Lead Scoring Case Study

Group Members

1. Gaurav Kumar Singh
2. Shrishti
3. Ankur

Problem Statement

- X Education Sells online courses to industry professionals.
- Once leads are acquired, marketing team approach them.
- X education gets a lots of leads but its conversion rate is very poor. For eg. Out of 100 leads only 30 gets 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.

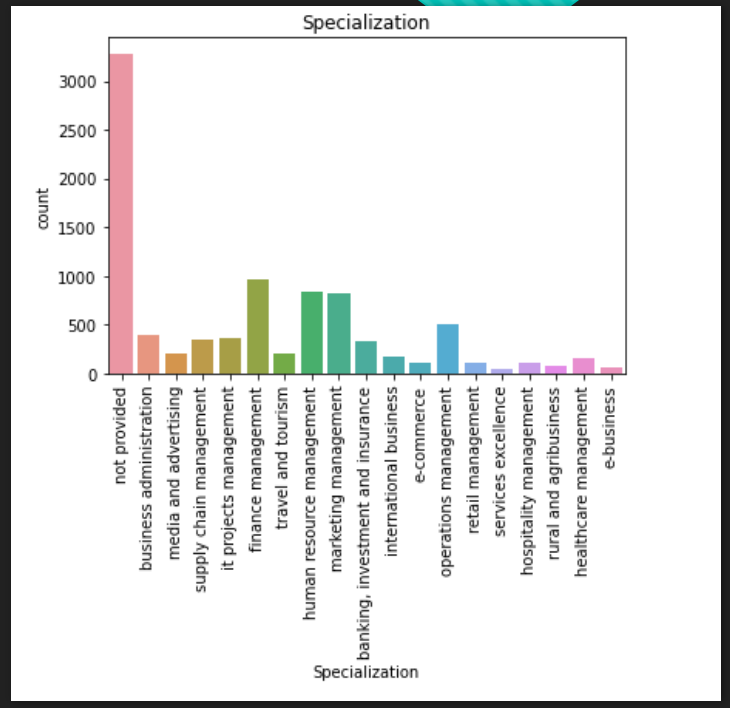
Steps

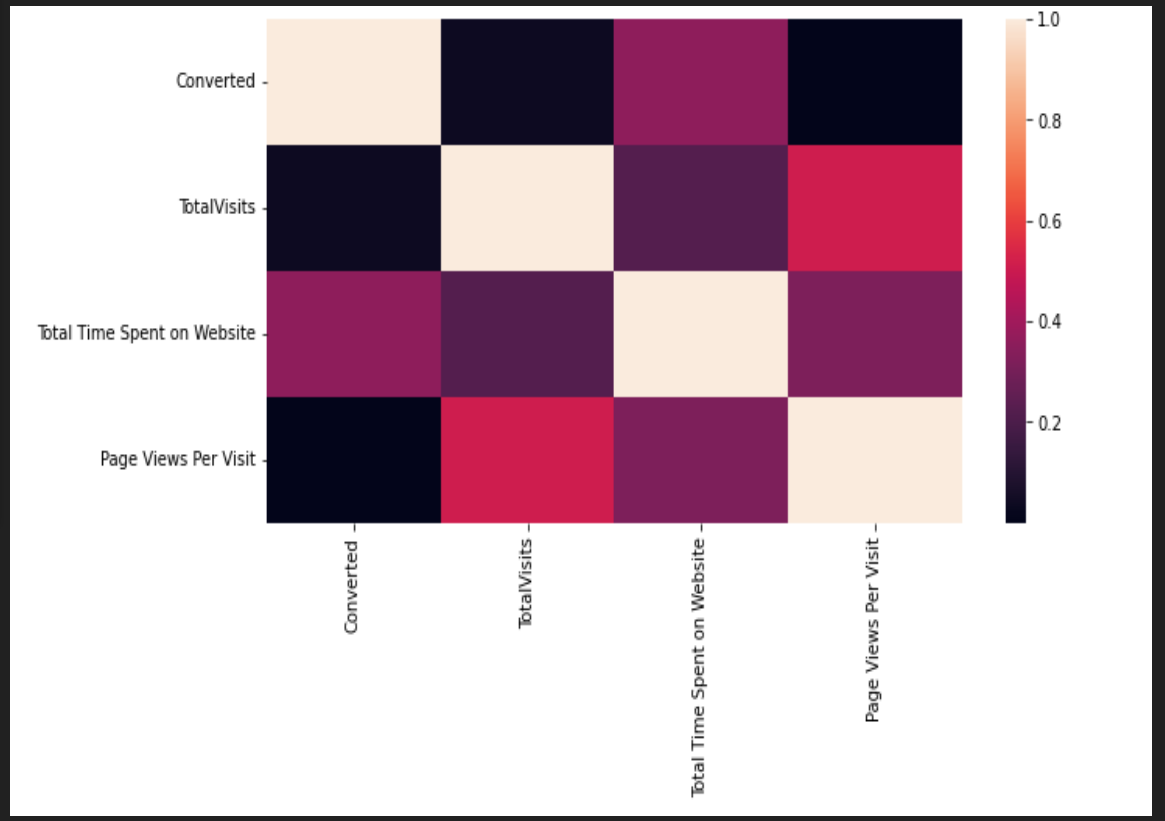
- Data Understanding
  - Load the data and understand various aspects of data
- Data Cleaning and Manipulation
  - Handle duplicate data
  - Handle missing values
  - Drop unnecessary columns
  - Imputation of values
  - Check and handle outliers in data
- Exploratory Data Analysis
  - Univariate Analysis: value counts and distribution of variables
  - Bivariate Analysis: Correlation between variables
- Feature Scaling and Dummy variables
- Model Building
  - Logistic Regression Model building
- Validation of Model through test set
- Model Presentation
- Conclusions and recommendations

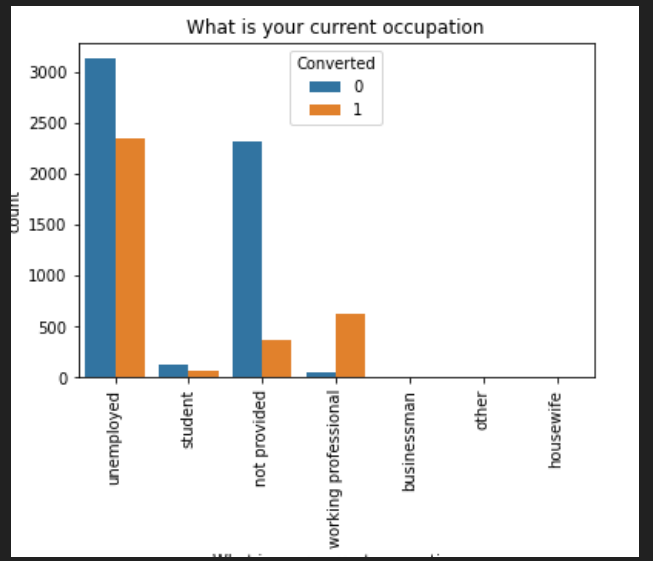
Data Cleaning and Manipulation

- Total number of rows and columns: 9240x37
- Removal of single value feature as they don't add much significance to the model like Magazine, Get updates on DM content etc.
- Remove the columns having more than 35% of missing values
- Filling rest NAN values as 'not provided'
- Removing and Prospect ID and Lead Number
- Removed columns having not enough variance such as Do no Call, Digital advertisement etc

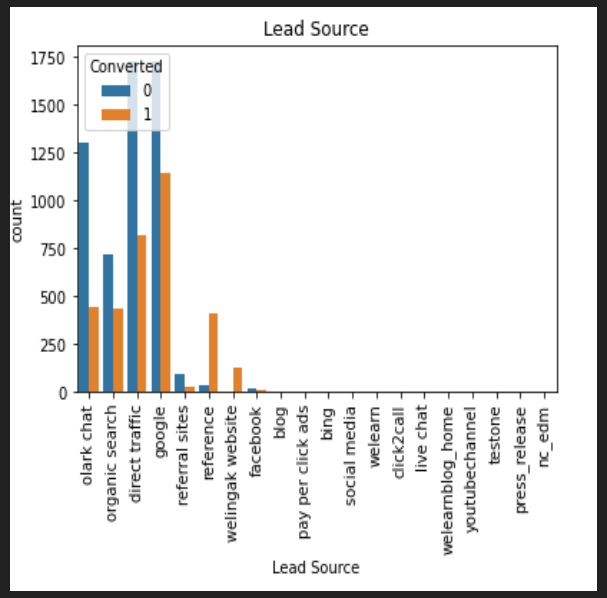
Exploratory Data Analysis

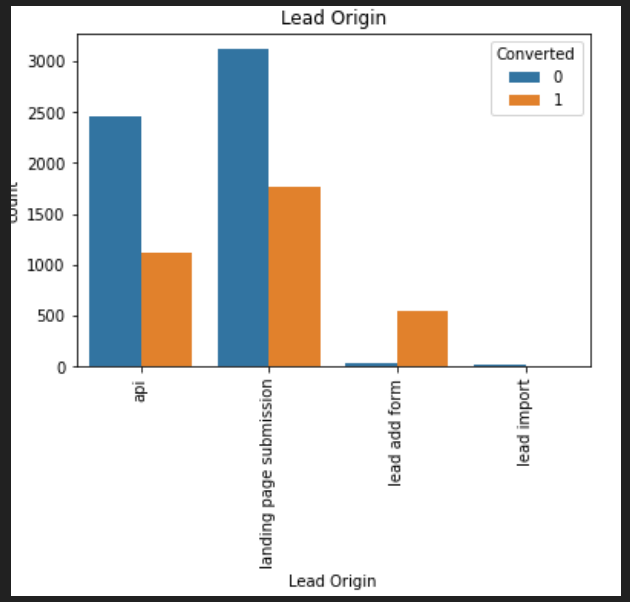


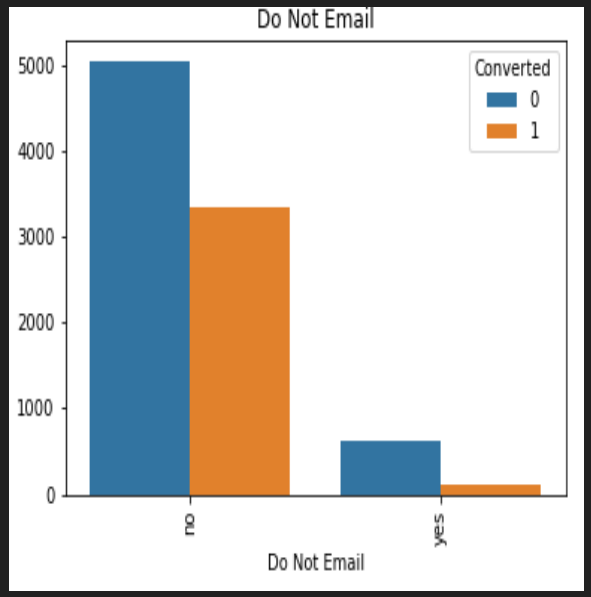


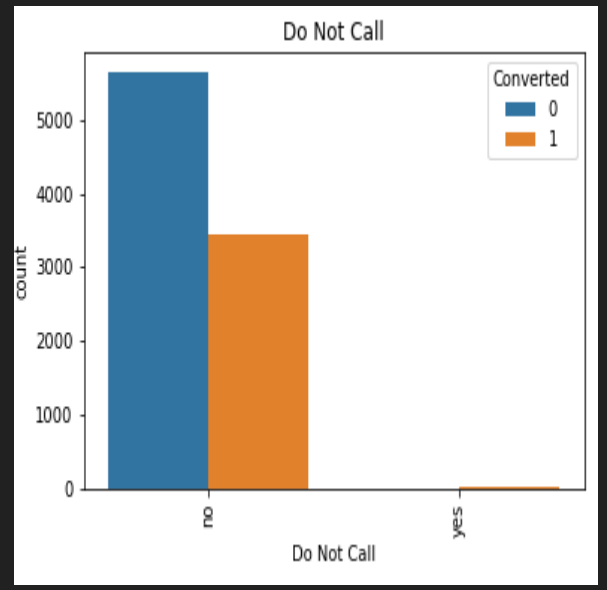


Categorical Variable





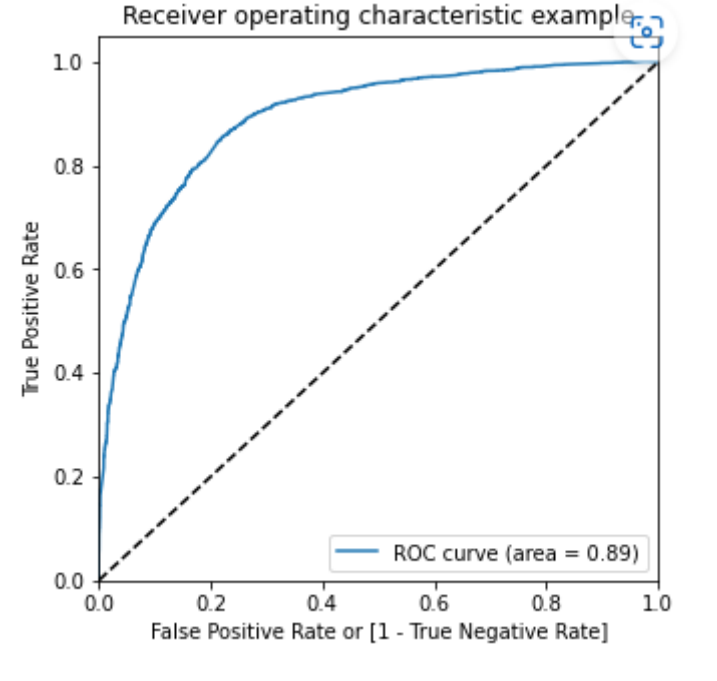




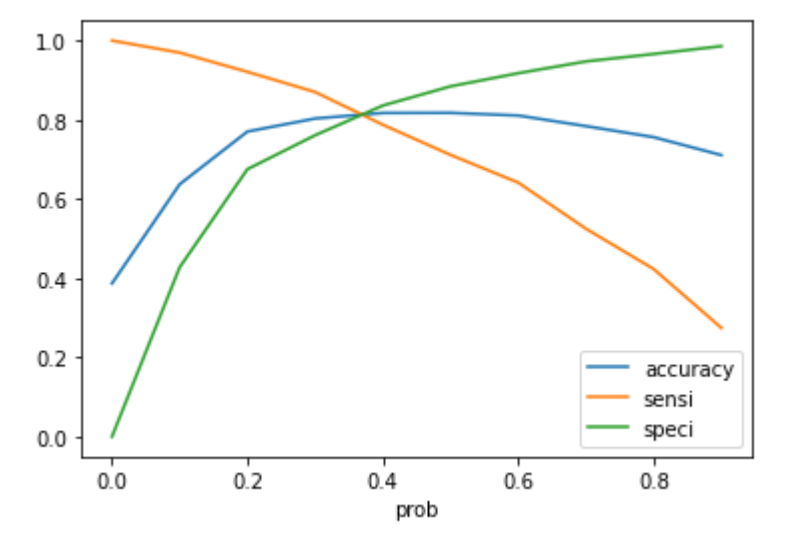
Model Building

- Splitting the data set into Train and Test Set in 70:30 ratio
- Scaling the variables using Min-Max Scaler
- Using Recursive Feature Selection to select variables with 15 variables as output
- Building the model by removing variables on the basis of p-values and VIF
- Prediction on Test Set
- Overall Model Accuracy: 81%

ROC Curve



- Finding optimal cutoff point at 0.35



Conclusions and Recommendations

Important Variables:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:
  a. Google
  b. Direct traffic
  c. Organic search
  d. Welingak website
- When the last activity was:
  a. SMS
  b. Olark chat conversation
- When the lead origin is Lead add format.
- When their current occupation is as a working professional.

Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.