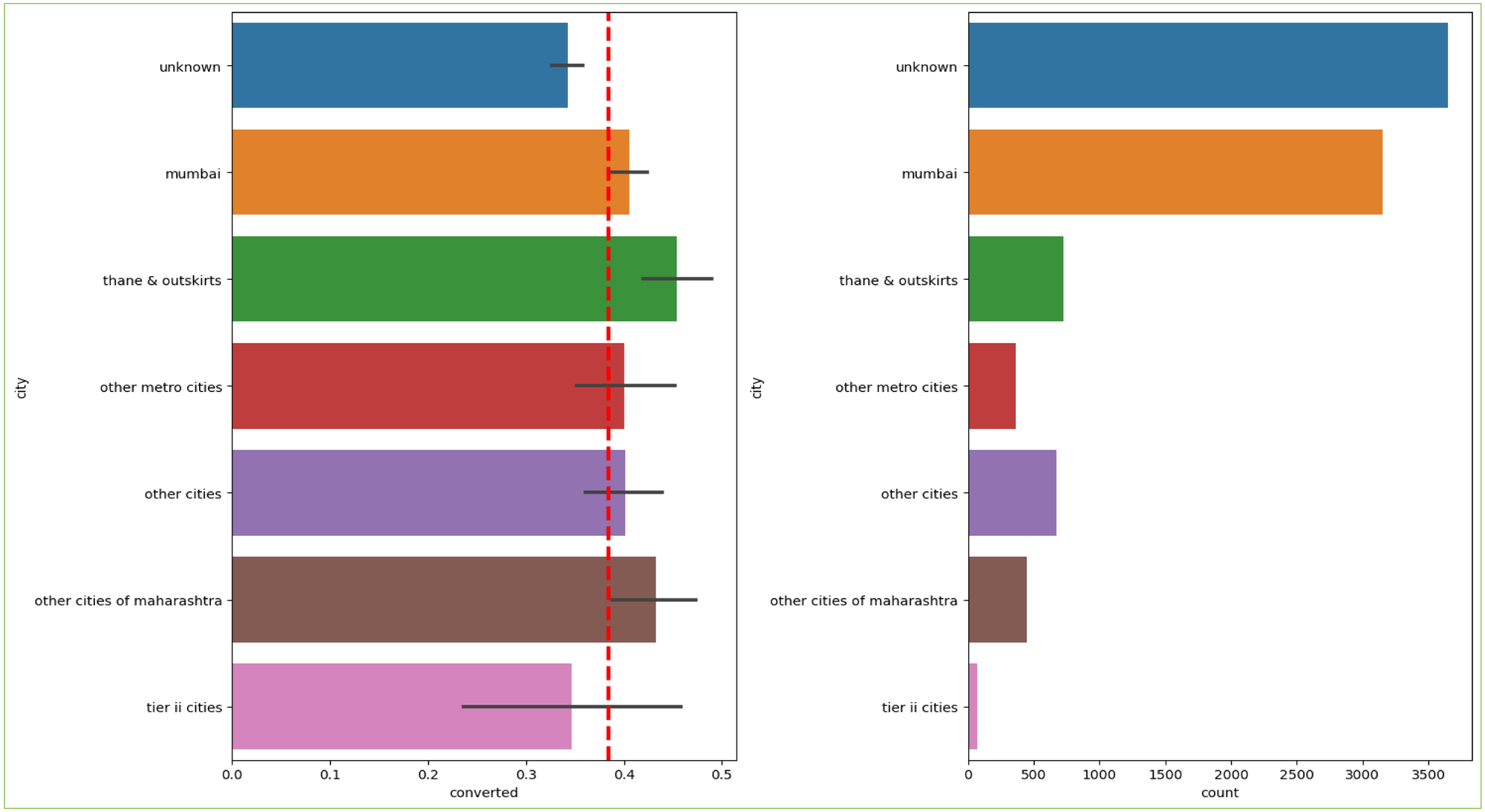
Solution Design

- Data Cleanup and Manipulation
  - Check whole dataset and handle duplicate data
  - NA and missing values will be handled
  - Drop columns containing nulls mostly
  - Handle Outliers in the data
- EDA
  - Univariate analysis
  - Bivariate analysis
- Encoding the data and dummy variables
- Classification: Using logistic regression model
- Build model and validation
- Conclusion and Recommendation.

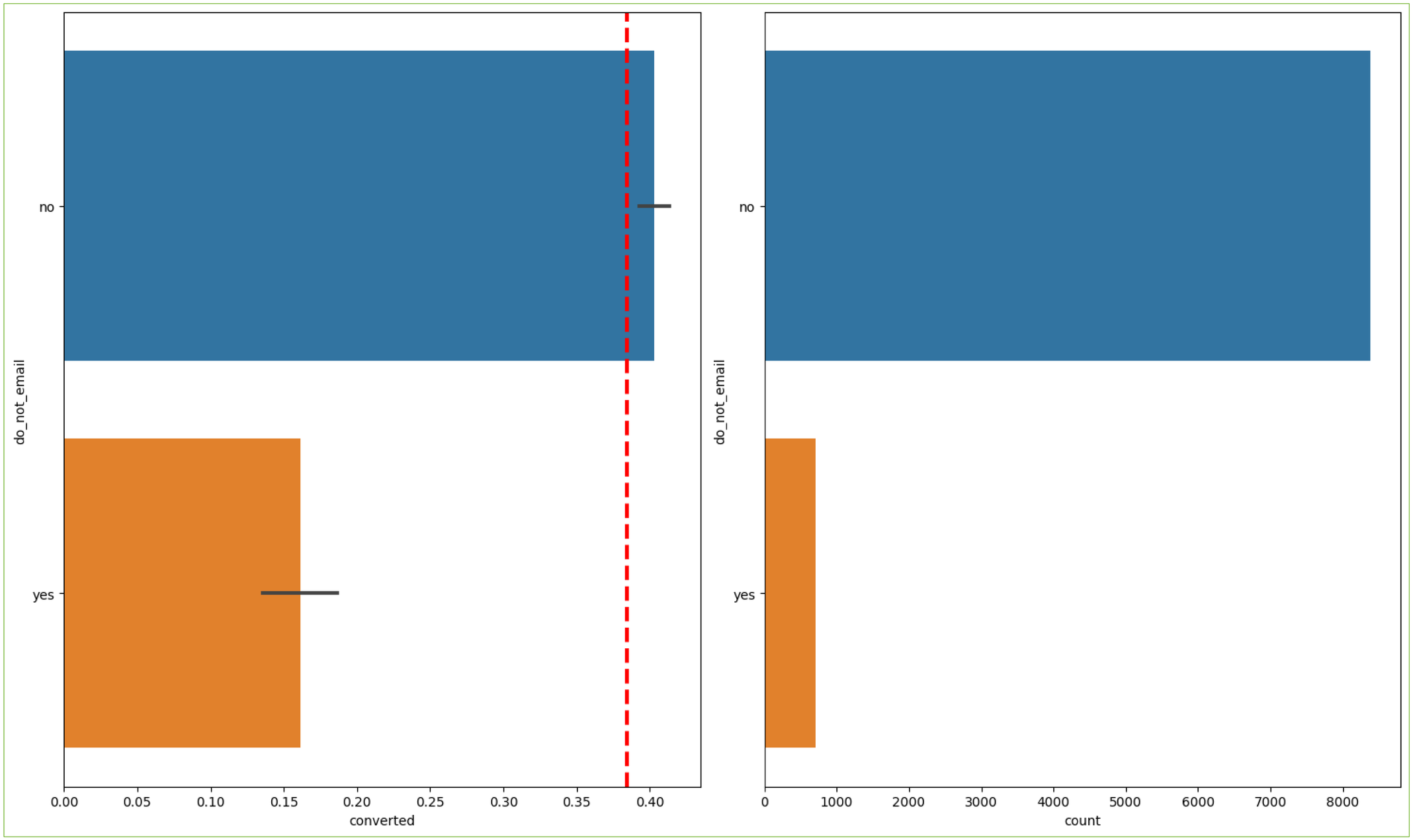
Data Cleanup and Manipulation

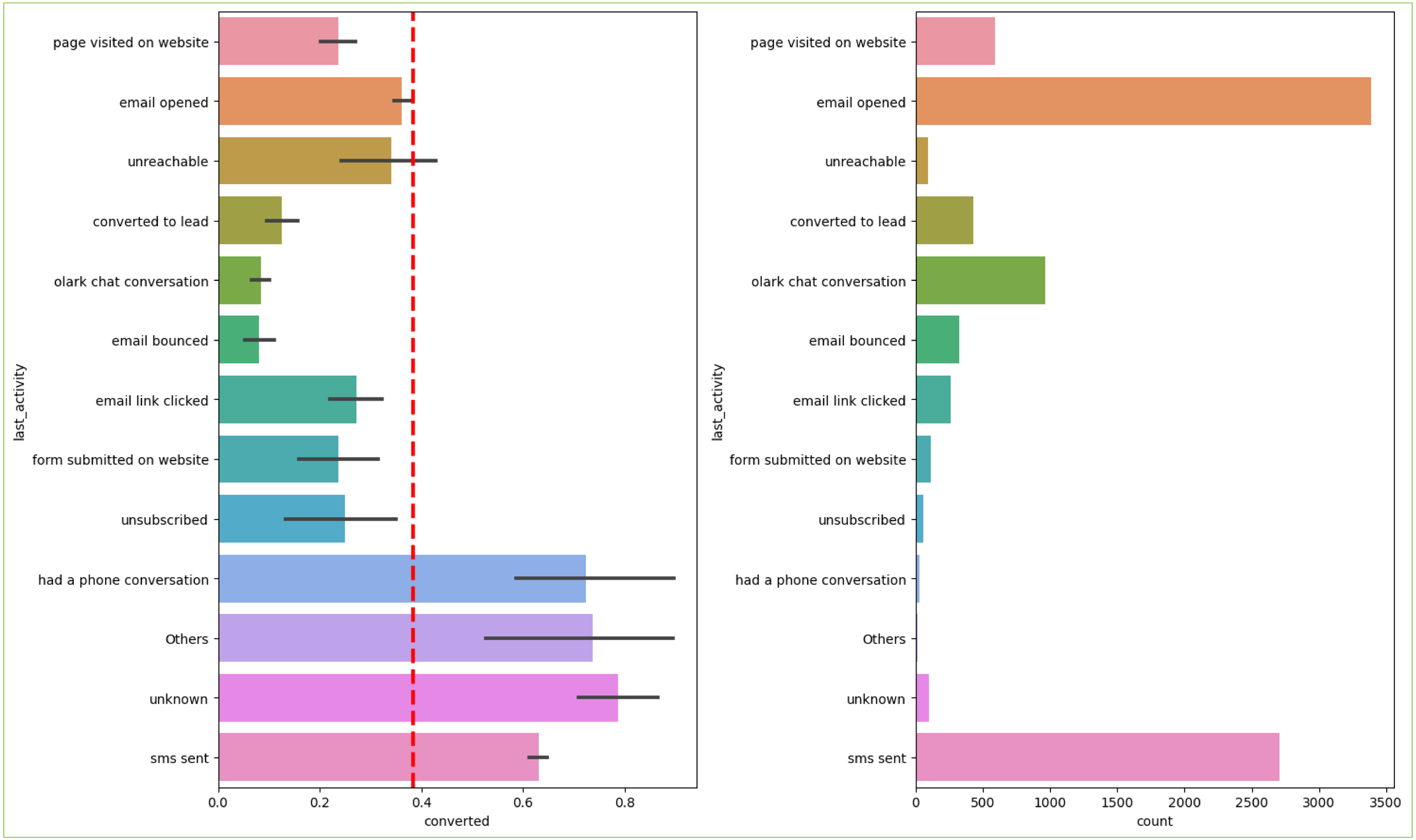
- Total Number of Rows is 9240 and Total Number of Columns is 37
- Dropping the columns having more than 40% missing values

•EDA [Categorical Variable Analysis]

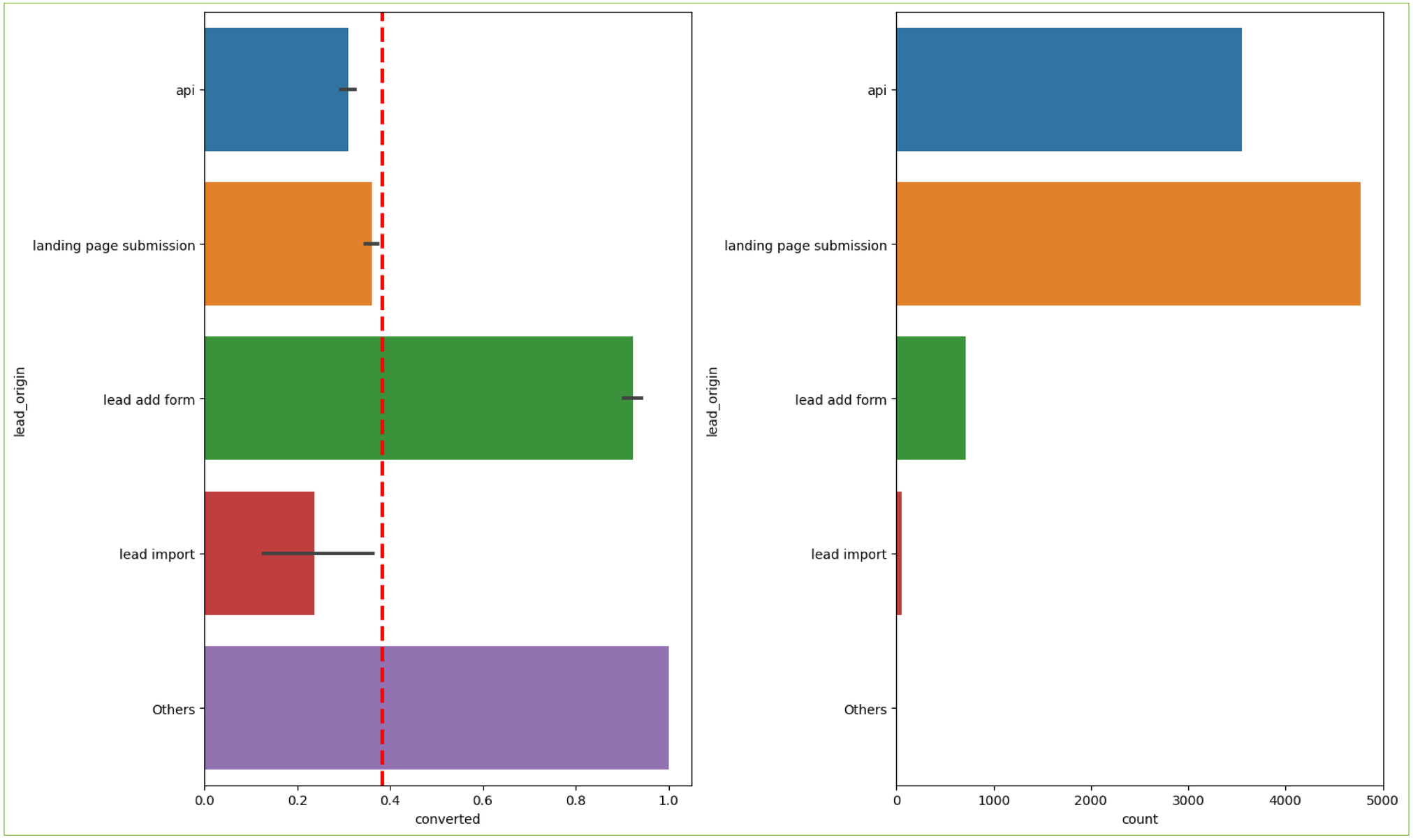


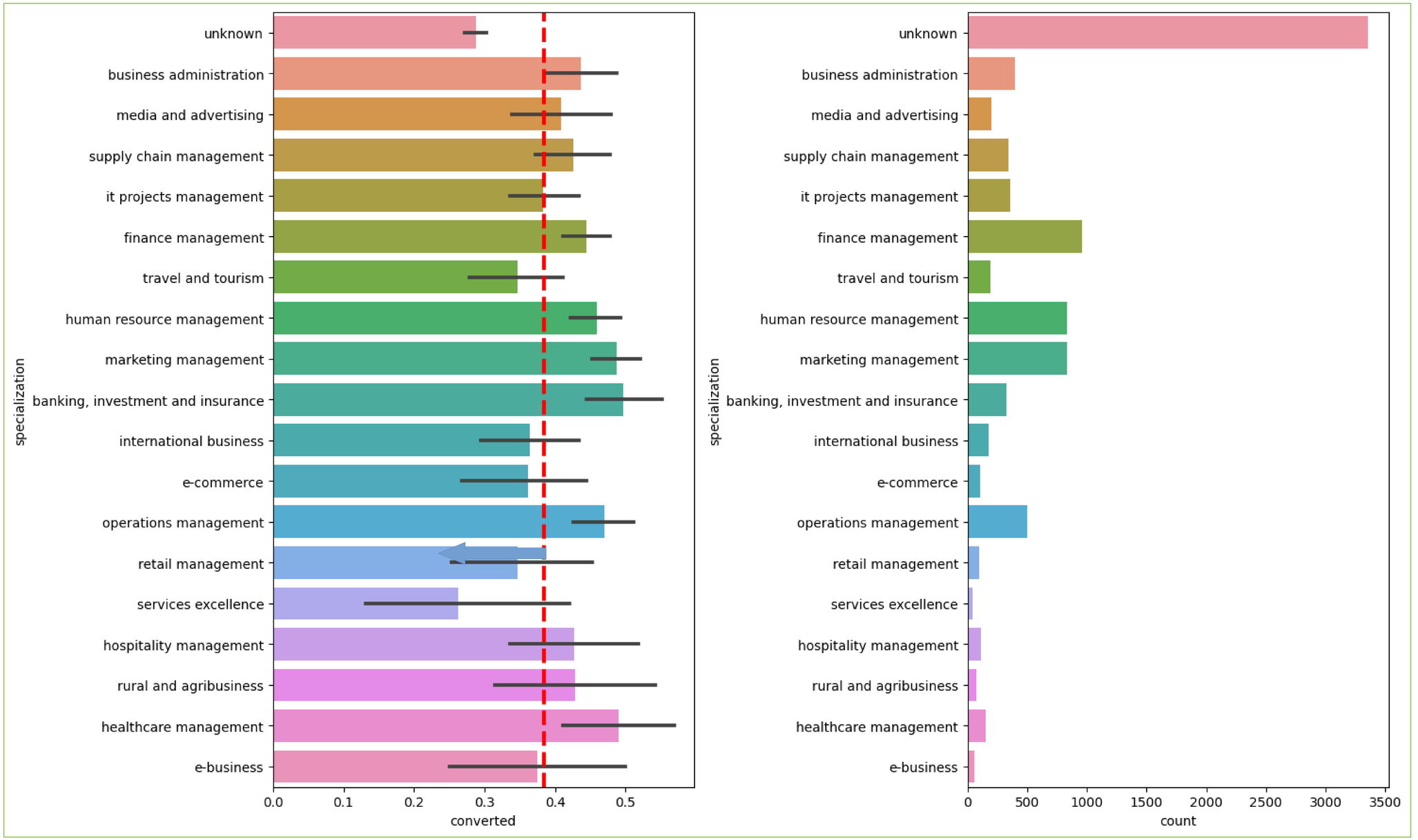
When City is not mentioned,the conversion rate has a drop


do_not_email
no
yes
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
converted
do_not_email
no
yes
0
1000
2000
3000
4000
5000
6000
7000
8000
count



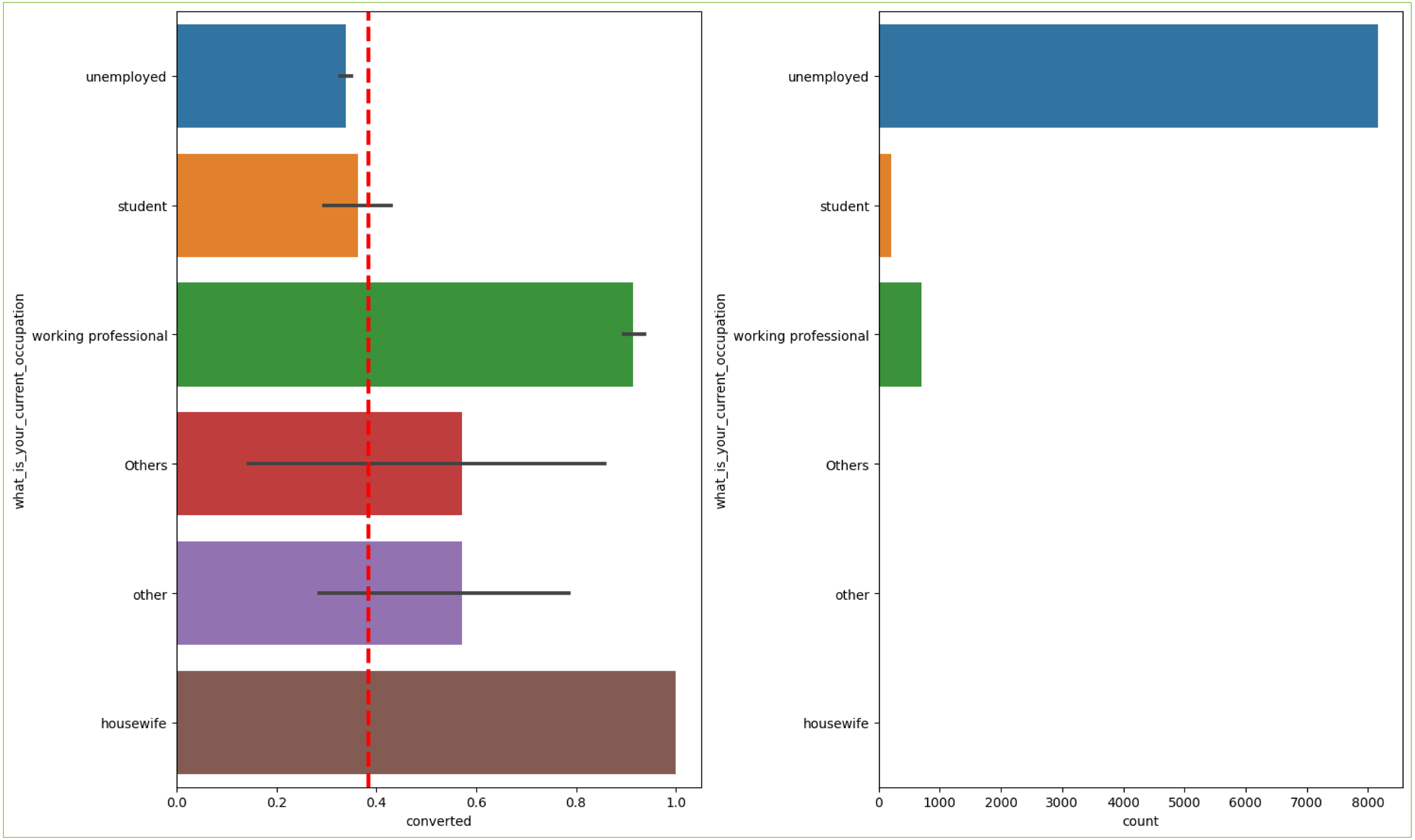
If last activity is sms sent, then there is high chance of conversion


lead_origin
api
landing page submission
lead add form
lead import
Others
0.0
0.2
0.4
0.6
0.8
1.0
converted
lead_origin
api
landing page submission
lead add form
lead import
Others
0
1000
2000
3000
4000
5000
count



1. If specialization is unknown then the conversion rate suddenly drops.

2. Management studies of all kind have a good conversion rate

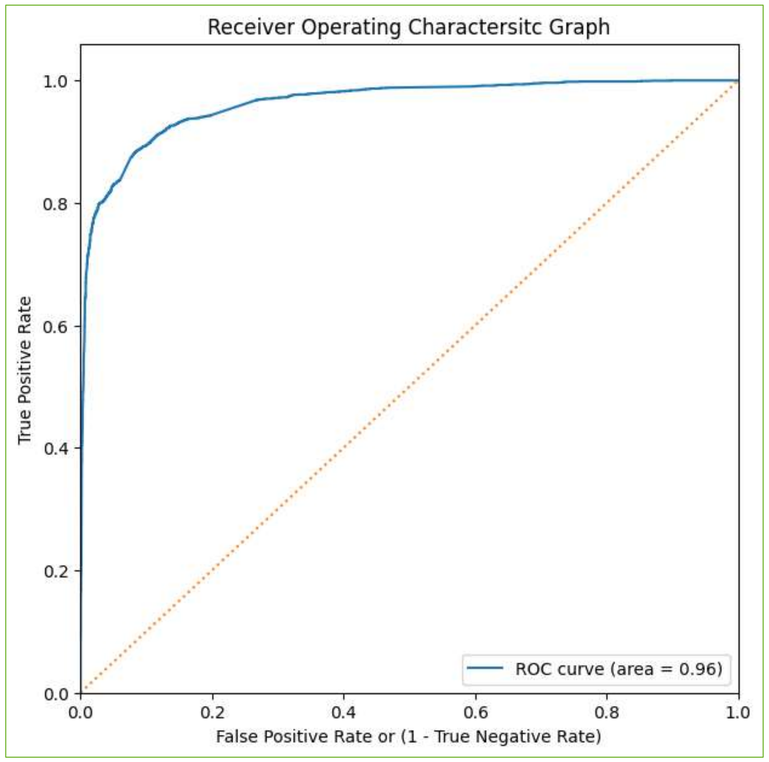


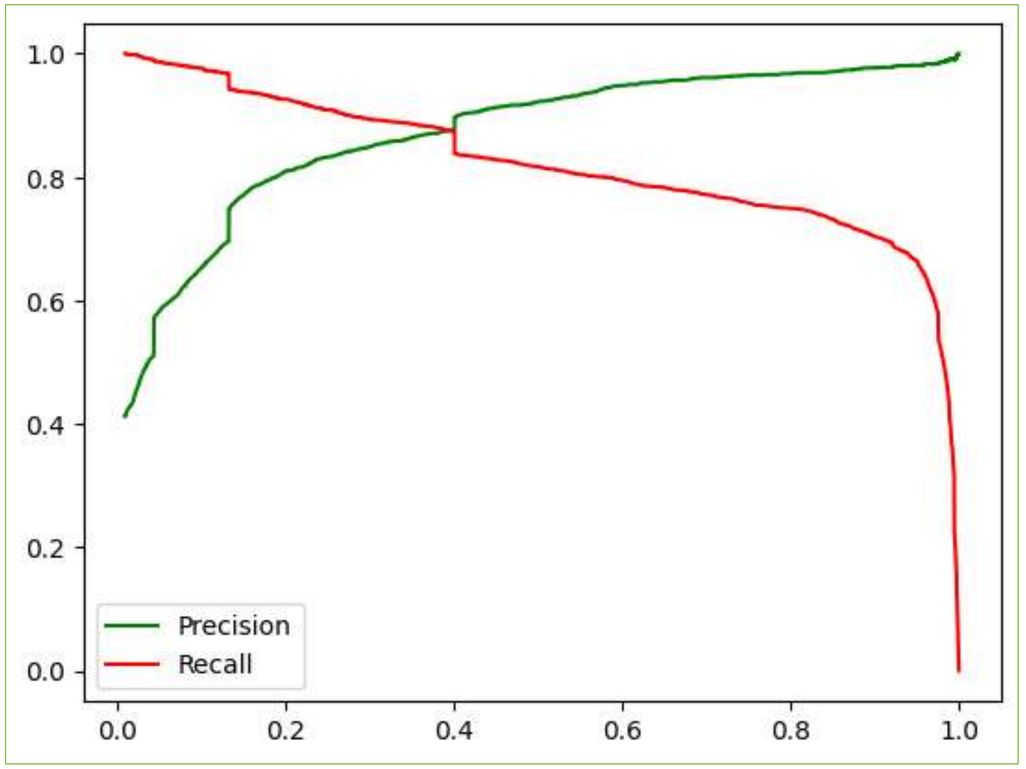
Working Professionals is good area to target

Model Building

- Split data into Train and Test Data
- The first step is to train-test split the data. We did this in ratio 67:33 ratio.
- Used RFE for Feature Selection
- Running RFE with 20 variables as output
- Build the model by removing high p-value and VIF values

•ROC Curve





- Finding Optimal Cut Off Point
- Optimal cut off point is that probability where we get balanced sensitivity and specificity
- From Second Graph we see optimal cut off is 0.40

•Conclusion

> **Train Data Score:**
>
> **Accuracy**: 90.14% **Sensitivty**: 90.10% **Specificity**: 90.22%
>
> **Test Data Score:**
>
> **Accuracy**: 90.14% **Sensitivty**: 90.10% **Specificity**: 90.22%

The Variables that have highest impact in identifying potential buyers are below (In descending order):

- When customer has tag: a. Will revert by email b. Closed by horizzon
- When last activity is sms sent
- Total time Spent on the Website
- When lead source was: a. Welingak b. Direct Trafic c. Google d. Organic Search

Keeping these in mind X Education can make their conversion rate better and get almost all potential buyers to buy their courses.