Lead Scoring
Case Study

Dhara KHAMAR

Deepanshu BENIWAL

Deepak GUPTA

[DS C52 – JAN’23 Batch]

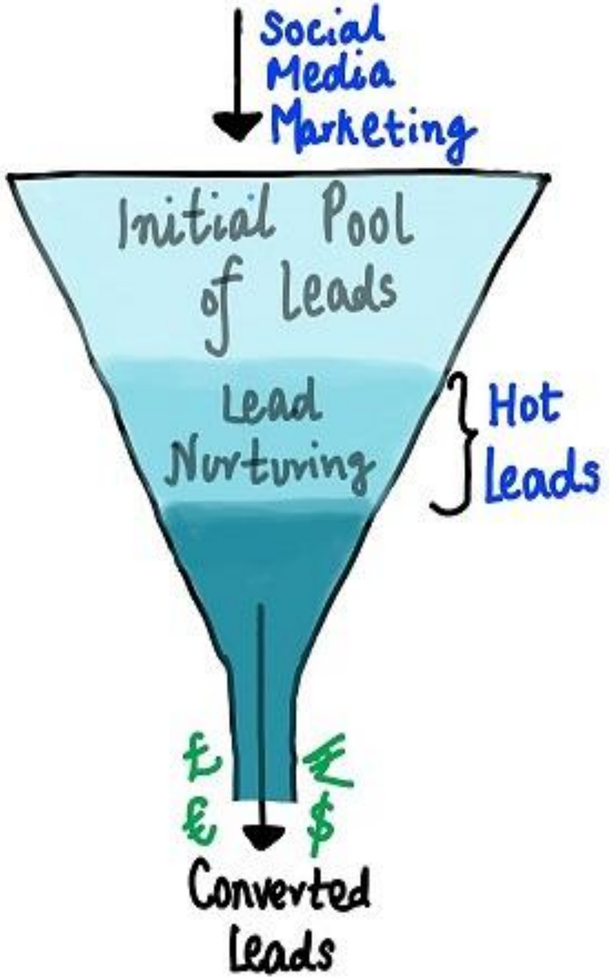
Problem Statement

❑ An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

❑ The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos.

❑ When these people fill up a form providing their email address or phone number, they are classified to be a lead. The typical lead conversion rate at X education is around 30%.

❑ To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.



Business Objective

❑ Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.

❑ A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Strategy

- ❑ **Data Reading and inspecting**
- ❑ **Data Cleaning**
- ❑ **Exploratory Data Analysis**
- ❑ **Data preparation for model building**
- ❑ **Train – Test Split**
- ❑ **Logistic Regression model building on train set**
- ❑ **Model evaluation by different measures and metrics**
- ❑ **Plotting ROC curve**
- ❑ **Finding Optimal Cutoff Point**
- ❑ **Testing model on test set**
- ❑ **Measure accuracy of model by other measures and metrics**
- ❑ **Calculating lead score of each lead**

Data Cleaning & Preparation

❑ **Handling the 'Select' level that is present in many of the categorical variables**

- Replace ‘Select’ with null value

❑ **Null values handling**

- Remove variables with null values higher than 35%
- Replace null values with mode for categorical variables
- Remove rows where null values are less than 2%

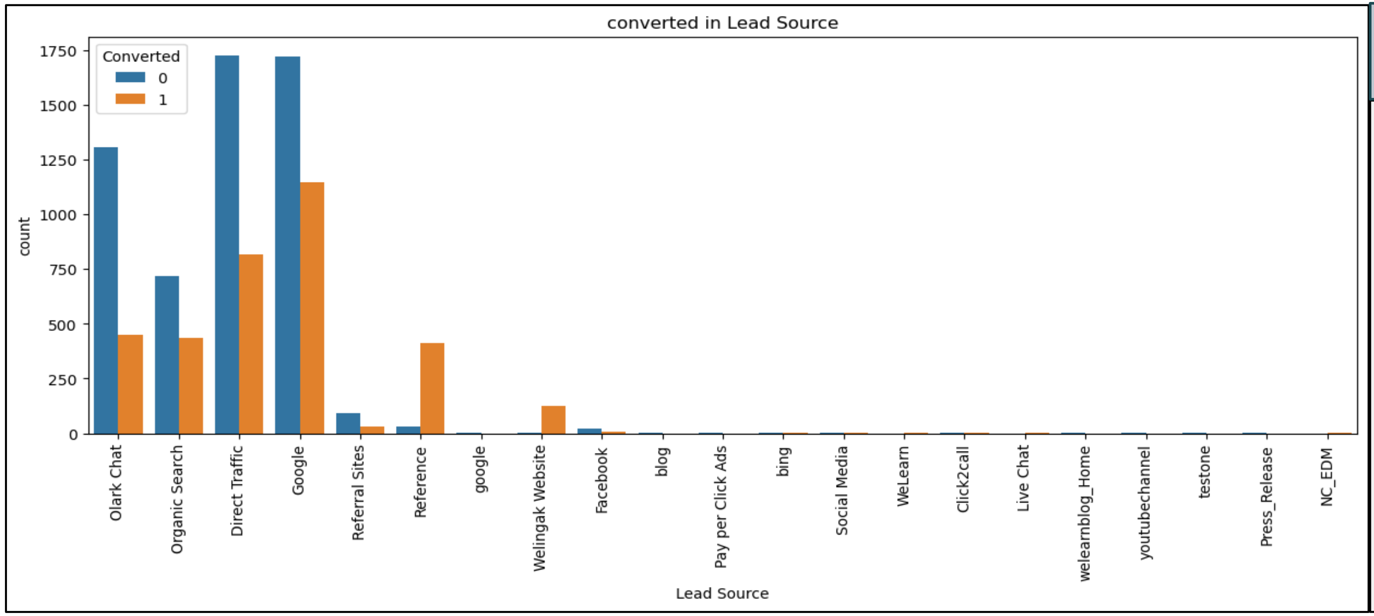
❑ **Outliers handling**

- Capping the outliers to 95% value

❑ **Data Preparation**

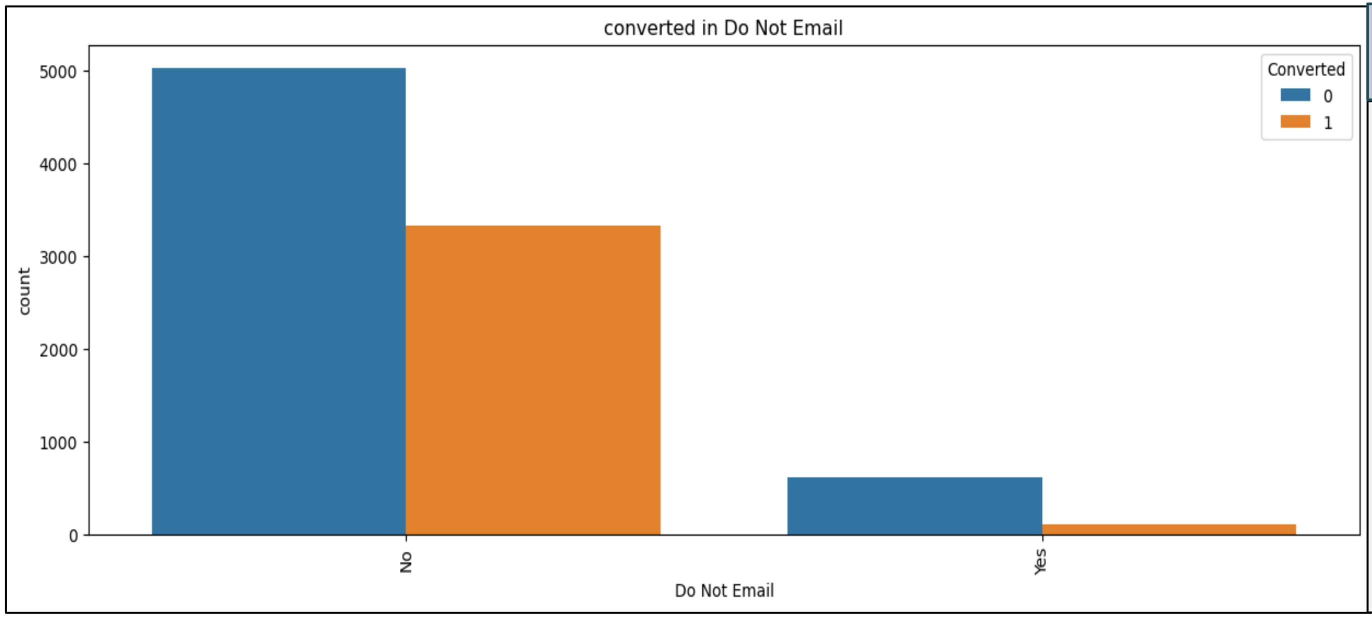
- Converting binary variables (Yes/No) to 0/1
- Creating Dummy variables and Dropping repeated variables

Exploratory Data Analysis



Lead Source VS Converted

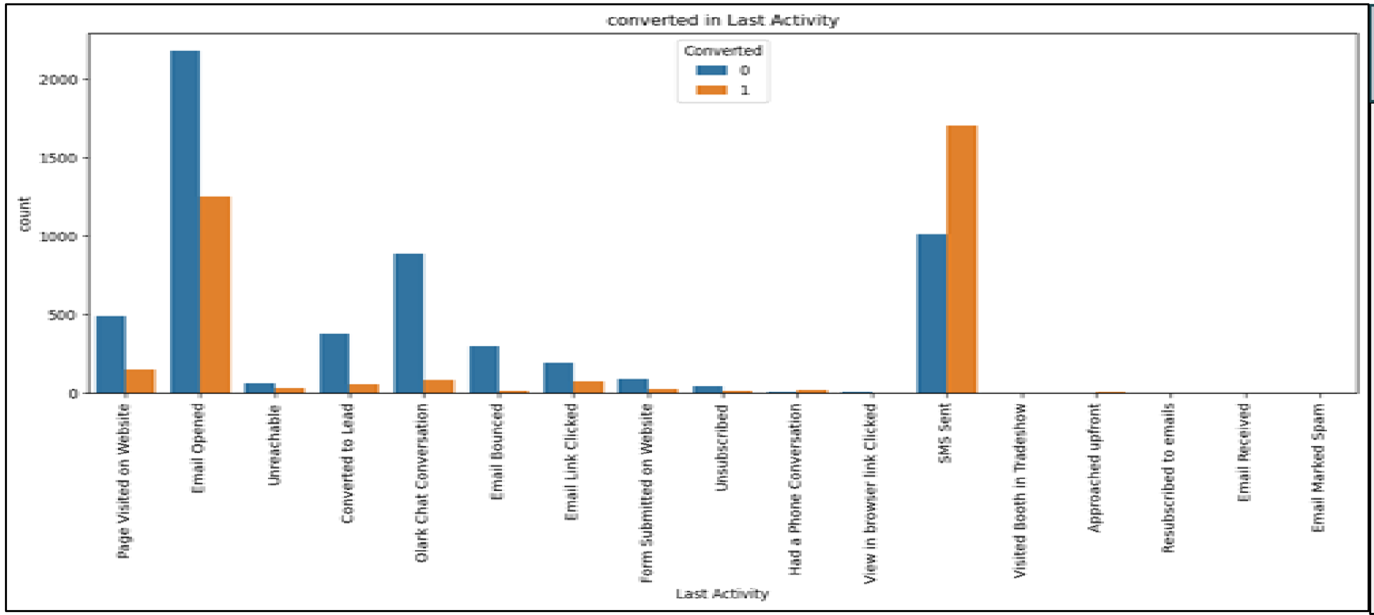
- High Conversion rate:
  - ✓ Reference
  - ✓ Welingak Website
  - ✓ Google searches



Do Not Email VS Converted

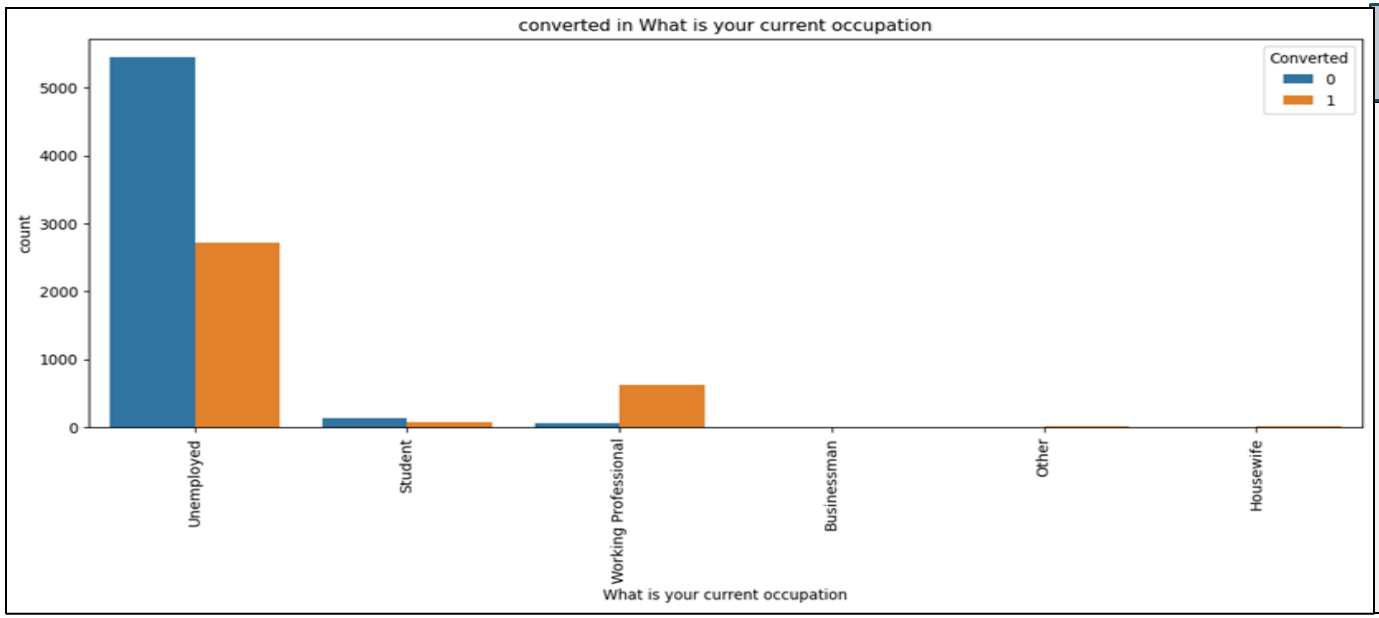
- These leads are not likely to be converted

Exploratory Data Analysis



Last Activity VS Converted

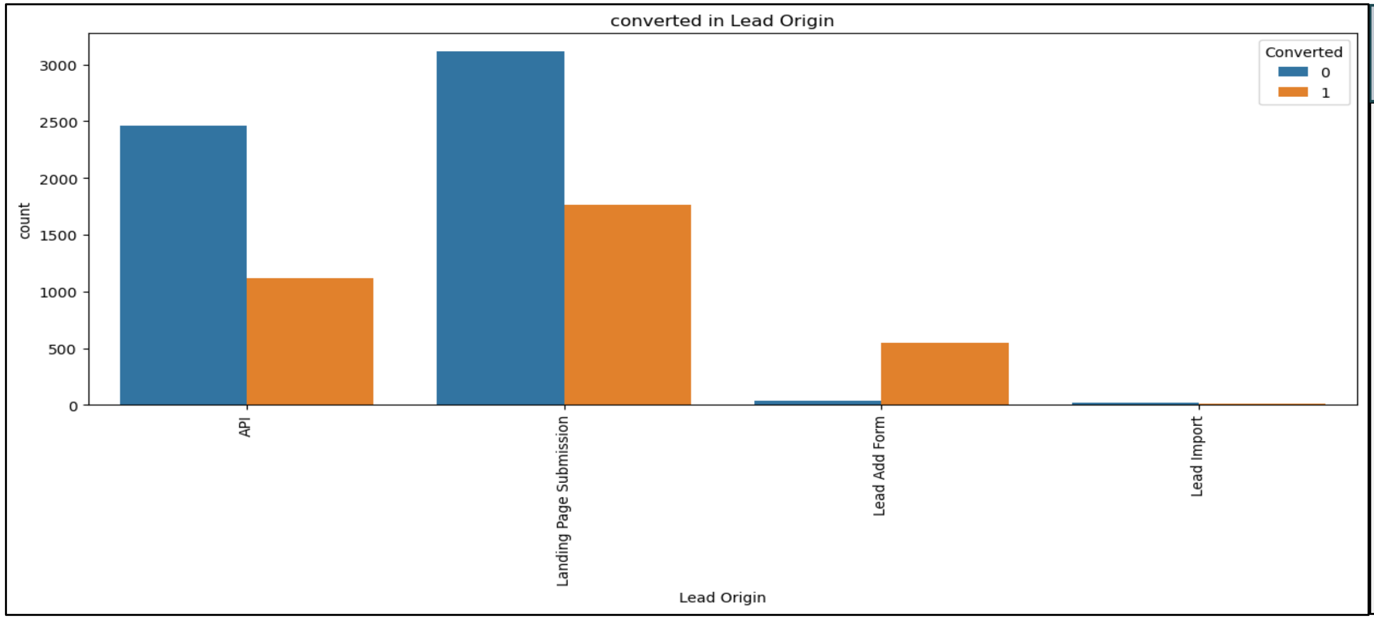
- Sending SMS and Email looks promising method to get higher confirmed leads



What Is Your Current Occupation VS Converted

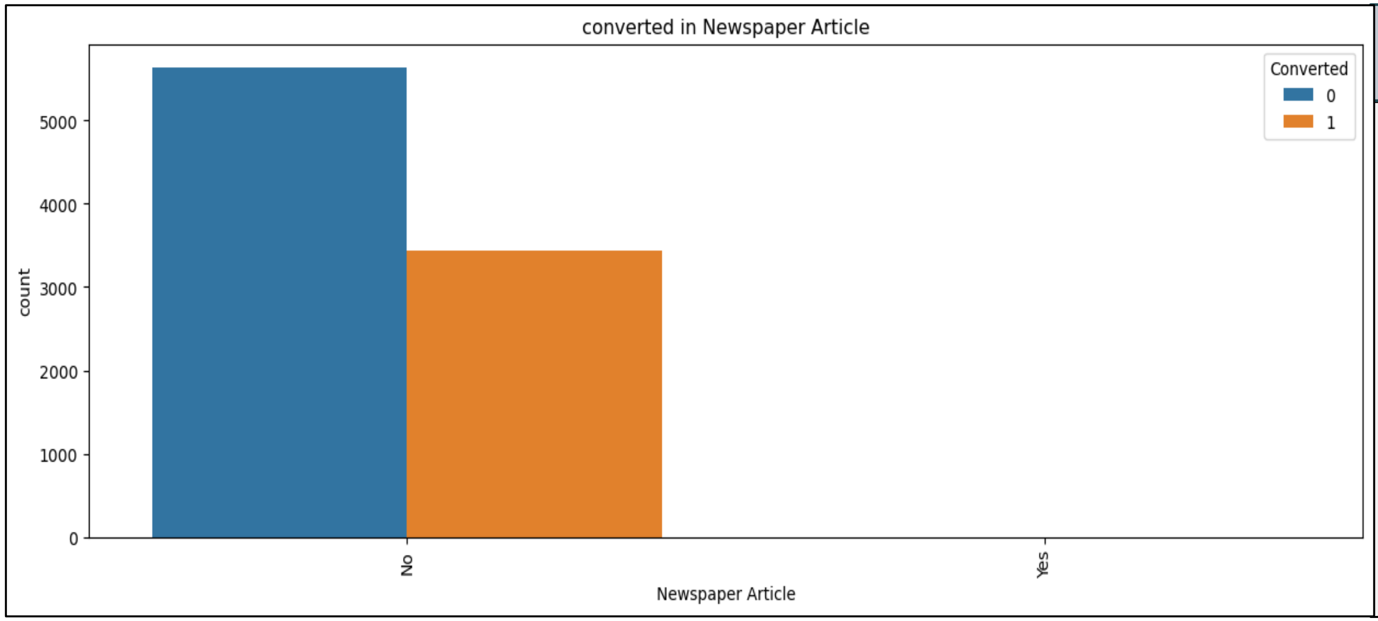
- Sending SMS and Email looks promising method to get higher confirmed leads

Exploratory Data Analysis



Lead Origin VS Converted

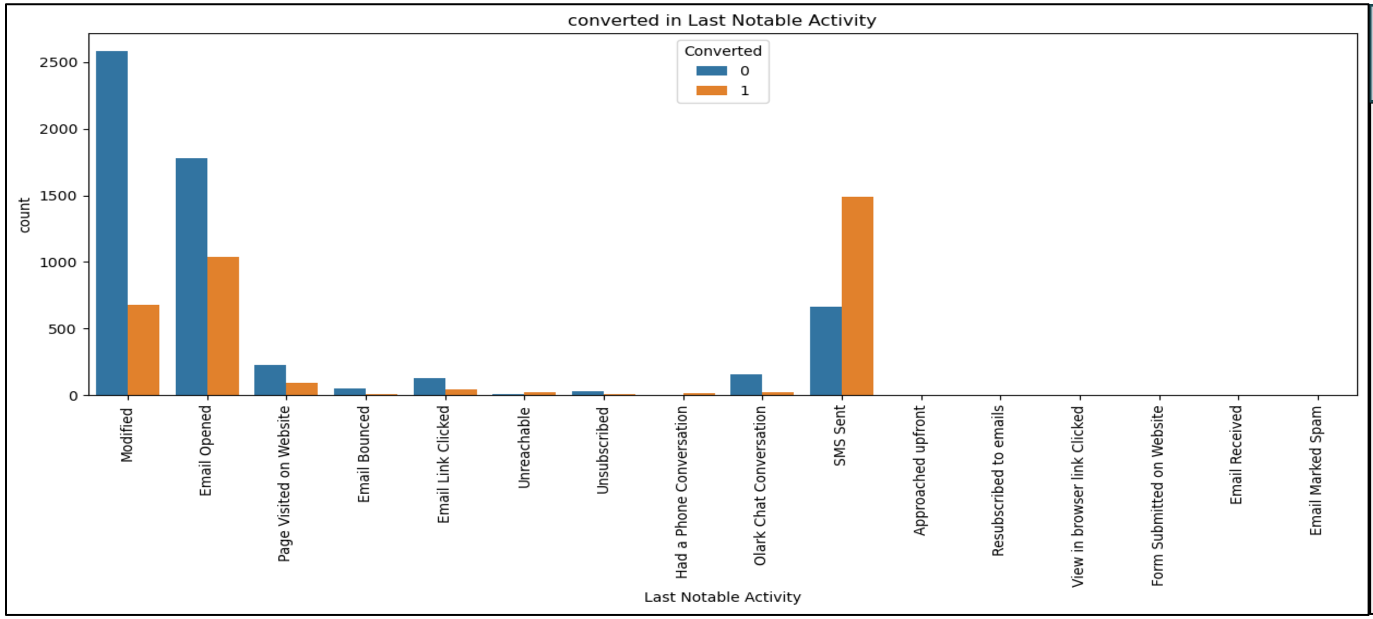
- Lead add form has higher conversion rate



Newspaper Article VS Converted

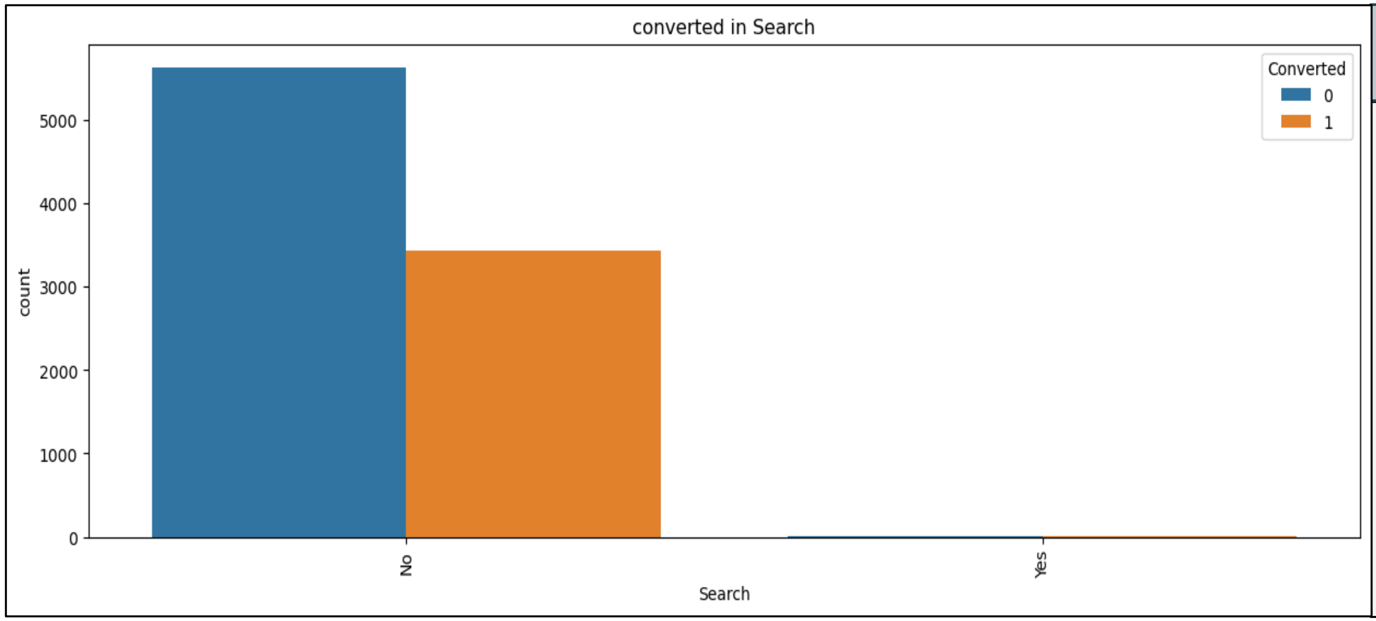
- Highly skewed feature and do not have promising leads

Exploratory Data Analysis



Last Notable Activity VS Converted

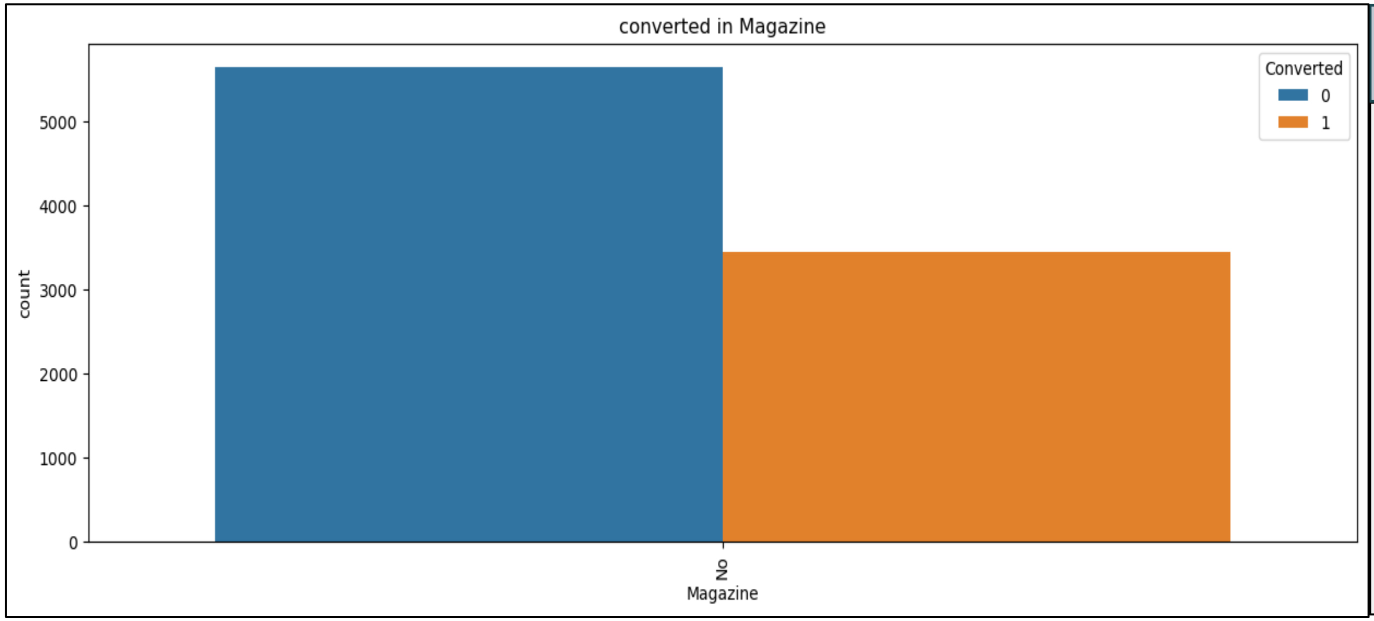
- Most leads are coming from messages and email communication



Search VS Converted

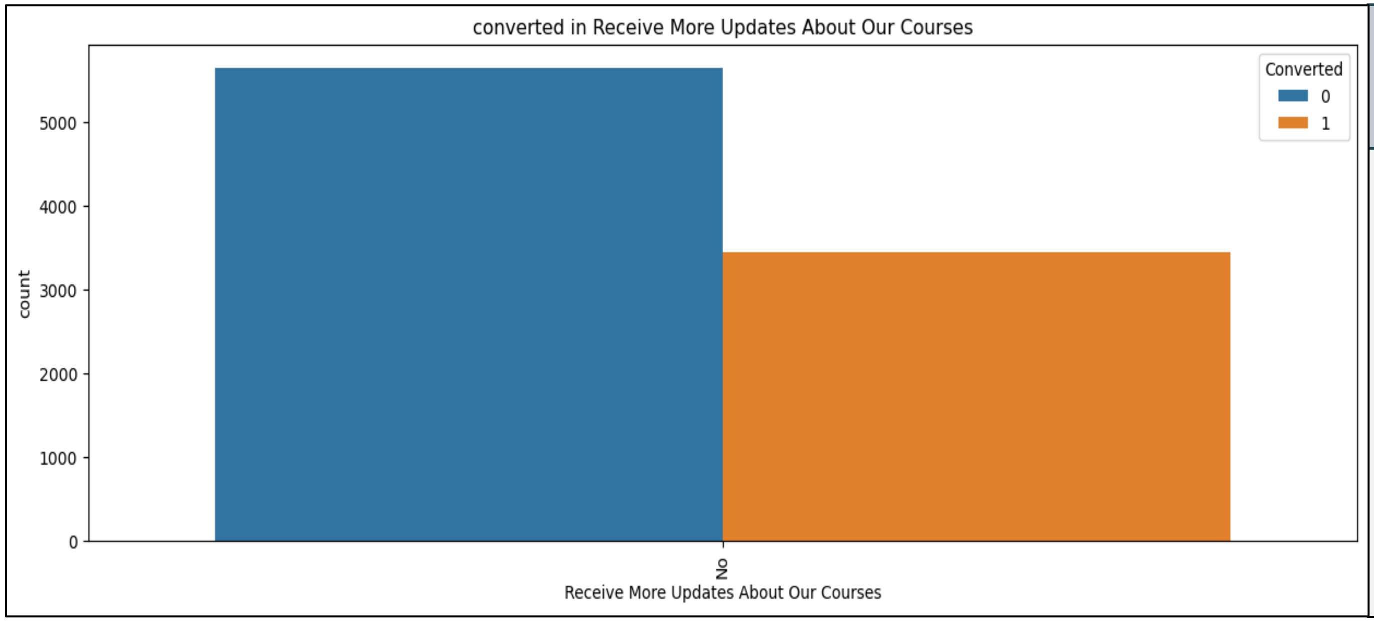
- Search is not good source of leads

Exploratory Data Analysis



Magazine VS Converted

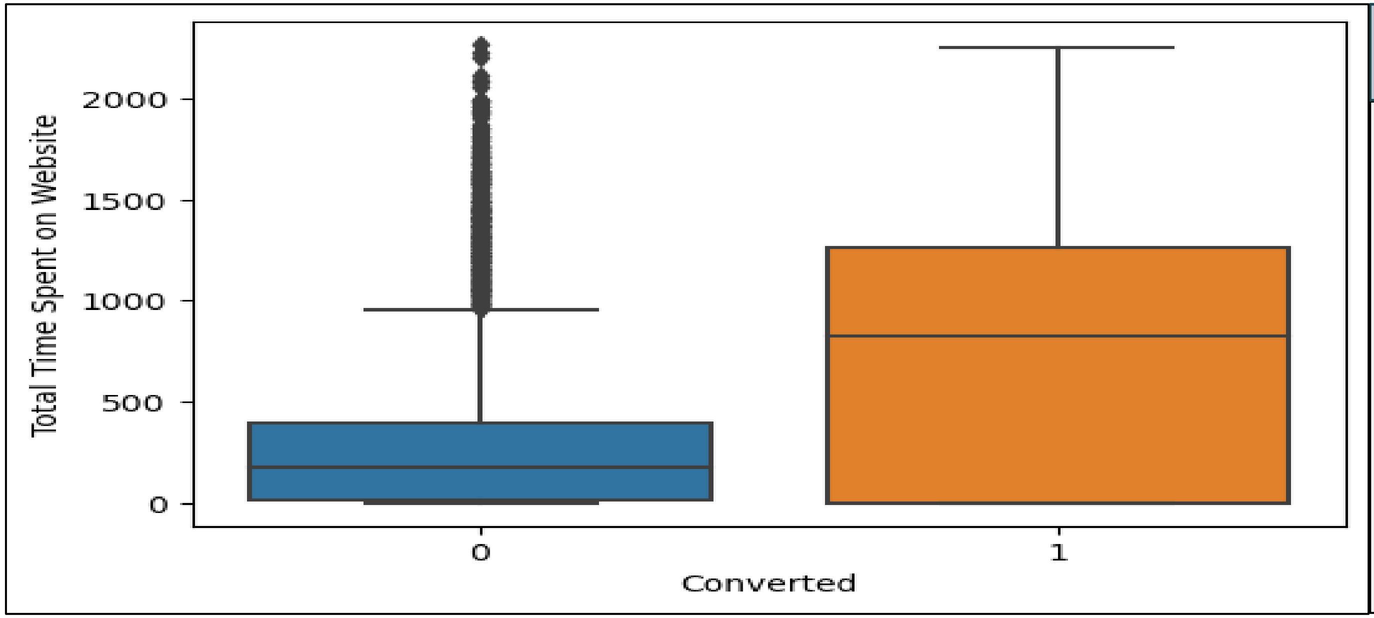
- Only one level is present so not useful for conversion of leads



Receive More Updates About Our Courses Vs Converted

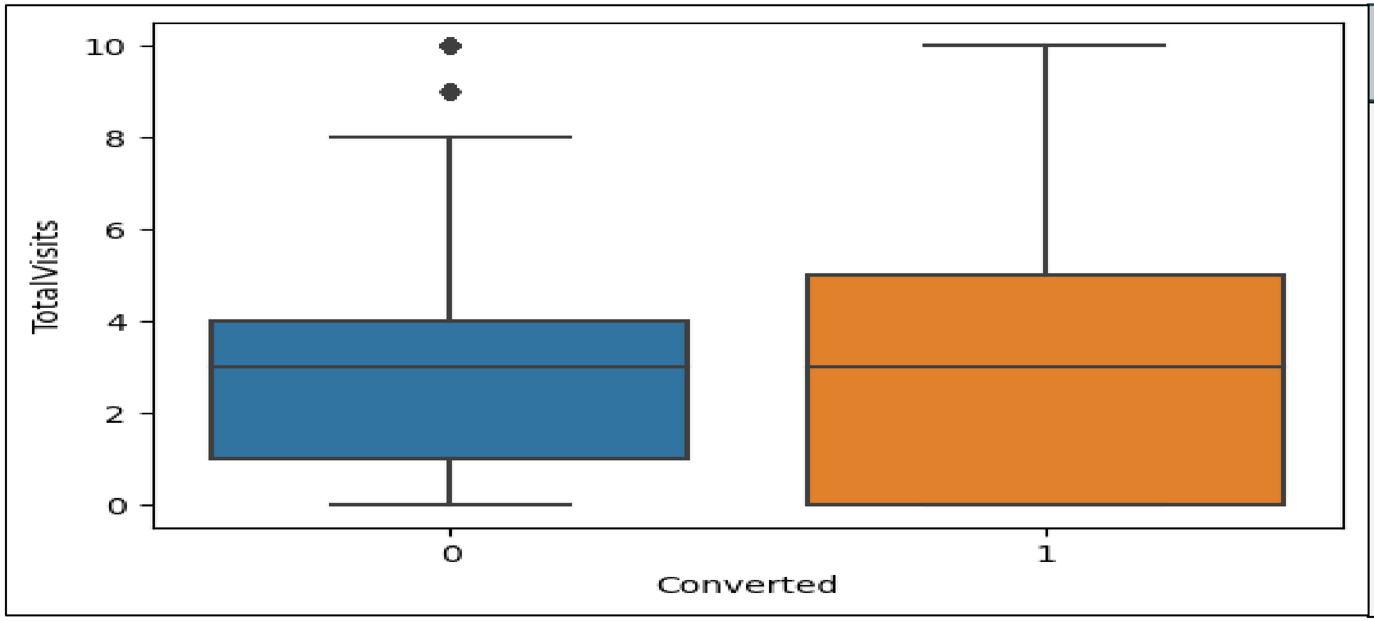
- Highly skewed and does not help to convert lead

Exploratory Data Analysis



Total Time Spent on Website VS Converted

- Leads who are spending more time on website are most likely to get converted



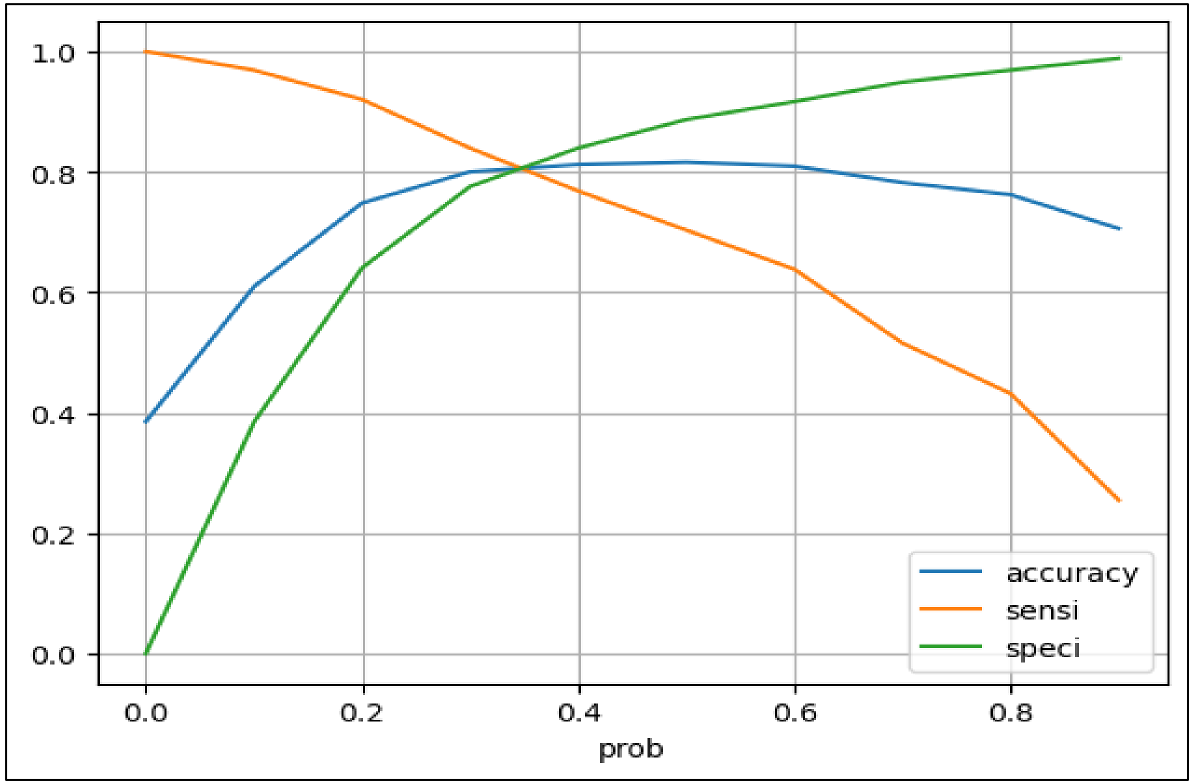
Total Visits VS Converted

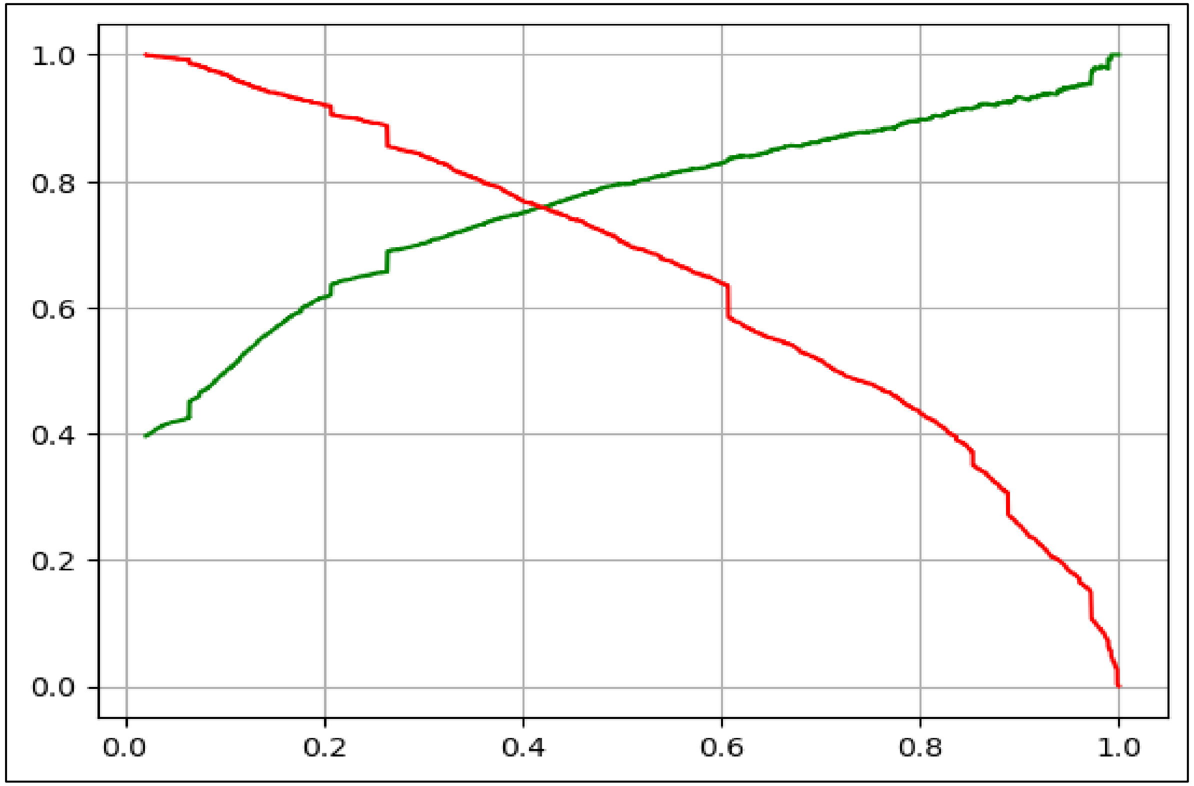
- Number of visits to website is also important factor to get converted

Model Building Process

- Splitting data into train and test sets
- Scale variables in train set
- Build model using variables selected by RFE
- Eliminate variable based on high P-value
- Build next model
- Check VIF value for all existing variables
- Predict using train dataset
- Evaluate accuracy and other metrics
- Plotting ROC curve
- Finding optimal cutoff point
- Predict using test dataset
- Accuracy, Precision, Recall analysis on test predictions

Model Evaluation – Train Dataset





Accuracy : **81.7 %**

Sensitivity : **70.3 %**

Specificity : **88.7 %**

Cutoff point : **0.35**

Confusion Metrix

3463	442
726	1720

Precision : **79.6 %**

Recall : **70.3 %**

Model Evaluation – Test Dataset

Accuracy : **87.4 %**

Sensitivity : **79.2 %**

Specificity : **81.0 %**

Confusion Metrix

1405	329
206	783

Precision : **70.4 %**

Recall : **79.2 %**

Conclusion

- Final model
  - **Accuracy is 81.7 %**
  - **Precision 79.6 %** & **Recall 70.3 %.**
- Optimal **Cutoff is 0.35** after checking accuracy, sensitivity and specificity metrics.
- Model works good on test dataset also **with Accuracy 80.4 %, Sensitivity 79.2 %, Specificity 81.0 %**
- Top variables in your model which contribute most towards the probability of a lead getting converted
  - **Total Time Spent on Website**
  - **Lead Origin - Lead Add Form**
  - **What is your current occupation - Working Professional**

Recommendations

- **The X company should focus on following kind of Leads,**
  - Spending more time on website
  - Working professionals
  - Leads coming through - Reference or welingak website
  - Leads whose last activity is – SMS and Email communication
  - Leads whose origin is either 'Lead Add Form' or 'Lead Import
  - The X company can also focus on leads whose number of visits to website is higher.

- **The X company should not focus on these Leads,**
  - Leads who prefer 'Do not Email' and 'DO not Call'
  - Unemployed and Student leads
  - Leads whose last activity was "Olark Chat Conversation"