Lead Score Case Study

- Mushir Malek

❖ Problem Statement

- An education company named X Education sells online courses to industry professionals.
- Many professionals who are interested in the courses land on their website and browse for courses
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if,
- say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most
- potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, 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.

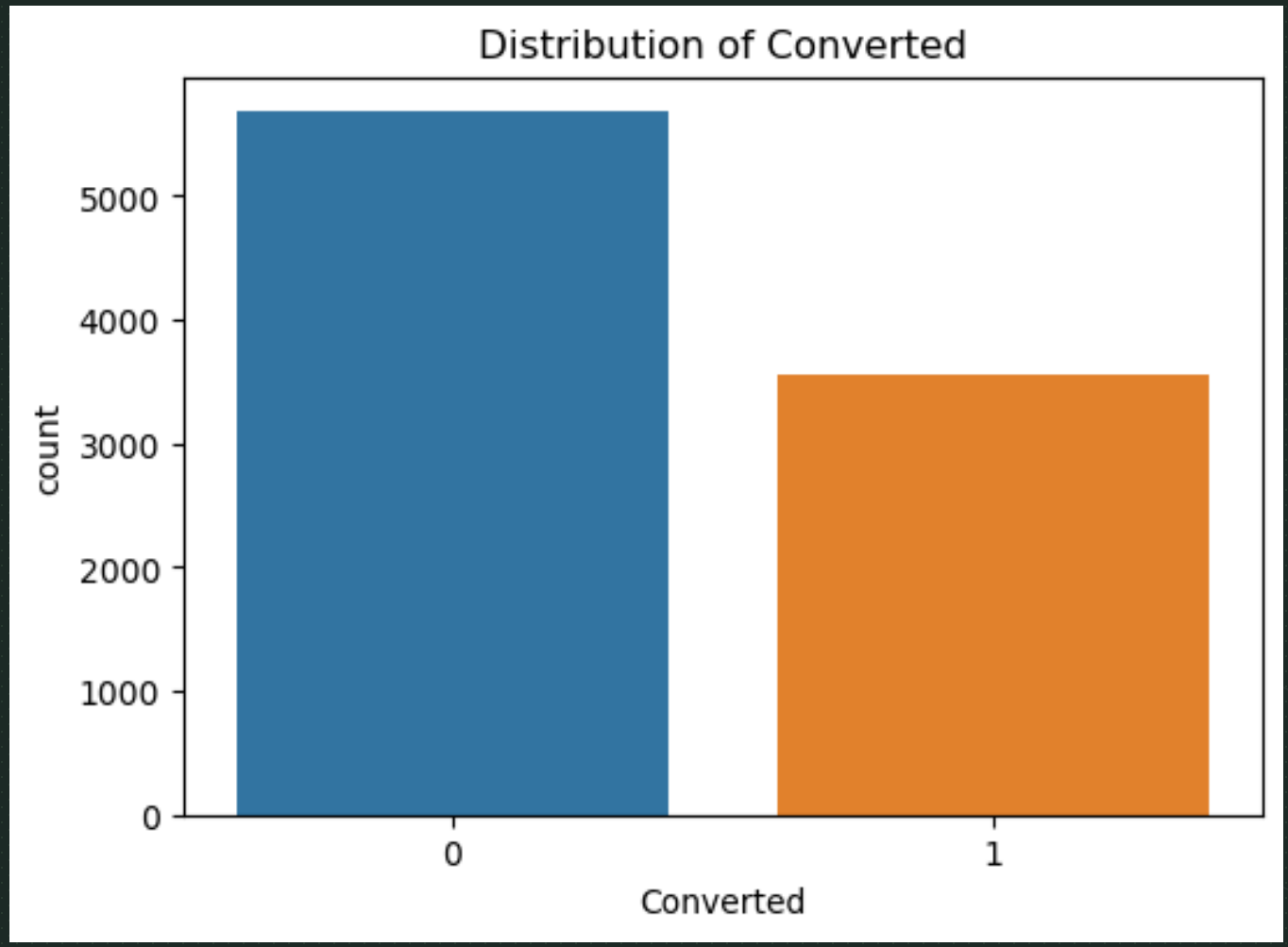
Data cleaning and data manipulation

- Check and handle NA values and missing values.
- Drop columns, if it contains large amount of missing values and not useful for the analysis.
- Imputation of the values, if necessary. 5
- Check and handle outliers in data.

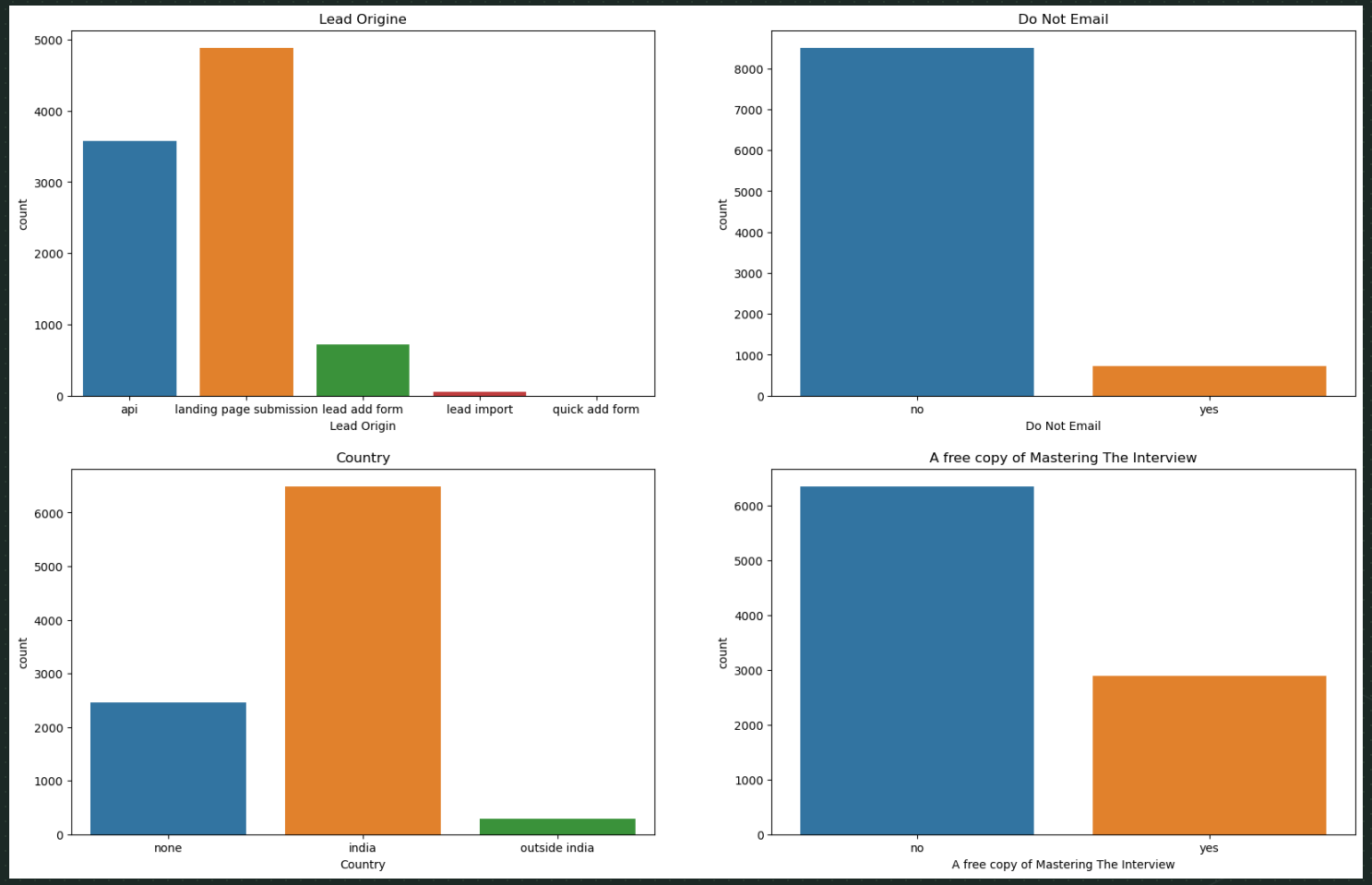
❖ EDA

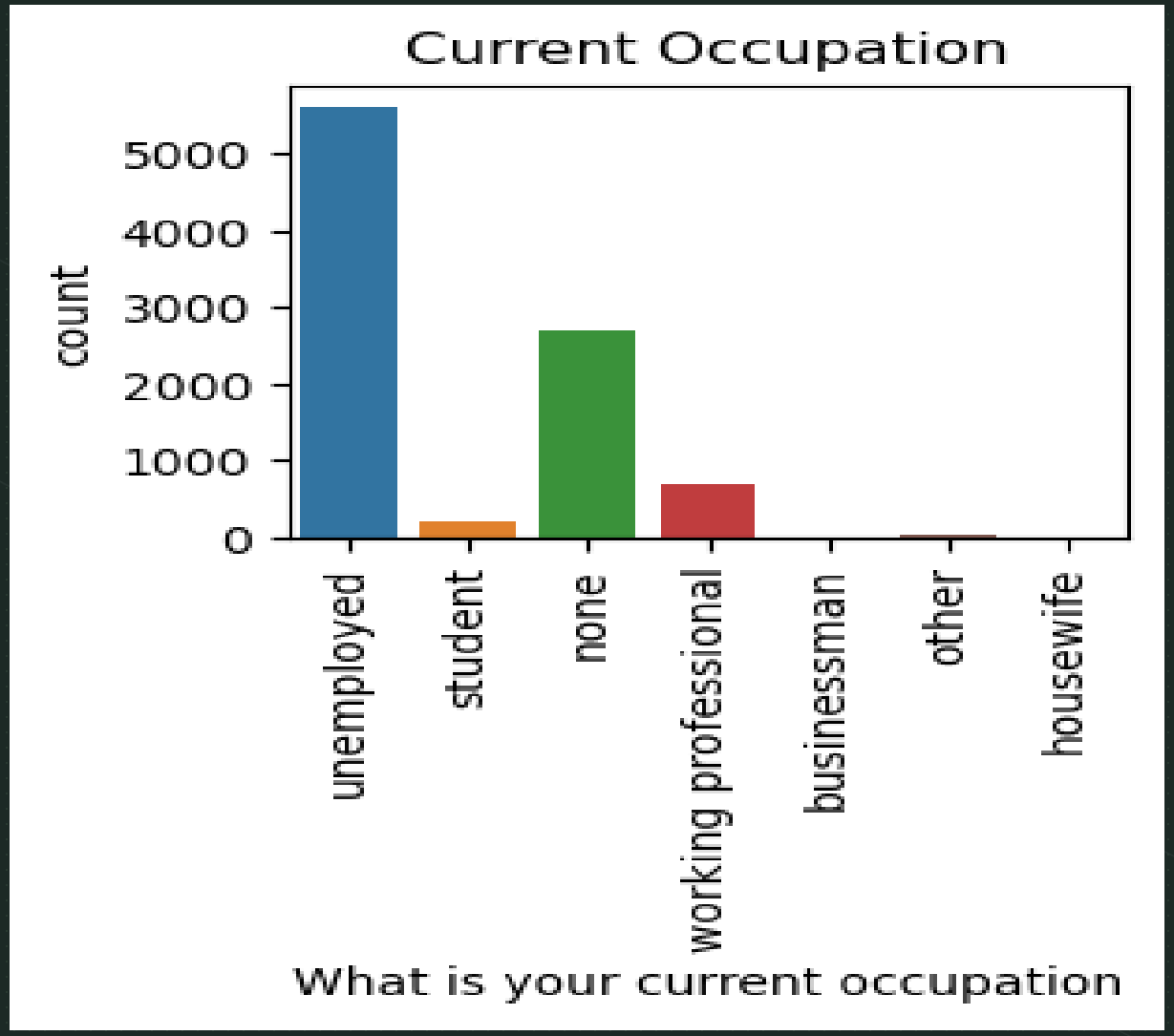
- Value count, distribution of variable etc.
- Correlation coefficients and pattern between the variables etc.
- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

❖ EDA

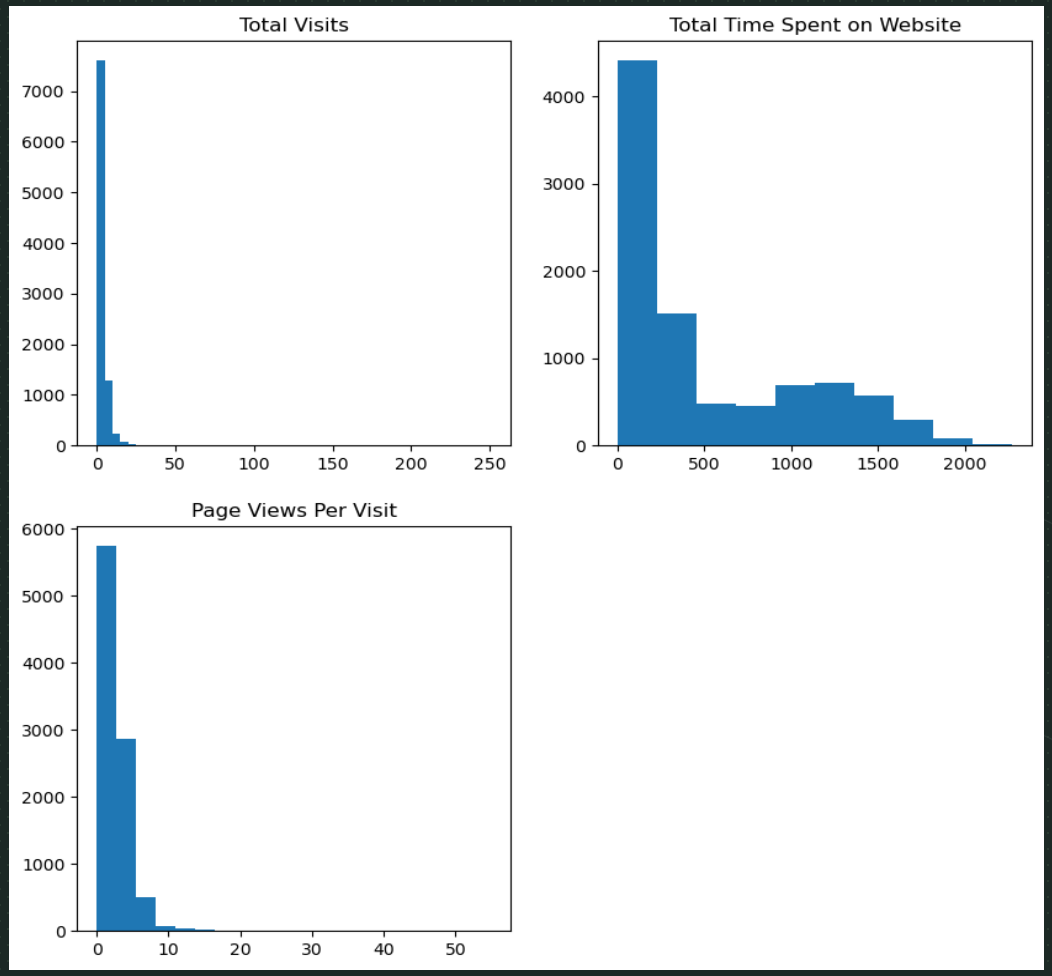

Distribution of Converted
count
5000
4000
3000
2000
1000
0
0
1
Converted

❖ Categorical Variables

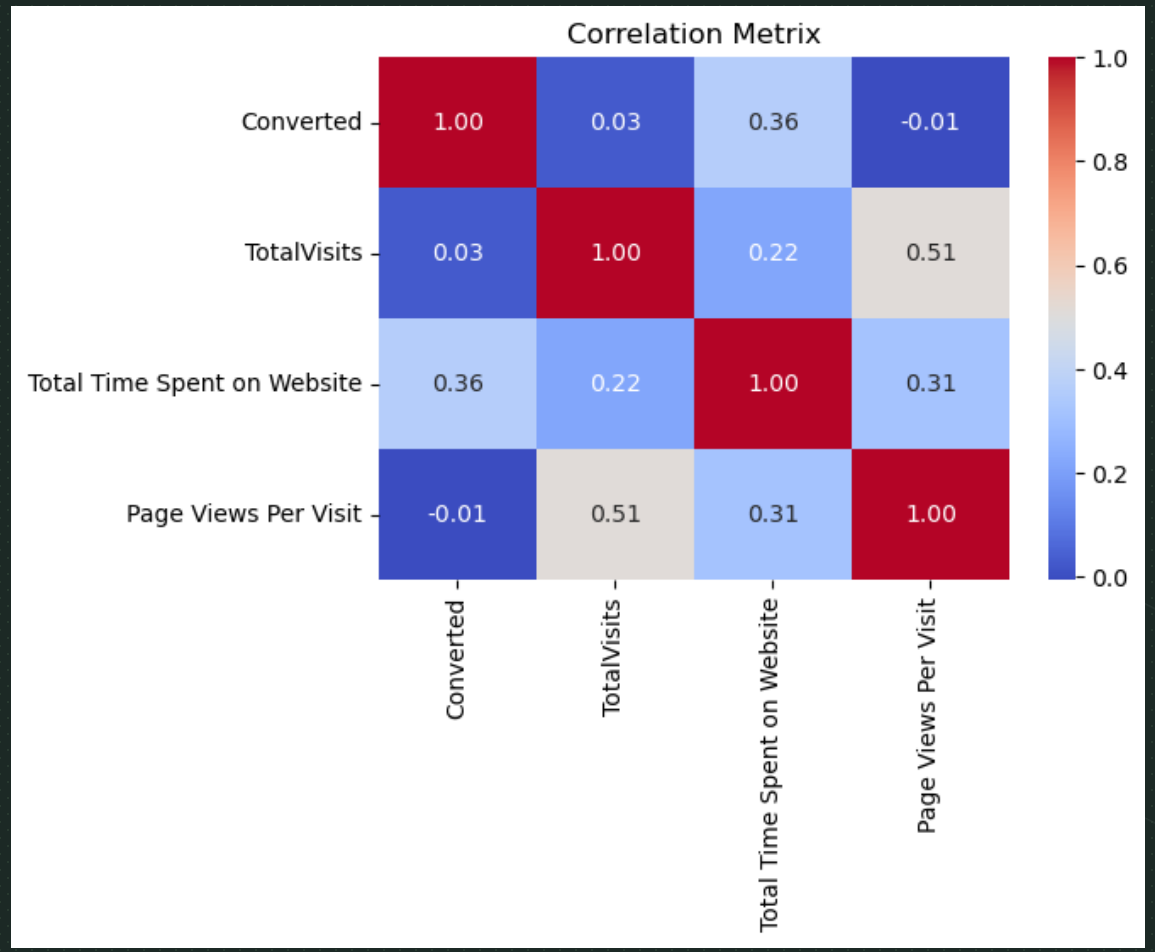



Current Occupation
count
5000
4000
3000
2000
1000
0
unemployed
student
none
working professional
businessman
other
housewife
What is your current occupation

❖ Numerical Variables



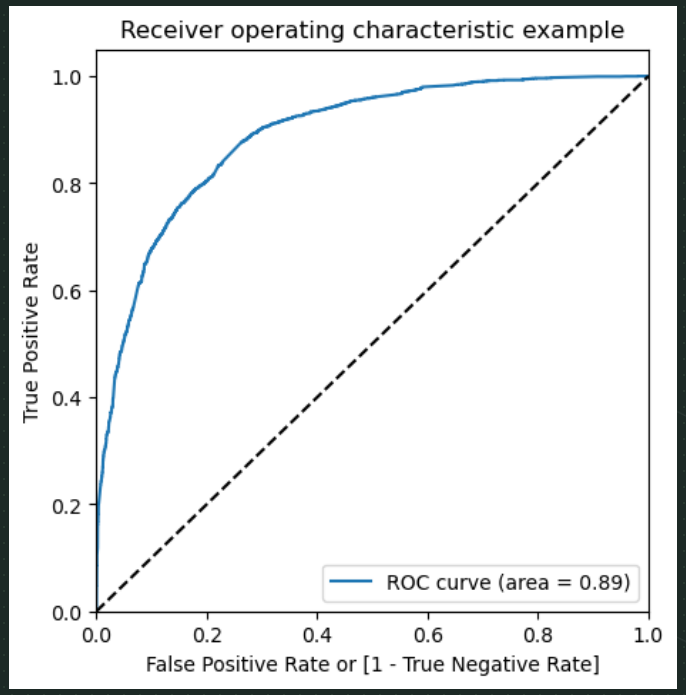
❖ Correlation Metrix.

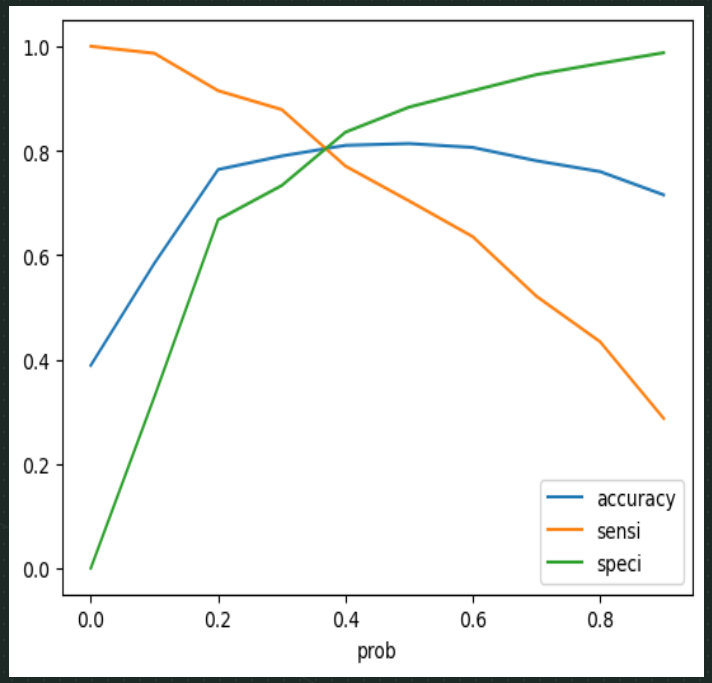


❖ Model Building

- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30
- ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output
- Building Model by removing the variable whose p- value is greater than 0.05 and vif
- value is greater than 5
- Predictions on test data set
- Accuracy 81%

❖ ROC Curve





- Finding Optimal Cut off Point
- Optimal cut off probability is that
- probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.