Lead Score Case Study

Vijay Kumar Jha
EPGPDS IIITB &
UPGRAD

Problem Statement

An education company named X Education sells online courses to industry professionals.

X Education gets a lot of leads, but its lead conversion rate is very poor ~30%. To make this process more efficient, the company wishes to identify the most potential leads, so the lead conversion rate should go up as the sales team will now be focusing more on potential leads rather than making calls to everyone.

As a data analyst, we are supposed to build a model wherein we will assign a lead score to each leads based of conversion probability and identify most promising leads, using a logistic regression model.

Solution Methodology

Data Cleaning and Sanitization:

- Dropping Irrelevant fields.
- Duplicate values handling.
- Null values handling – imputation and category creation.
- Fields sanitization – New Category creation.
- Checking Outliers.

EDA (Exploratory data analysis):

- Univariate Analysis – Distribution of variable, count of values, category comparison.
- Bivariate Analysis – Correlation and pattern b/w variables.

Model Building & Evaluation:

- Creating Dummy Variables, Train Test Split & Feature Scaling.
- Classification Techniques – Logistic Regression & Predictions.
- Model Validation & Evaluation - Accuracy, Sensitivity, Specificity, Precision and Recall.

Conclusion of the model

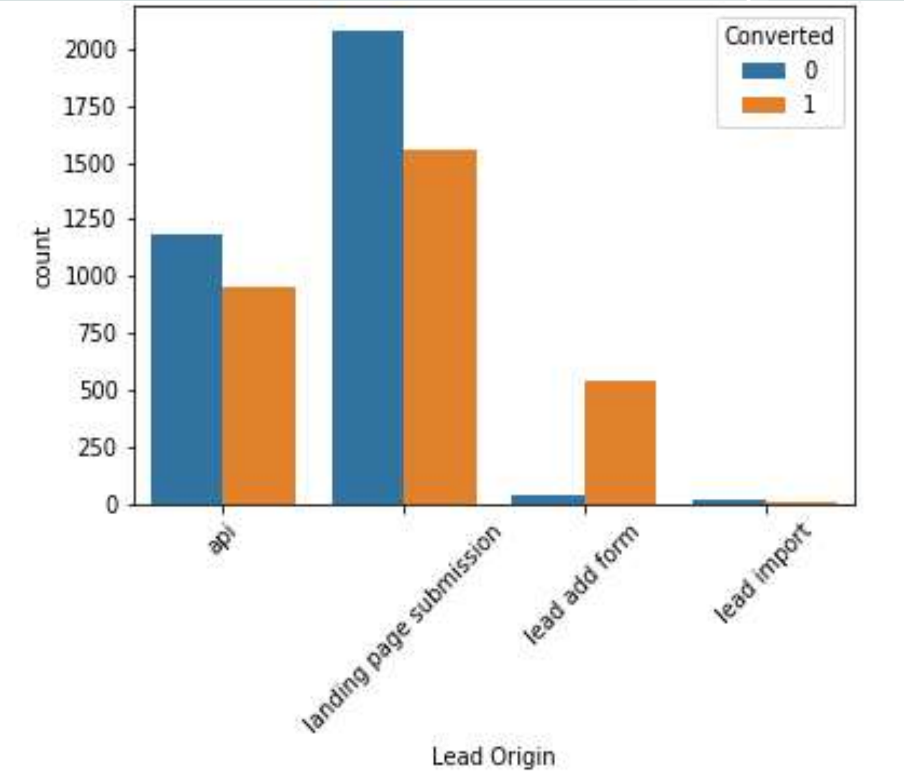
Data Cleaning and Sanitization

- Dropped the columns with only one unique value from the Data Frame.
- Where customer did not select any option, replaced 'Select' as null.
- Removed less relevant fields having null values more than 30% or without condition.
- For Important fields replaced null with 'not Sure’.
- Divided field into less no of categories.

EDA (Exploratory data analysis)

Lead Origin vs Converted

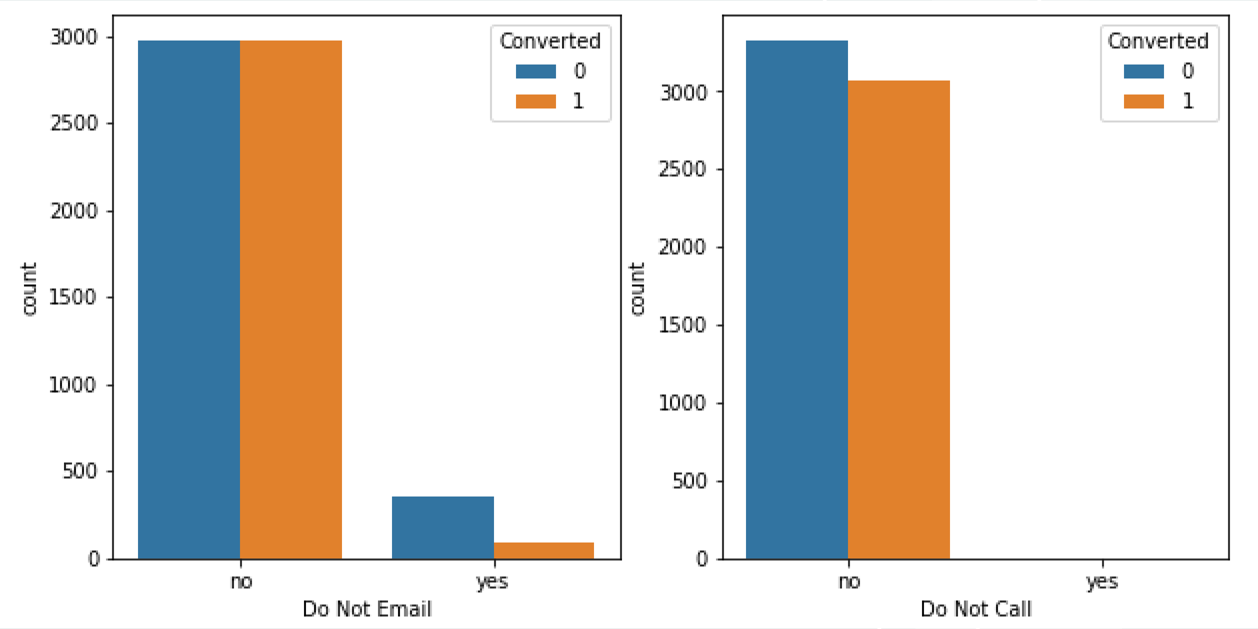
Highlights:
1. 'Landing Page Submission' is having best conversion% followed by API
2. 'Lead Import' is negligible
3. 'Lead add form' conversion is high but lead counts looks minimal
Suggestion:
1. Since 'Landing Page Submission' has best conversion business should focus more on to it
2. Also should focus on increasing leads from 'Lead Add From' that will also add up good conversion



EDA (Exploratory data analysis)

Do Not Email and Do Not Call vs converted

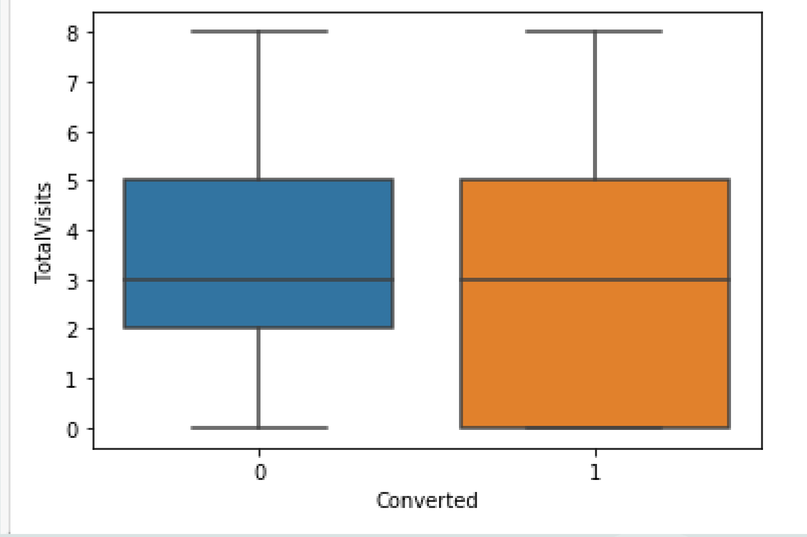
Highlights:
1'Do Not Call' & 'Do Not Email' are not making any significant negative mark on conversion hence it can be ignored



EDA (Exploratory data analysis)

Total Visits

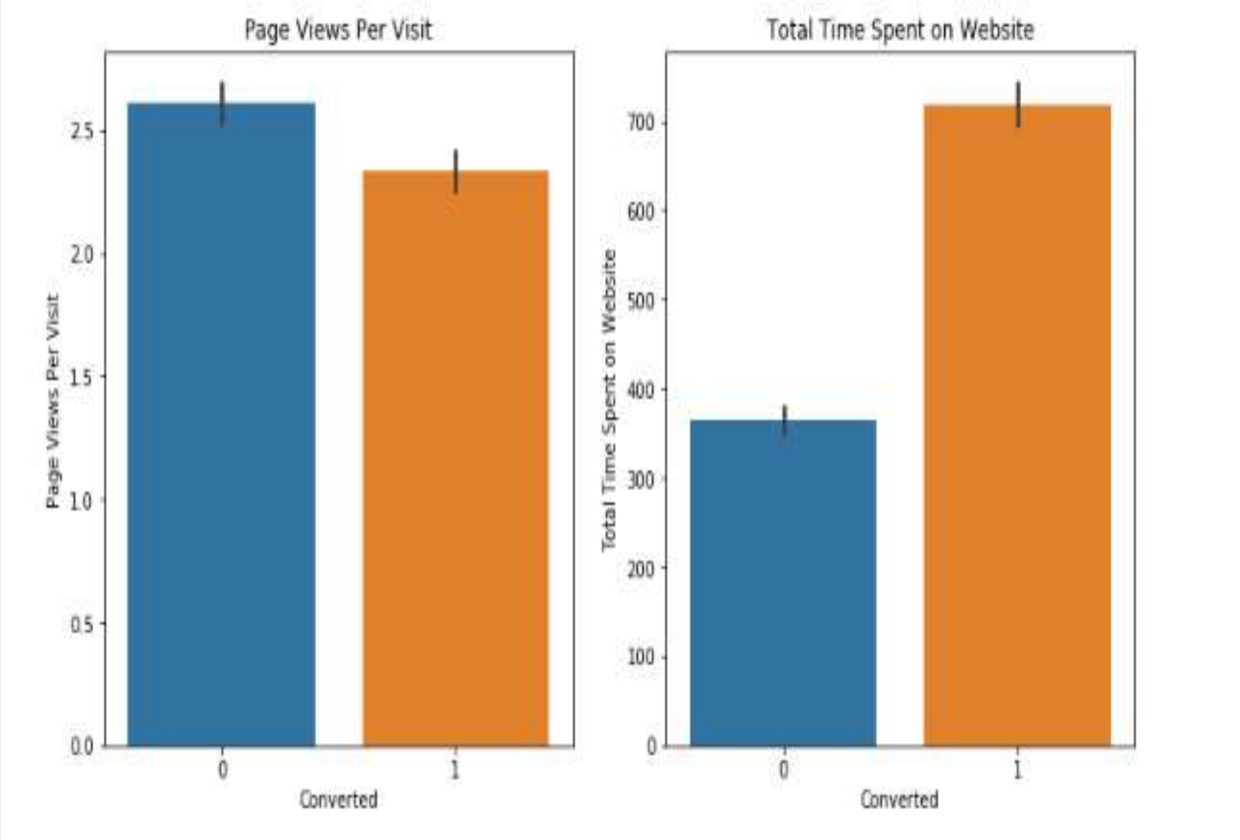
Highlights:
1 Median lines are almost similar for both converted and non converted hence ignore the 'Total Visits'



EDA (Exploratory data analysis)

Total Time Spent on Website & Page Views Per Visit vs Converted

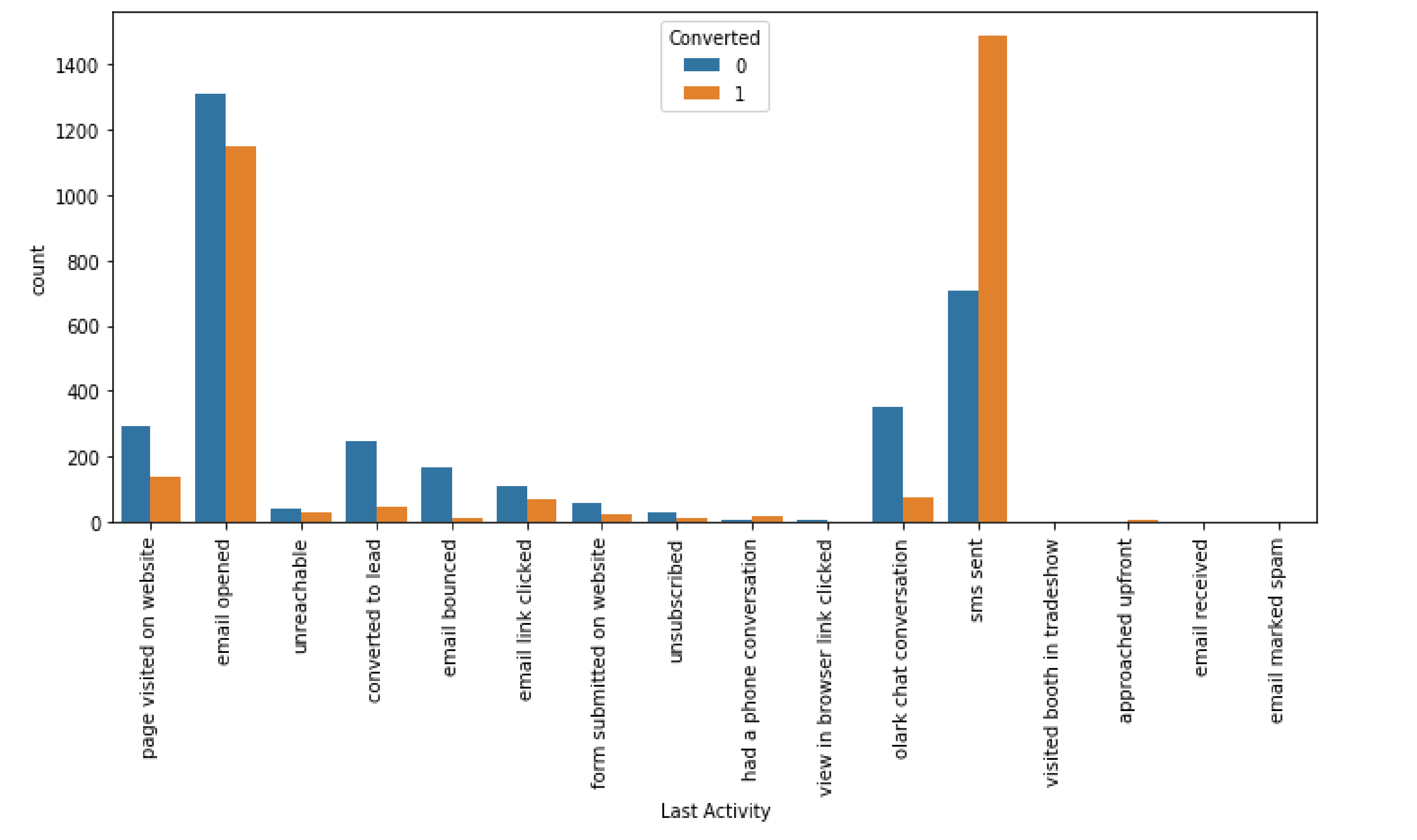
Highlights:
1. 'Page Views Per Visit' are similar
2. Conversion is high on 'Total Time Spent on Website' should increase the website visibility



EDA (Exploratory data analysis)

Last Activity

Highlights:
1. 'email opened' and 'sms sent' are the top 2 activities

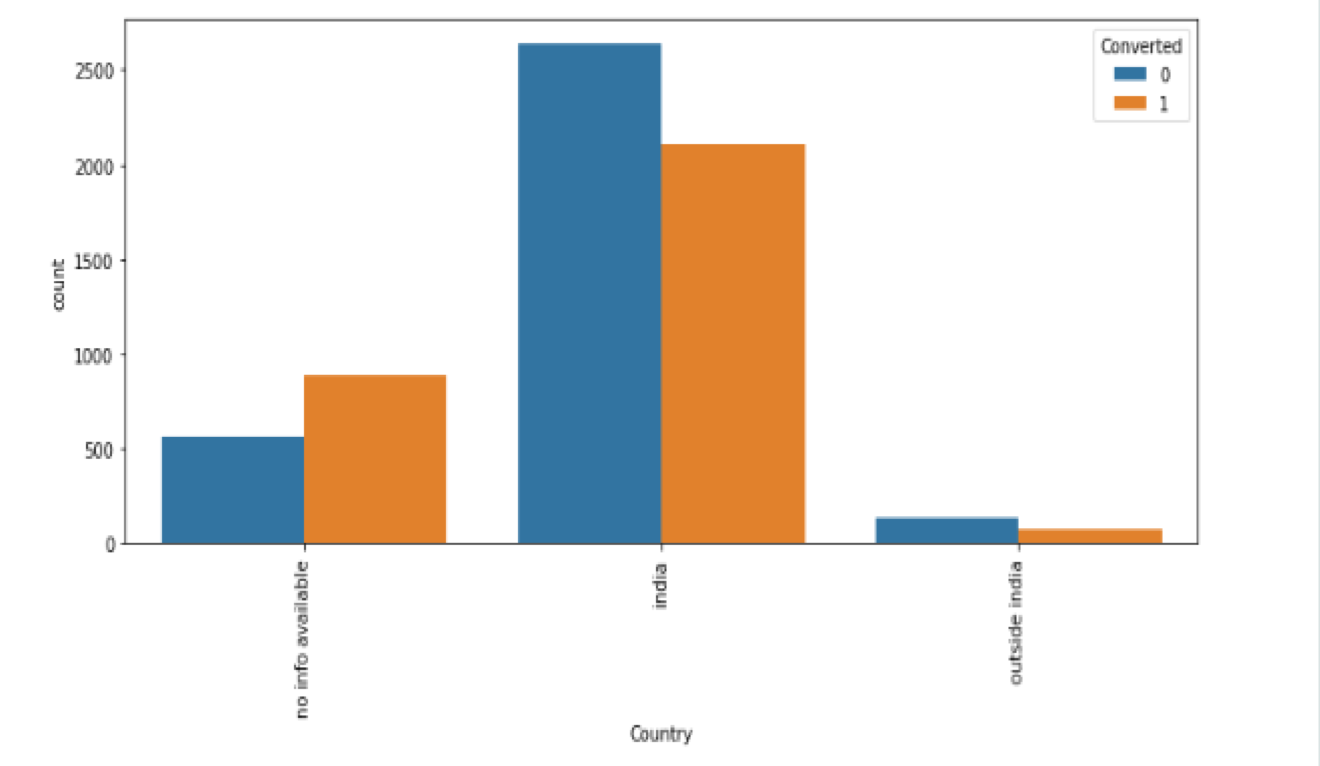


EDA (Exploratory data analysis)

Country

Highlights:

1. India it is, no major conclusion from this



EDA (Exploratory data analysis)

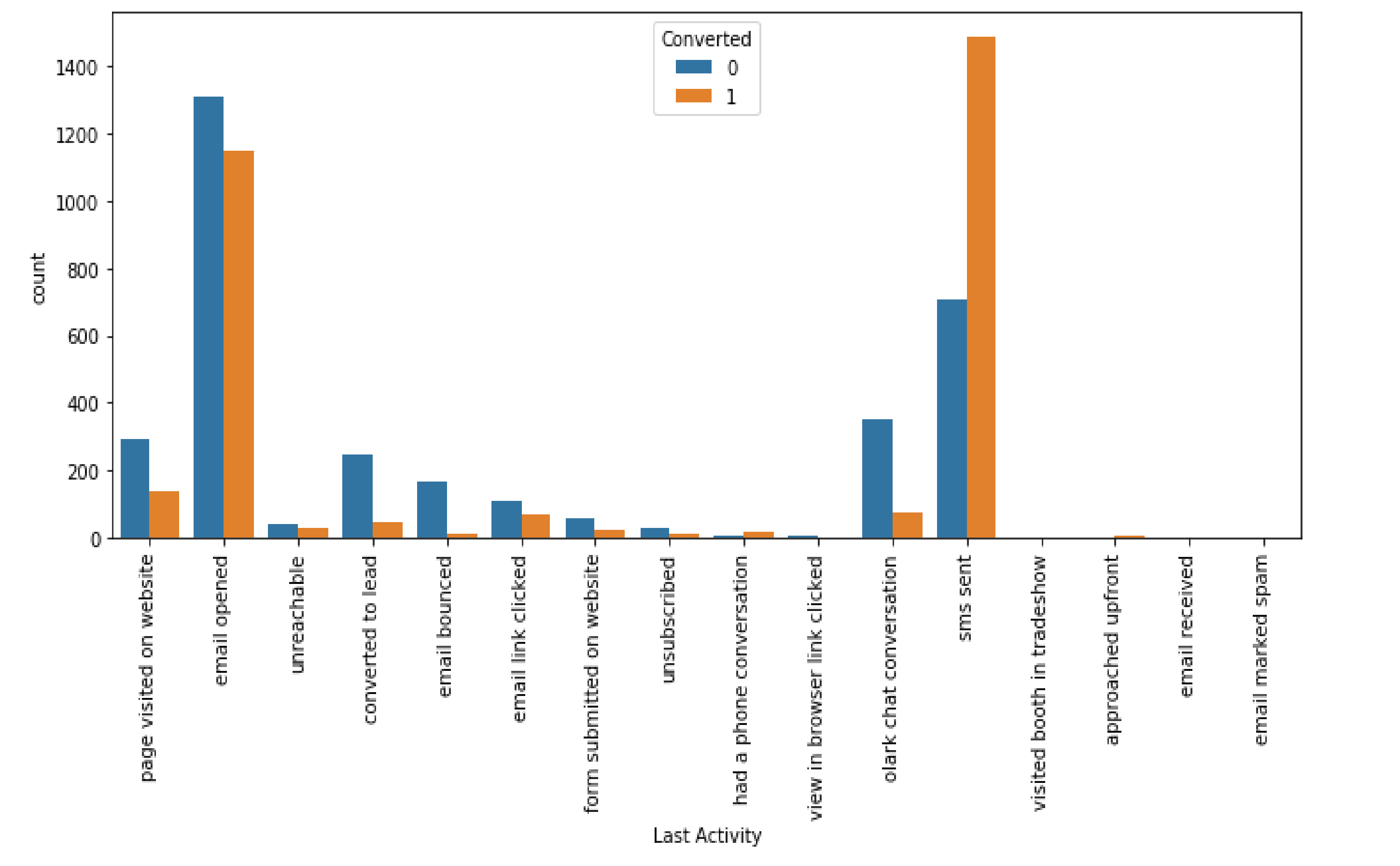
Occupation

Highlights:

1. Conversion is great for working professional
2. Unemployed leads are the top in numbers conversion rate isn't that great.

Suggestion:

1. Should focus for increasing leads on working professionals
2. Also should focus on unemployed conversion for conversion%

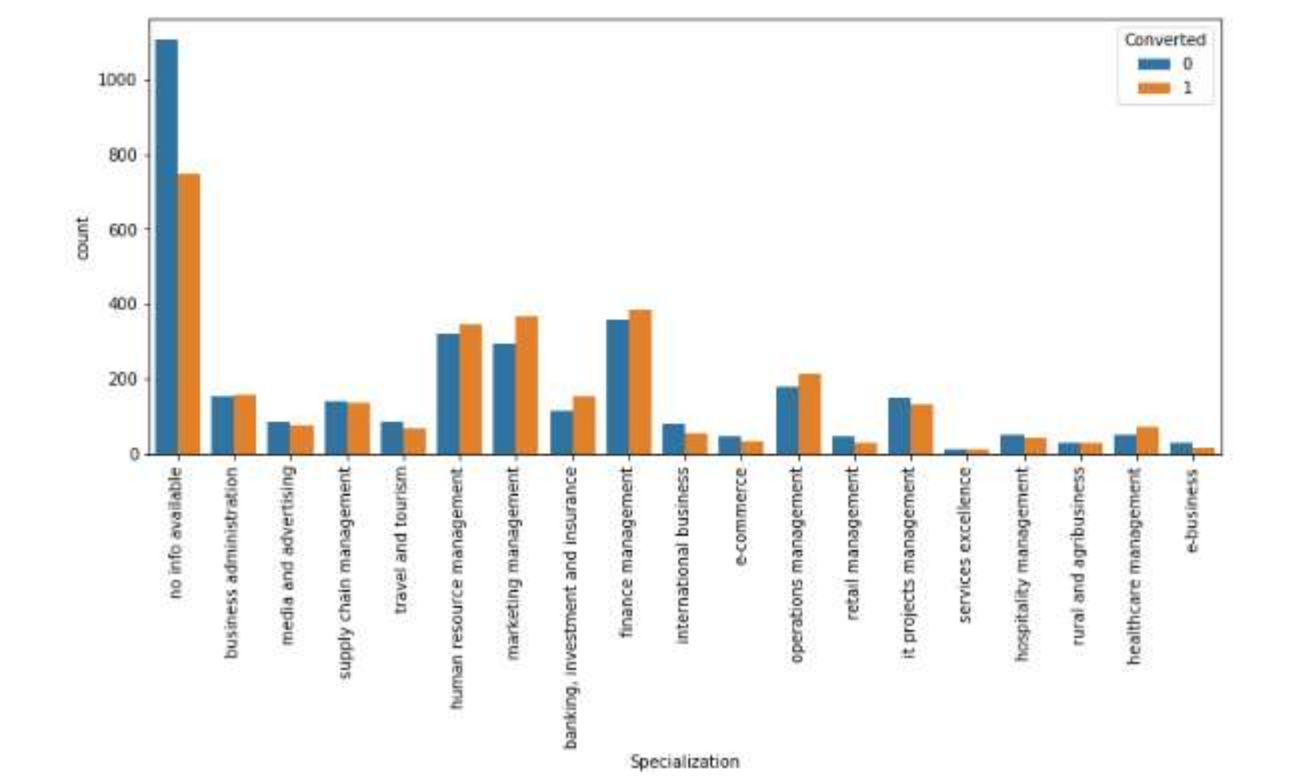


EDA (Exploratory data analysis)

Specialization

Highlights:

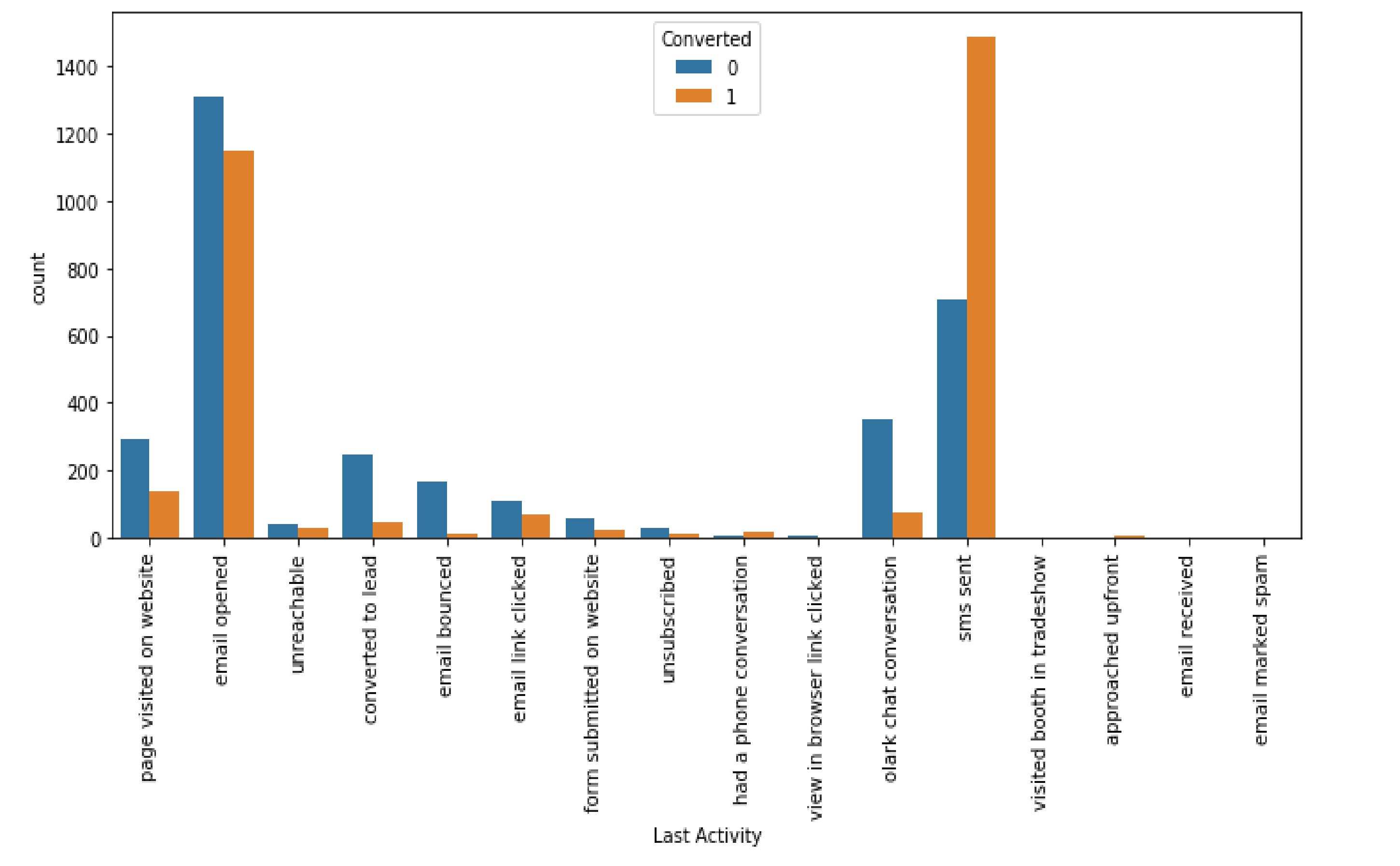
1. There is not any specific 'Specialization' that is popping out but if we consider management as a category, it contributing more
2. Can focus more on management 'Specialization' but before that should deep down the analysis on it



EDA (Exploratory data analysis)

Lead Quality

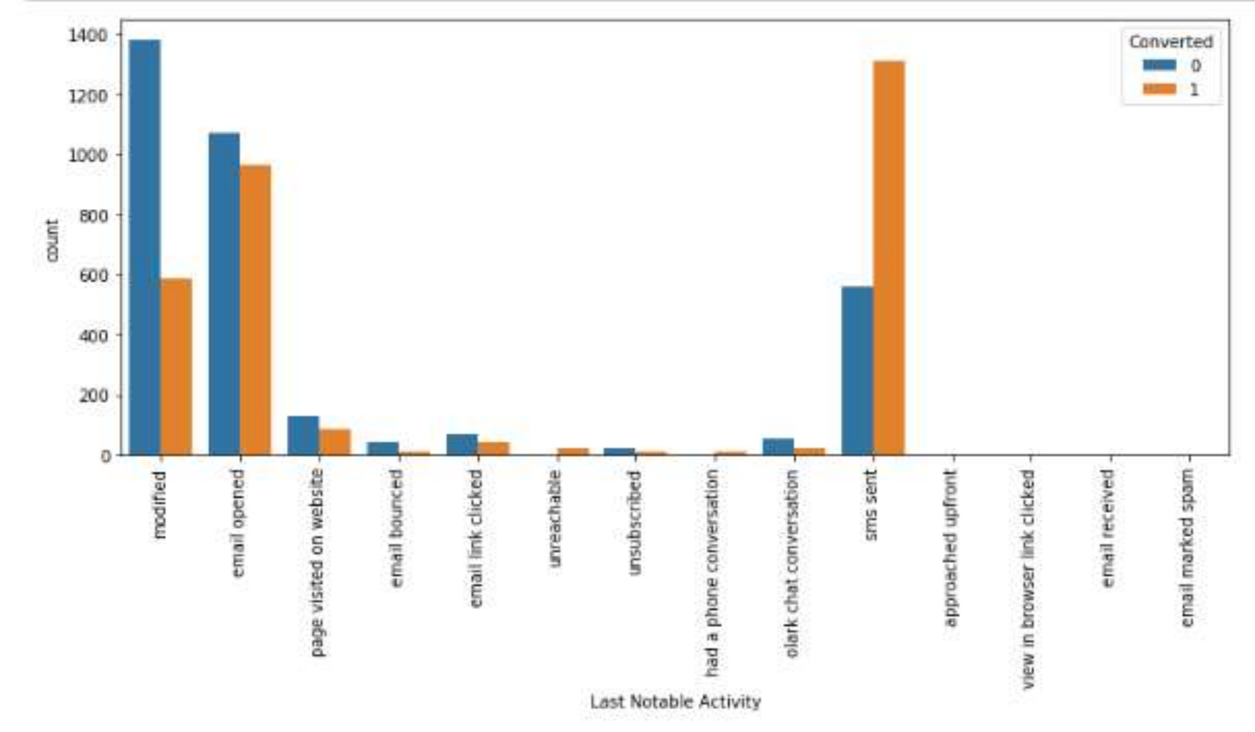
Highlights:
1. No Insight as most cases are 'not sure’



EDA (Exploratory data analysis)

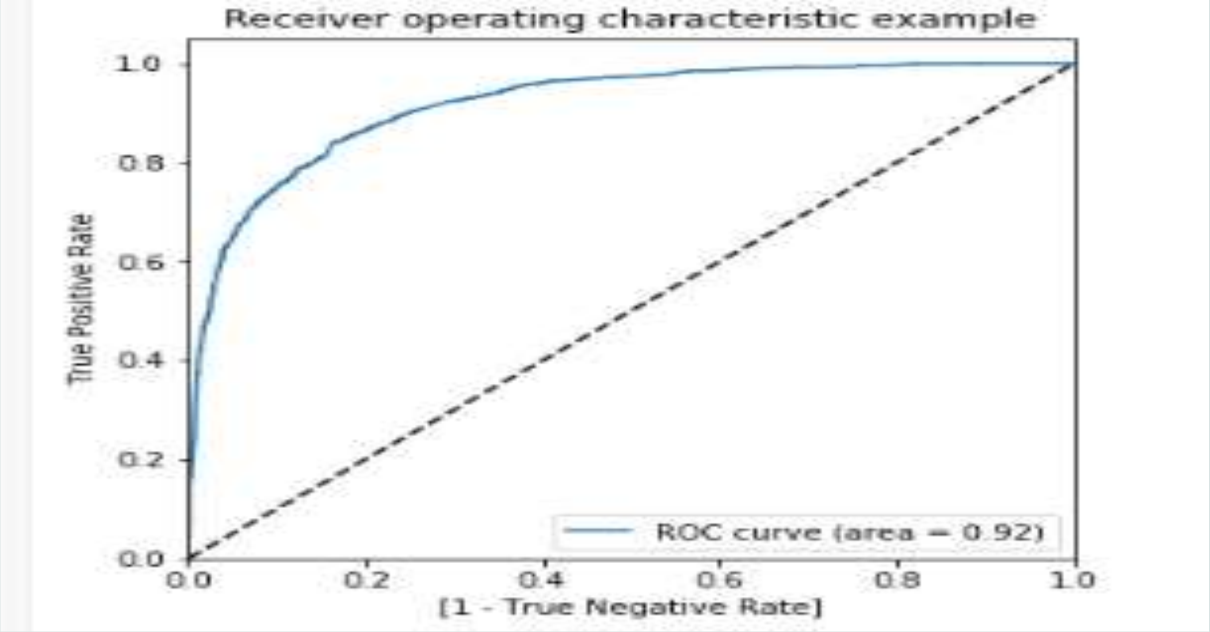
Last Notable Activity

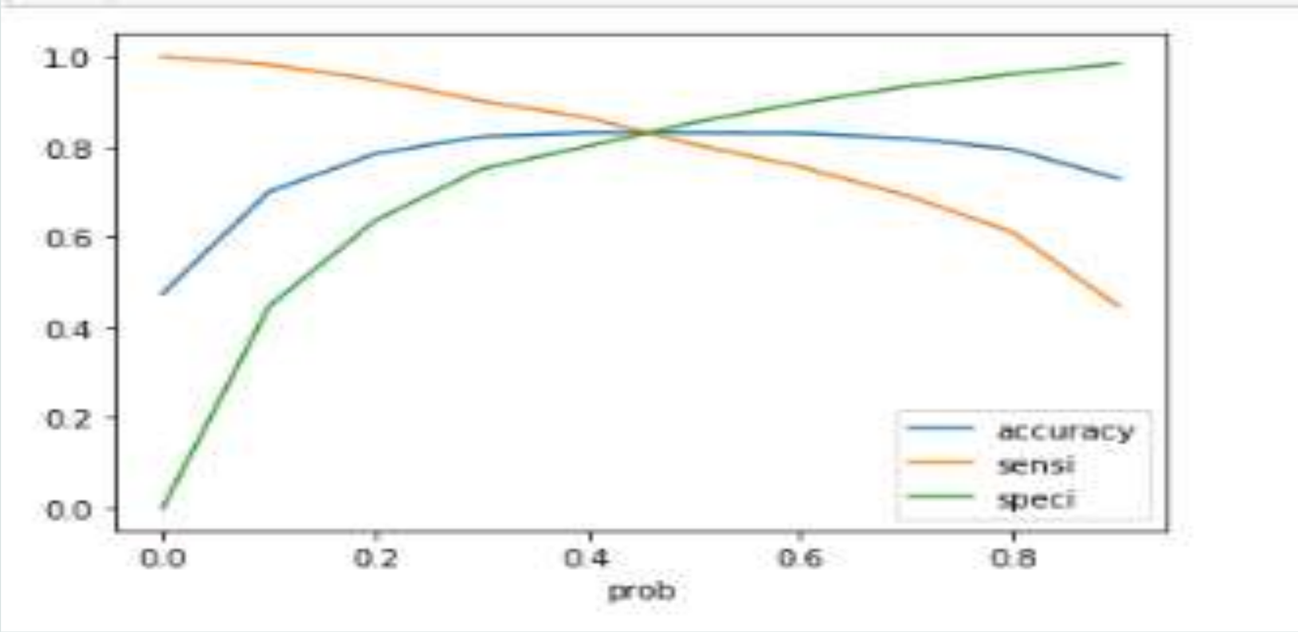
Highlights:
1. 'email opened' and 'sms sent' are the top 2 activities
'modified' has many leads but conversion is not great



Model Building and Evaluation

- Creating Dummy Feature
- Splitting the data into Train and Test set by 30-70 ratio
- Used REF for feature selection (top 15)
- Used variance inflation factor for multicollinearity
- By using VIF & P value removed variable
- Used Confusion metrics
- Overall accuracy, Sensi, Speci >80
- Use ROC curve for more precise cut off
- On optimal cutoff 0.45 - Accuracy, Sensi improved Speci was same





Conclusion and Findings

As per our model variable that are important in converting a lead are:

1. Where Lead Origin is 'lead add form'

2. Where Lead Source are:

- olark chat’
- 'welingak website'

3. Occupation is 'working professional'

4. Last Notable activity is 'Modified'

5. Last Activity is 'SMS Sent'

6. Lead Quality is 'Not Sure' and 'might be' (not sure are the cases where sales team did not fill anything)

7. Also some less important variables are that are behaviorally looking off:

- Last Activity email bounced
- Lead Quality worst
- Last Notable Activity unreachable

Thank you

Vijay Kumar Jha
EPGPDS IIITB & UPGRAD
August 2023