Lead Scoring Case Study

SUBMITTED BY:

KAJOL ROHRA

TEJASHREE MC

Problem Statement

- X Education sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses and 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. Moreover, the company also gets leads through past referrals.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

Business Goal

- X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Analysis Approach

- Reading and Understanding Data
- Data Cleaning
- Exploratory Data Analysis
- Creating Dummy Variables
- Test Train Split
- Feature Scaling
- Feature selection using RFE
- Finding the Optimal Cutoff Point
- Computing the Precision and Recall metrics
- Making Predictions on Test Set

Data Understanding and Cleaning

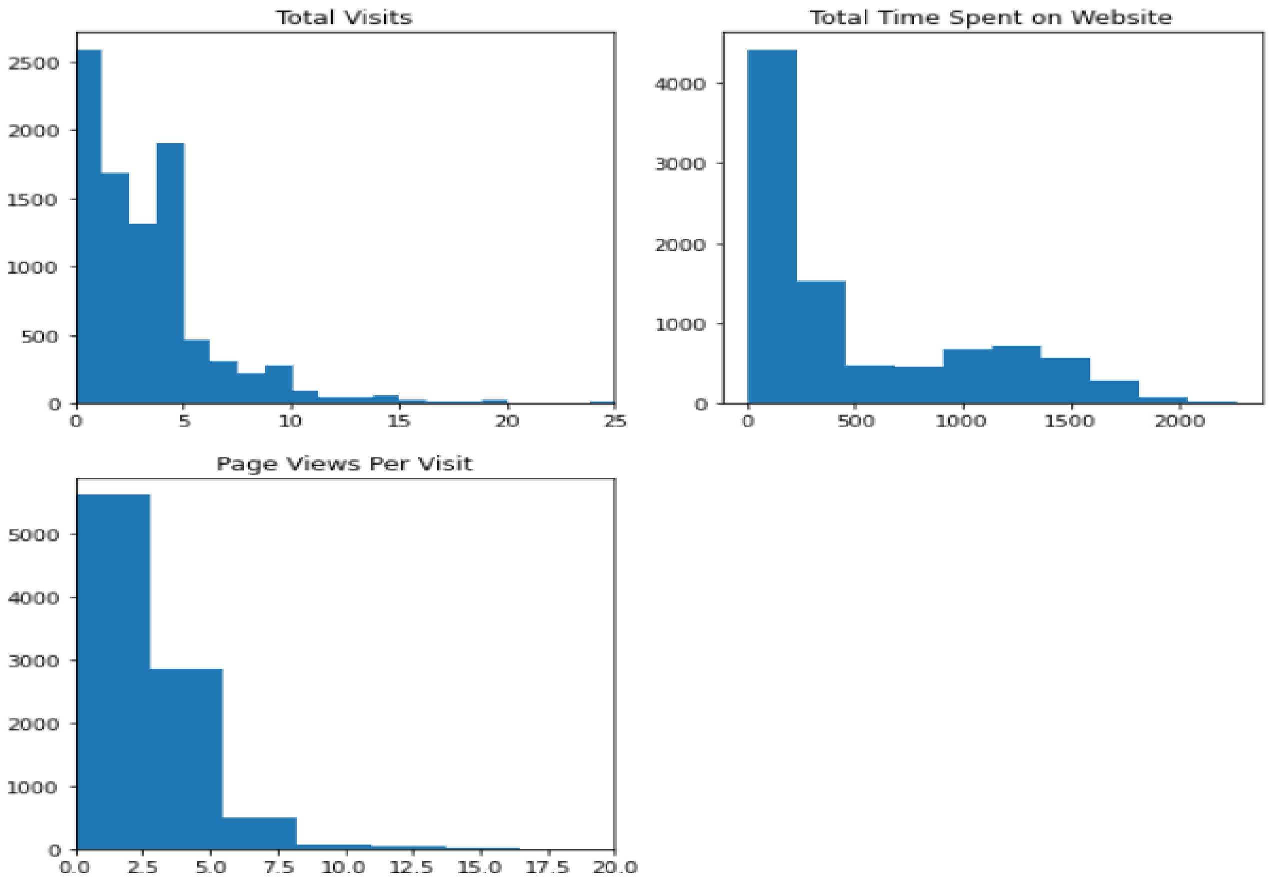
- The dataset Leads.csv has around 9240 entries with 37 attributes.
- The given data has entries as “Select” which are blank values and are not selected by the user while filling the data. These values are treated as null and are handled
- Other missing values are imputed accordingly.
- Dropping columns/rows with high missing values.
- Dropping columns with high data imbalance for a single category
- Minority categories are clubbed into one category
- Outlier analysis
- EDA

Data Preparation

- Binary variables of Yes or No are mapped to 1 or 0
- Dummies are created for categorical variables.
- Data is split into Train and Test in the ratio 70:30
- Scaling is performed on this data

Exploratory Data Analysis

- Univariate Analysis



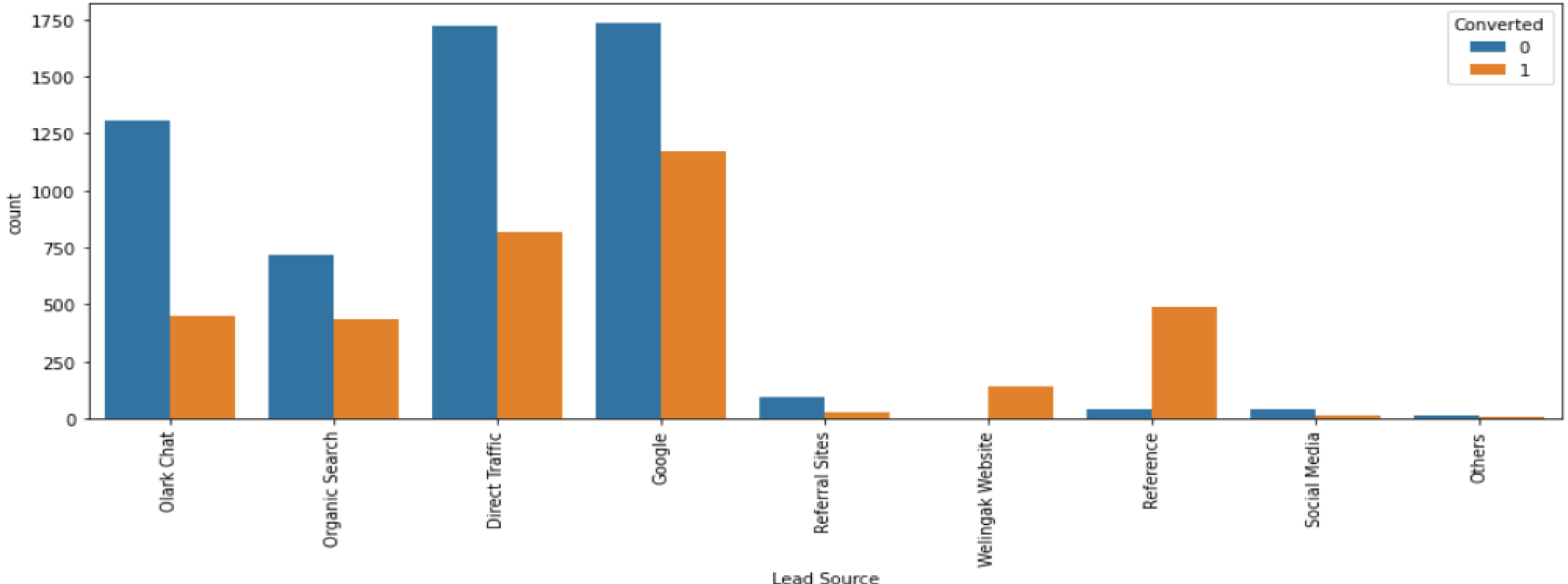
Total Visits

Total Time Spent on Website

Page Views Per Visit

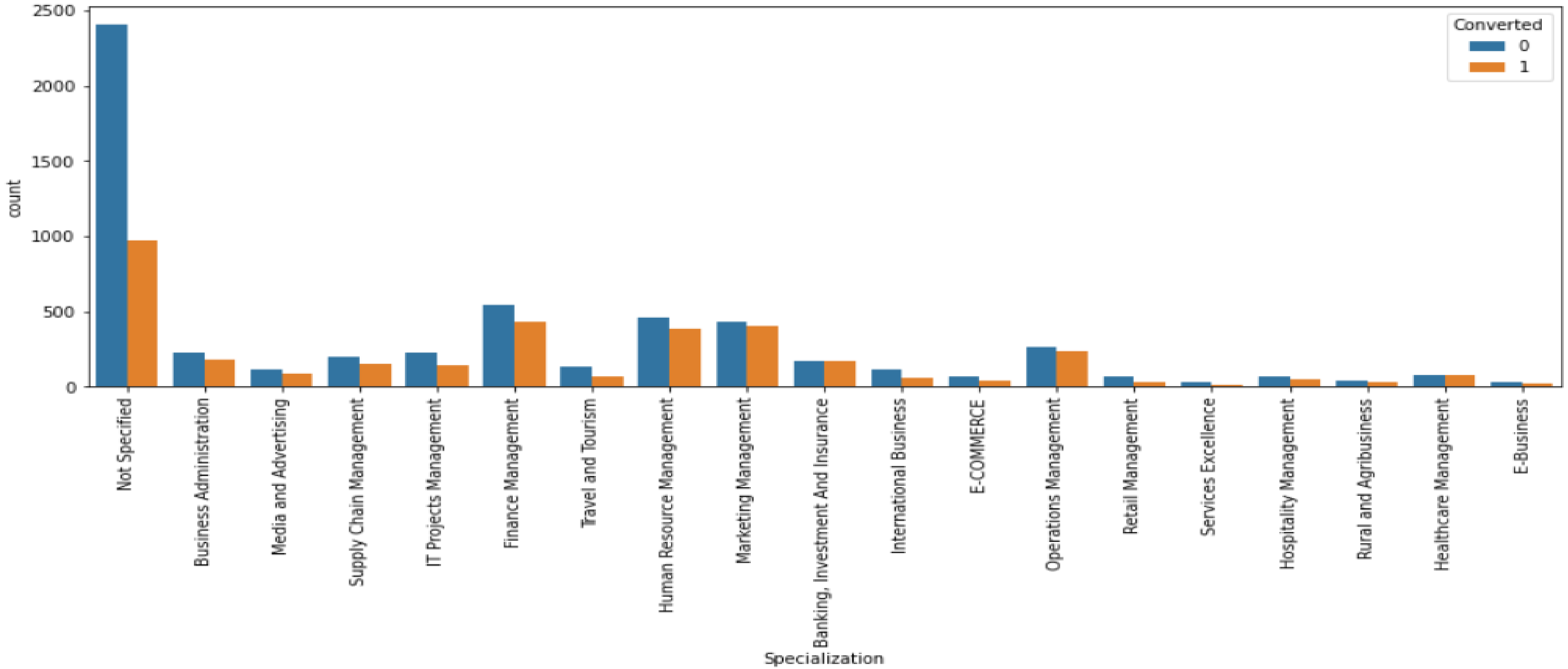
Exploratory Data Analysis

Categorical Variable Analysis: Lead Source



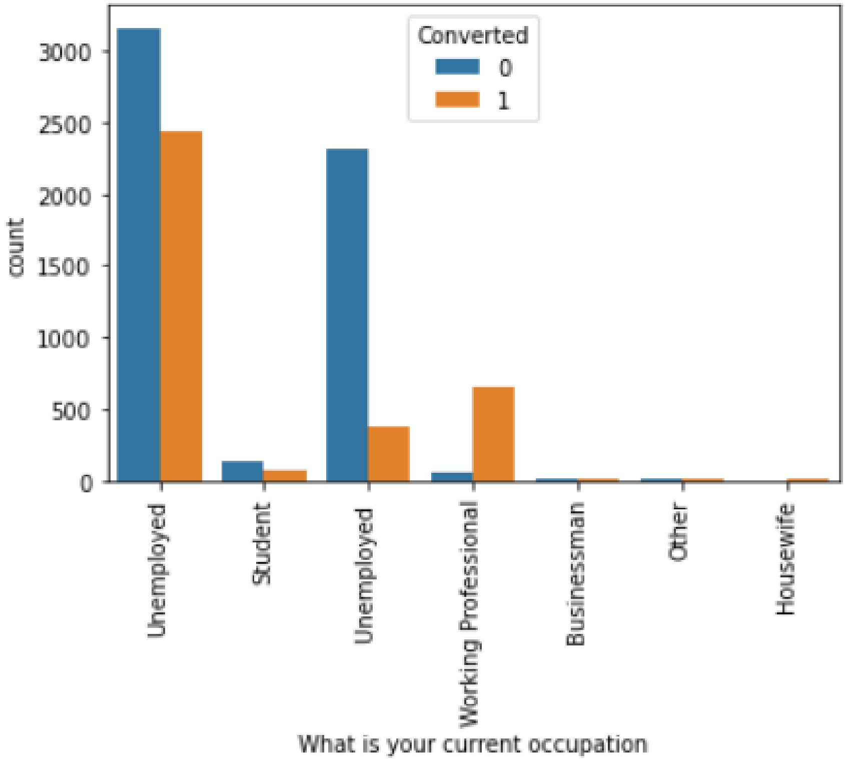
Exploratory Data Analysis

Categorical Variable Analysis: Specialization



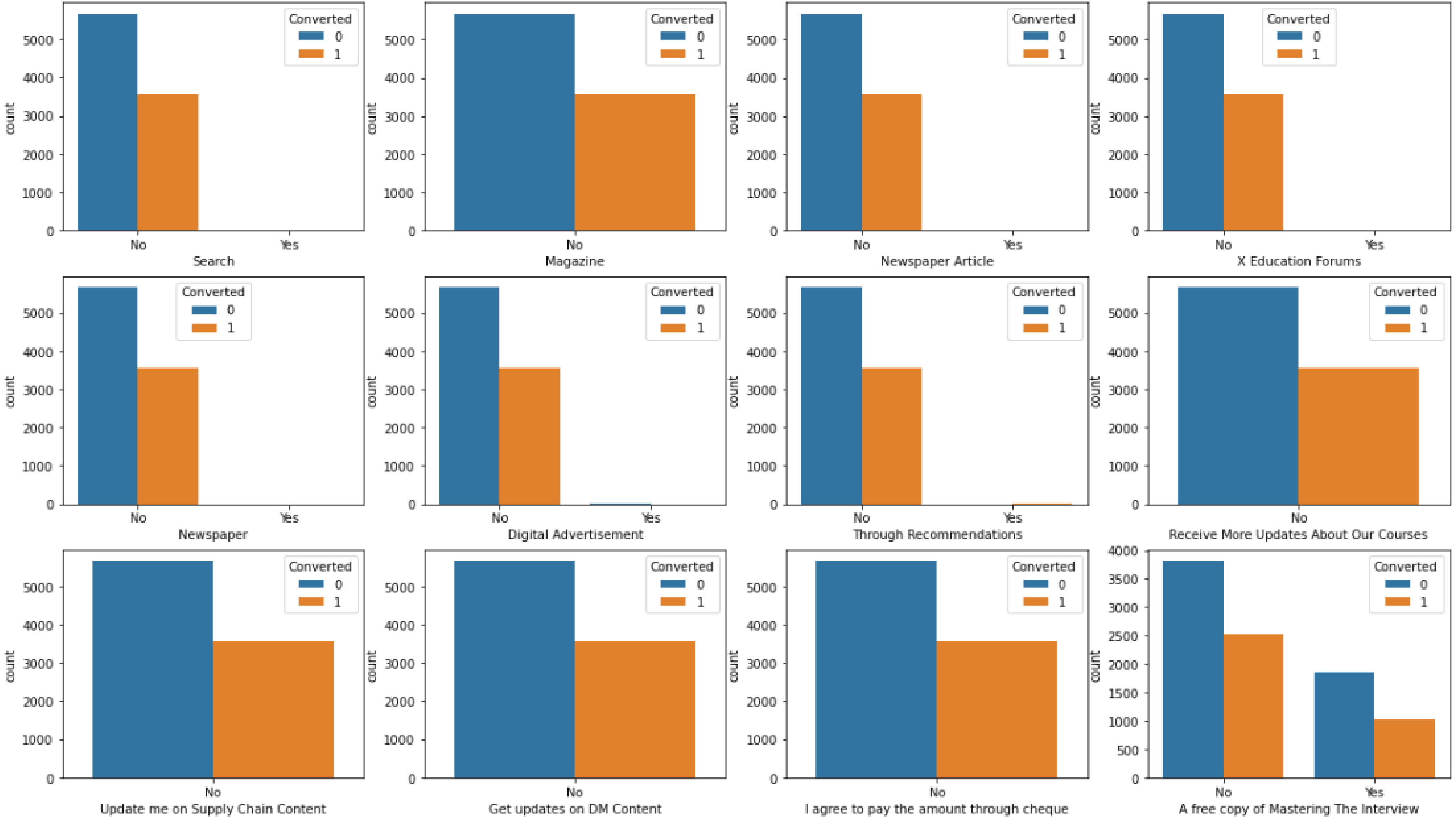
Exploratory Data Analysis

Categorical Variable Analysis: What is your current occupation



Exploratory Data Analysis

Imbalance Analysis



Variables Impacting the Conversion Rate

- Do Not Email
- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Origin_Lead Import
- Lead Source_Direct Traffic
- Lead Source_Google
- Lead Source_Organic Search
- Lead Source_Reference
- Lead Source_Referral Sites
- What is your current occupation_Other
- What is your current occupation_Student
- What is your current occupation_Unemployed
- What is your current occupation_Working Professional

Model Evaluation

Accuracy, Sensitivity and Specificity on Train Dataset:

- Accuracy : 76%
- Sensitivity: 83%
- Specificity: 72%

Model Evaluation

Precision and Recall on Train Dataset:

- Precision: 65%
- Recall: 83%

Model Evaluation

Accuracy, Sensitivity and Specificity on Test Dataset:

- Accuracy : 77%
- Sensitivity: 81%
- Specificity: 74%

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 77%, 81% and 74% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%.

Recommendation

Important features responsible for good conversion rate or the ones' which contributes more towards the probability of a lead getting converted are :

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- What is your current occupation_Working Professional