Lead Score case study

GROUP MEMBERS:
1.Sravya Simhadri
2.Ananya Utkarsh

Problem Statement

- X Education sells online courses to industry professionals
- X Education gets a lot of leads, but 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’.
- 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.

Business Objectives

- X education wants to know most promising leads.
- For that they want to build a Model that identifies the hot leads.
- Deployment of the model for the future use.

Solution Approach/Methodology

- Data cleaning and manipulation.

1. Check and handle duplicate data.
2. Check and handle NA and missing values.
3. Drop columns, if it contains large amount of missing values which are not useful for the analysis
4. Imputation of the values, if necessary.
5. Handling outlier in data

- EDA analysis

1. Univariate data analysis: value count, distribution of variable etc.
2. Bivariate data analysis: distribution/relation between two variables etc.
3. Multivariate data analysis: correlation coefficients and pattern between the variables etc.

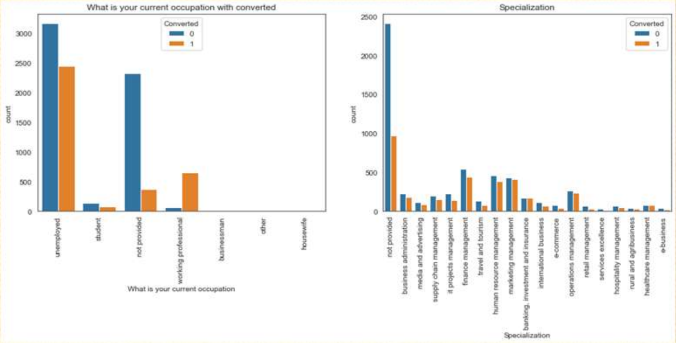
Solution Approach/Methodology

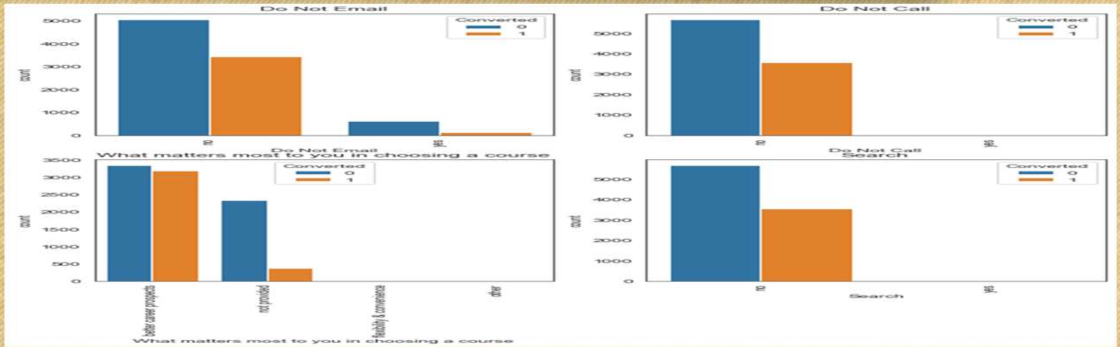
- Feature Scaling & Dummy Variables and encoding of the data
- Classification technique: logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

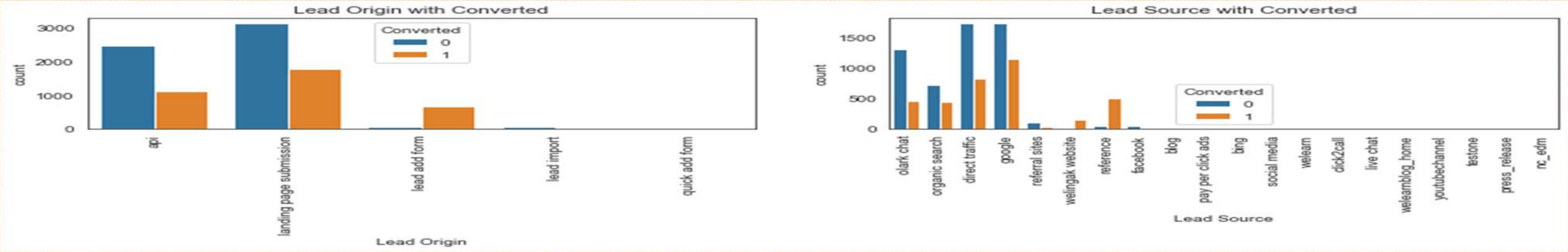
Data Cleaning and Manipulation

- Total Number of Rows =37, Total Number of Columns =9240
- Removing columns :How did you hear about X Education, Lead Profile, Lead Quality, Asymmetrique Profile Score, Asymmetrique Activity Score, Asymmetrique Activity Index, Asymmetrique Profile Index as they had more than 45% data/values missing
- City and tags columns does not have any useful information and more than 35% columns are null values so we can drop them
- Dropping prospects id and Lead number as they are having only unique ID values.
- 'Magazine', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque’ as they have very less unique values which won’t help in analysis
- Imputing high null value data with not provided in columns like: 'Specialization’, 'Country’, 'What is your current occupation’ etc.

EDA

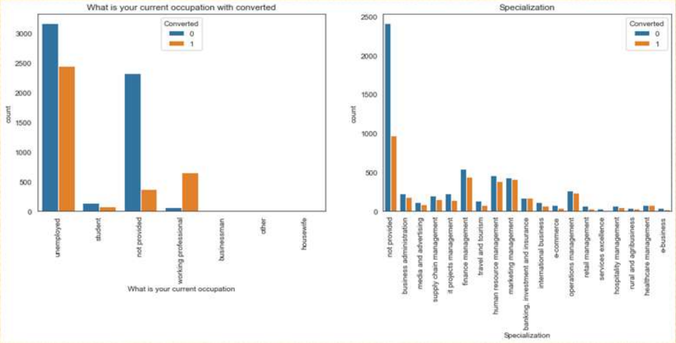


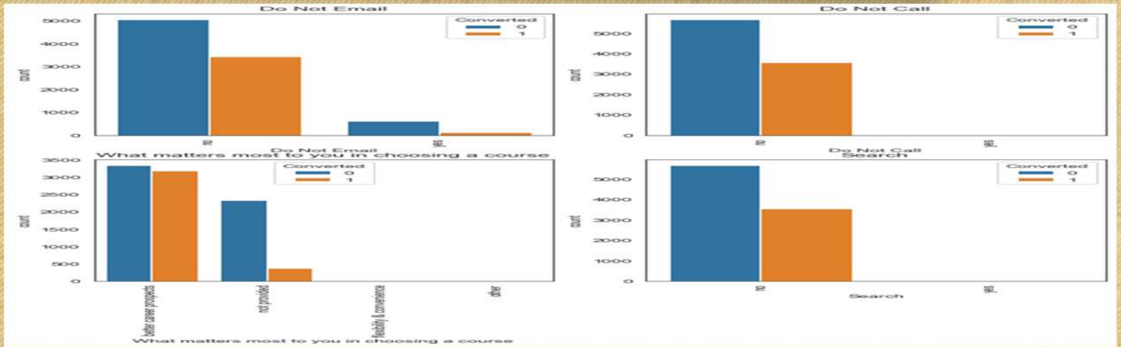


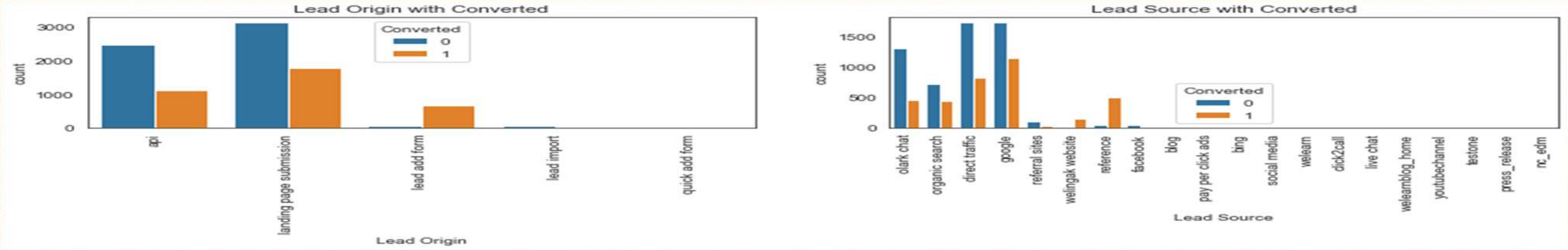


We can see from eda analysis how values of different variables/features correspond to conversion of customer

EDA

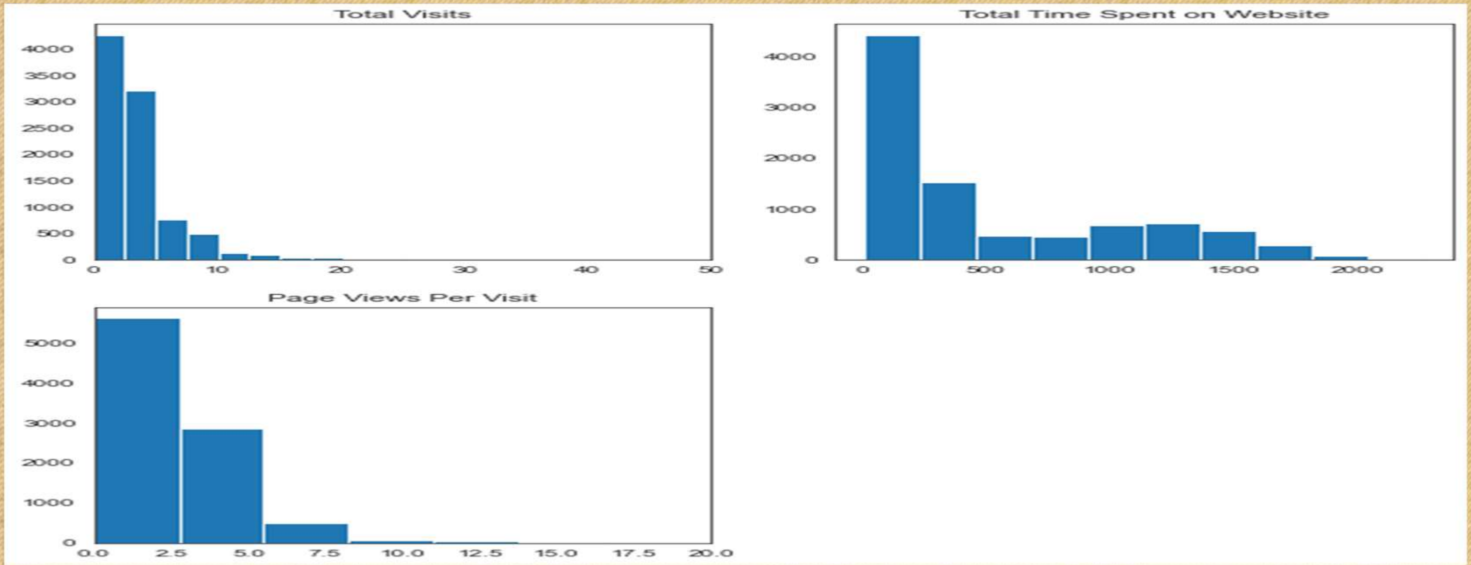






We can see from eda analysis how values of different variables/features correspond to conversion of customer

EDA



Checking for the distribution of values in a particular column/feature

Data conversion and dummy variable creation

- Creating dummy variables for categorical data
- Dropping redundant/duplicate columns from which dummy variables were created

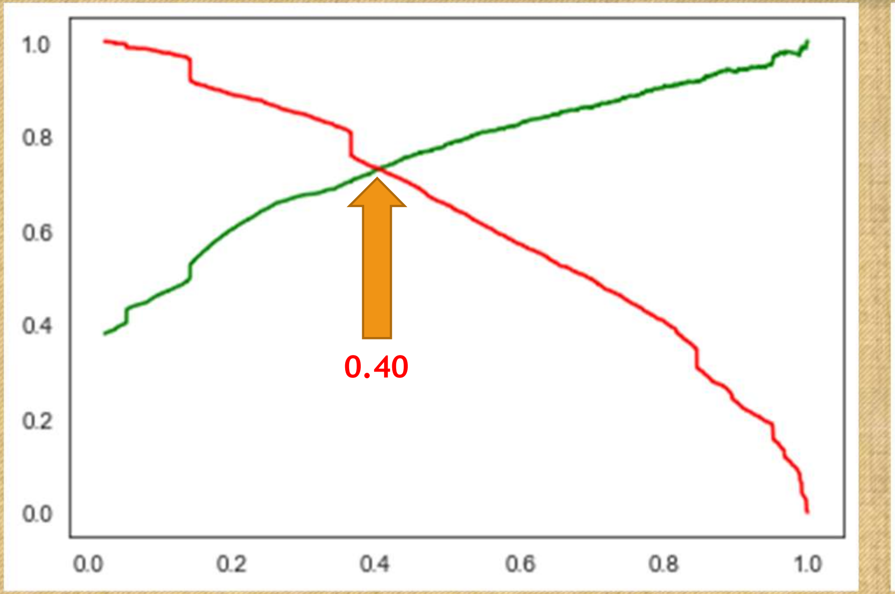
Module building and Evaluation

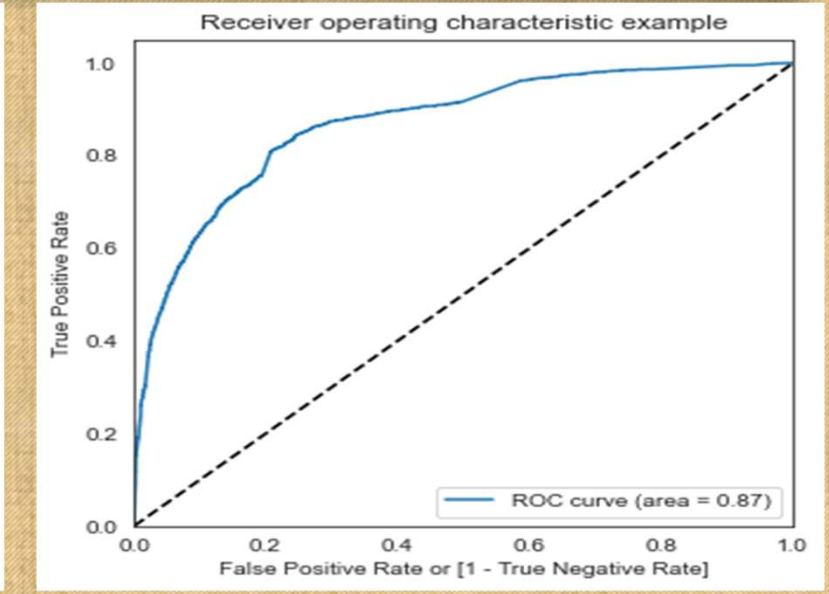
- Splitting the Data into Training and Testing Sets
- The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- Use RFE for Feature Selection
- Running RFE with 15 variables as output □
- Finalizing Model by removing the variable whose p-value is greater than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

  accuracy score(Training dataset):0.795

  accuracy score(Test dataset):0.7875

ROC CURVE





Finding Optimal Cut off Point :
Optimal cut off probability is the probability where we get balanced sensitivity and specificity. From the first graph it is visible that the optimal cut off is at **0.40**

Conclusion

The major contributing variables are:

- 1.The total time spend on the Website.
- 2.Total number of visits.
- 3.When the lead source was google, direct traffic and organic search.
- 4.When the last activity was SMS and Olark chat conversation.
- 5.When their current occupation is unemployed.
- 6.And it is better to send emails and sms to the customers rather than direct calls.