Lead Scoring Test Study

By – Kamlesh Atara

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. 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%.
- Now, although X Education gets a lot of leads, its lead conversion rate is very poor. 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. A typical lead conversion process can be represented using the following funnel:
- The company requires to build a model wherein we need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a 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%.

Expectation from the model

- Want to get list of most important variable company should consider while making calls
- Want to improve the lead conversion rate

Solution Steps

- **Preprocessing of I/p Data**

1. Inspect I/P data
2. Handle Null values, Unwanted columns, Outliers

- **EDA/Visualization**

1. Categorical univarient Analysis
2. Numerical univarient Analysis
3. Bi-Varient Analysis - Numerical - Numerical Analysis
4. Bi-Varient Analysis - Categorial to Converted Analysis

- **Model Building**

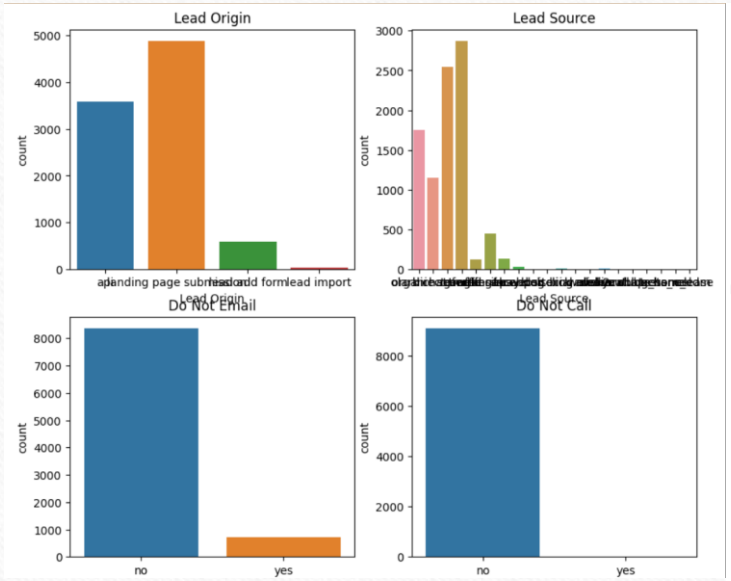
1. Feature Scaling and getting dummy variables
2. Using RFE for Feature Selection
3. Model Building using Logistic Regression
4. Predicting and validating the model using different parameters
5. Using model to predict test data
6. Conclusion

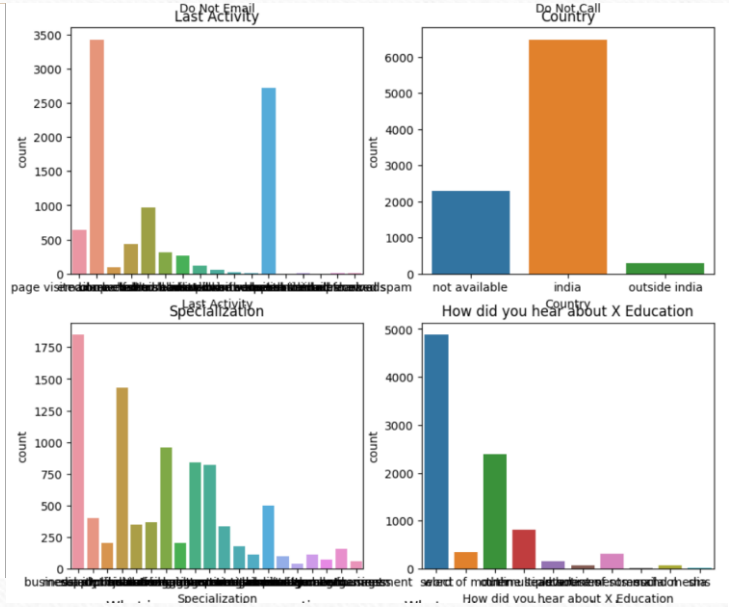
Preprocessing of I/p Data

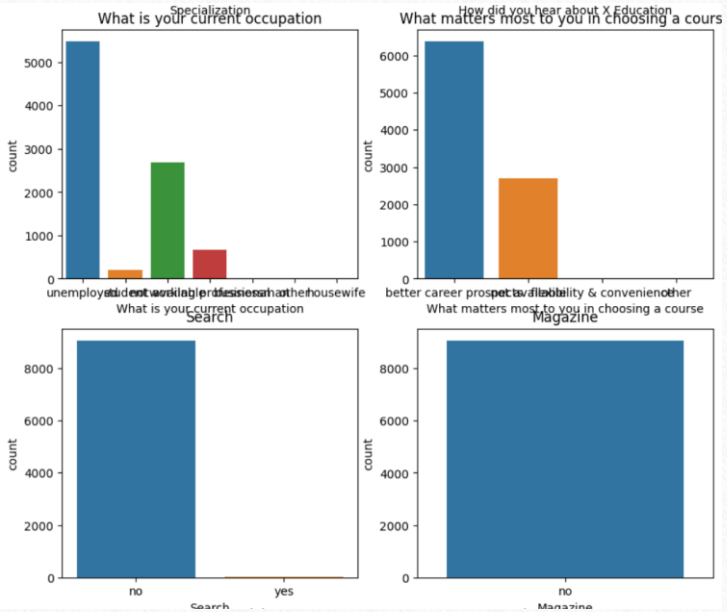
- Handle Null Values by using different methods like creating new category – Not available, removing columns or rows
- Removing unwanted columns based on different criteria like has only single value, unique for each rows
- Handling outliers by dropping outlier rows

EDA/Visualization

- Categorical Univariant Analysis

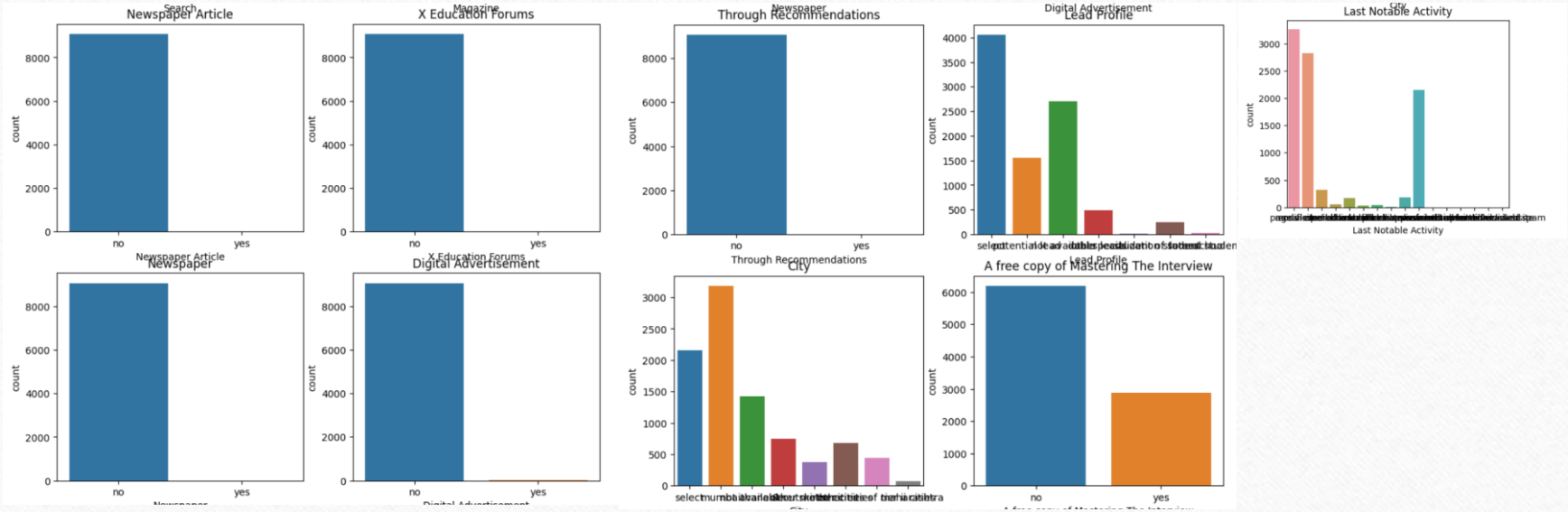






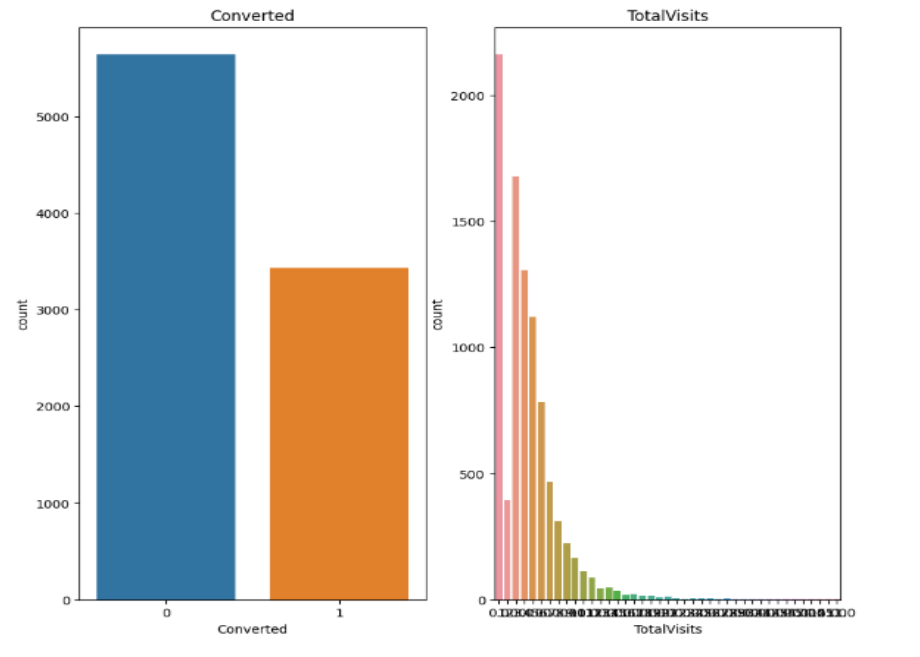
EDA/Visualization – Continue…

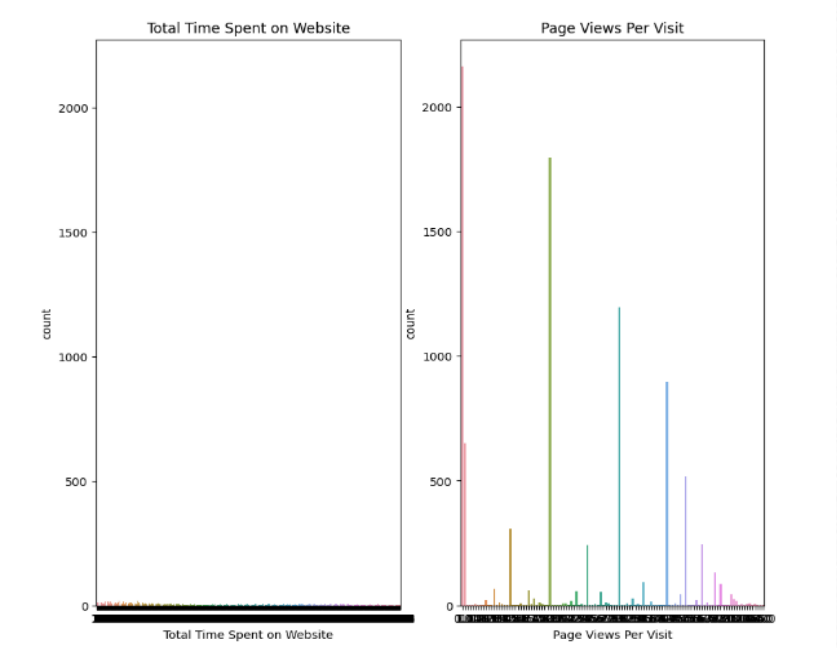
- Categorical Univariant Analysis



EDA/Visualization – Continue…

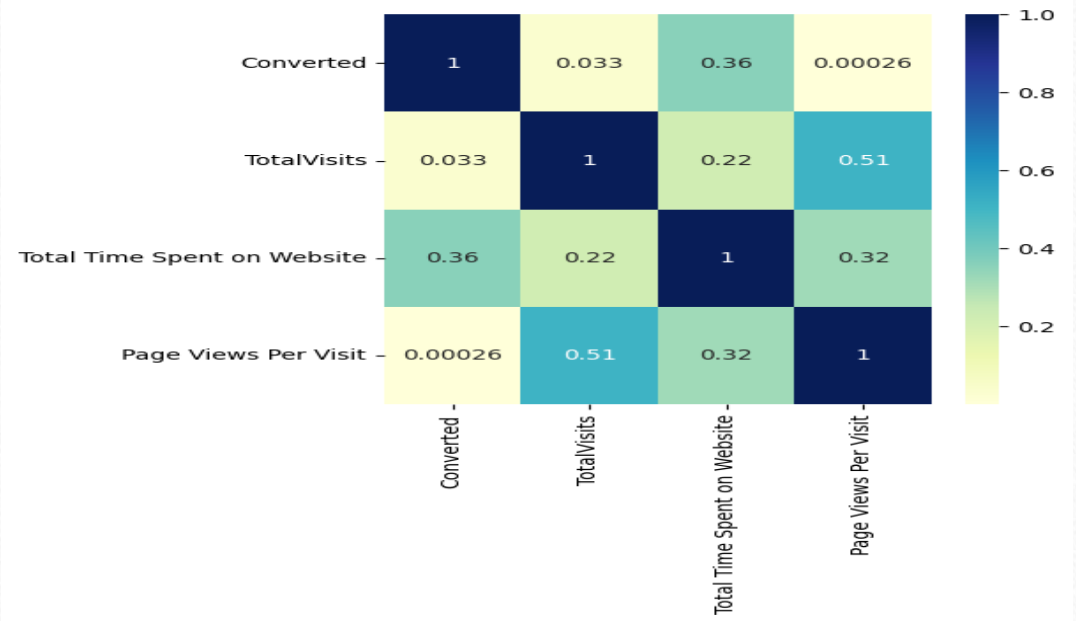
- Numerical Univariant Analysis





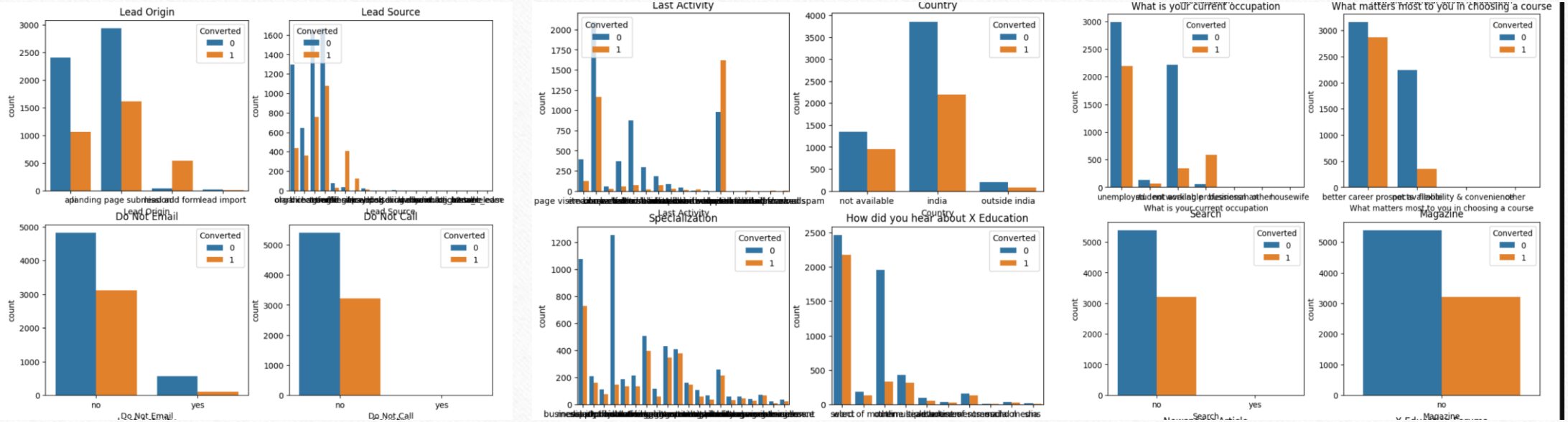
EDA/Visualization – Continue…

- Bi-Varient Analysis - Numerical - Numerical Analysis



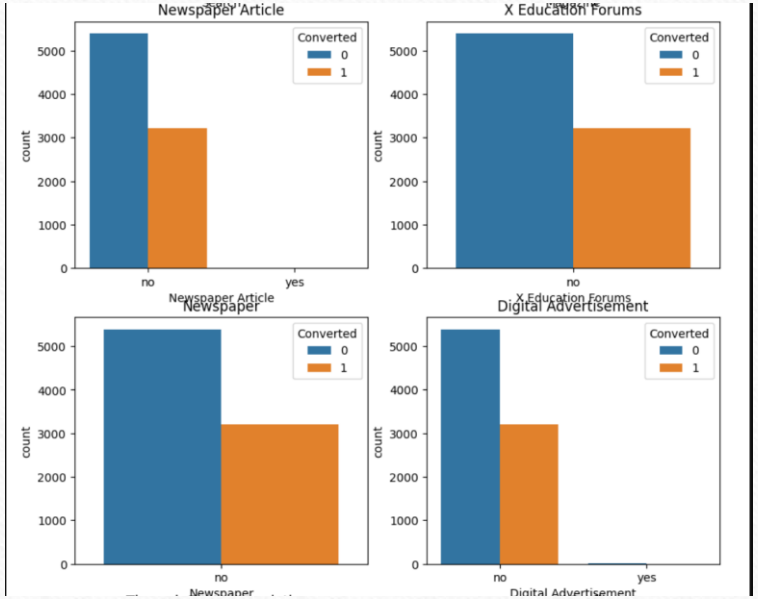
EDA/Visualization – Continue…

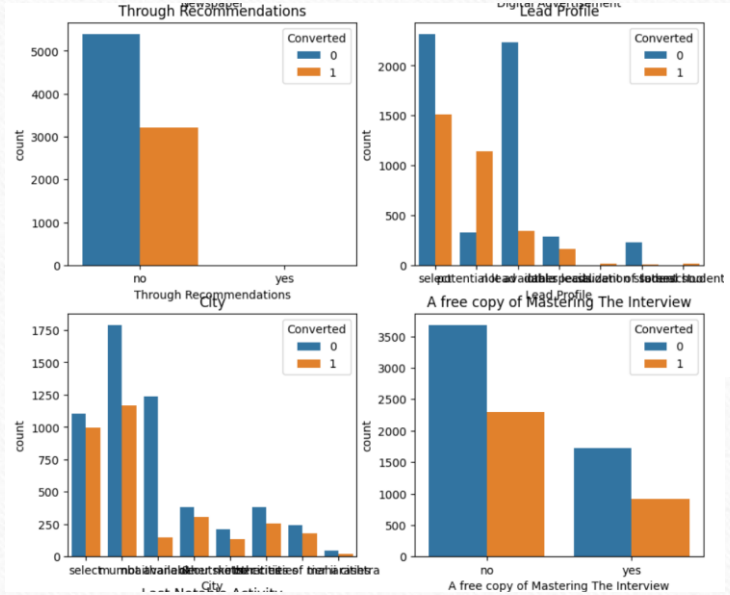
- **Bi-Varient Analysis - Categorial to Converted Analysis**

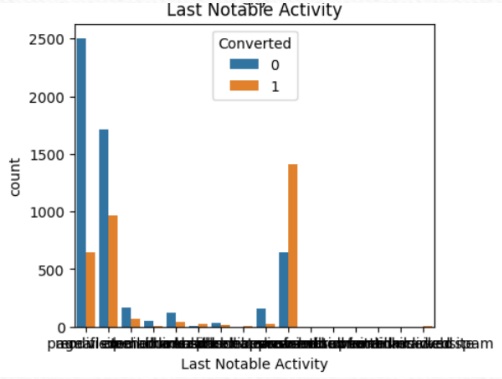


EDA/Visualization – Continue…

- **Bi-Varient Analysis - Categorial to Converted Analysis**







Model Building

- Convert Binary Categorical Variable into 1 and 0
- Using get_dummies split categorical variables
- Drop unwanted columns
- Train Model

1. Using train_test_split, split the dataset into train and test data

Train row, col = 6024,110

Test row, col = 2583, 110

Model Building Cont…

- Use Feature scaling for numeric variables
- Considering number of fields, correlation is difficult so use RFE to identify top 20 variables
- Use logistic regression to build model on train data set
- Evaluate model and use VIF and P-value to remove unwanted variables
- Calculate accuracy, sensitivity, specificity on final model
- Plot ROC curve and find optimal cutoff plot
- Use precision and recall tradeoffs to get good cutoff value
- Use model to predict value on test data set

Model Building Cont…

- **Train Data set**
- Accruacy is 83.4%
- Sensitivity is 72.8%
- Specificity is 89.7%

- **Test Data Set**
- Accuracy of test model - 81.8%
- Sensitivitiy of test model - 76.5%
- Specificity of test model – 85.7%

Conclusion

- Top 3 variables are

1. Lead Origin - lead add form
2. What is your current occupation - working professional
3. Lead Profile - potential lead