Lead Scoring Case Study

Members:

1. Shakti Singh Chauhan
2. Sheeba
3. Mothiki Siva Surya

Problem statement

- X Education sells online courses to professionals in the field, but while receiving a lot of leads, it has a relatively low lead conversion rate.
- For instance, only approximately 30 of 100 leads they could gather in a day might actually be converted.
- The goal of the business is to find the most promising leads, commonly referred to as "Hot Leads," in order to increase the efficiency of this process.
- The lead conversion rate should increase if they are successful in locating this group of leads because the sales staff will be spending more time speaking with potential leads rather than calling everyone.

Business objective

- X Education aims to construct a model that can identify the hot leads in order to find out which leads are the most promising.

Methodology

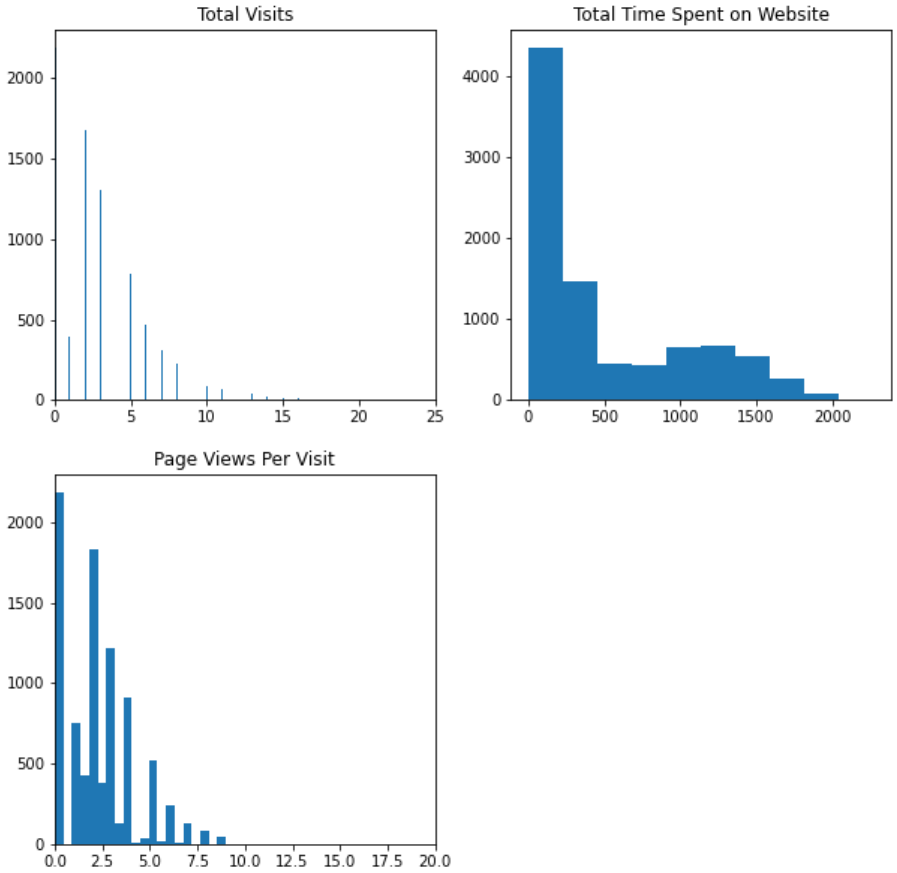
- Data cleaning
    1. Check and handle duplicate data, missing values, unique values and outliers.
    2. Drop columns, if it contains large number of missing values and not useful for the analysis.
    3. Imputation of the values, if necessary.
- EDA
    1. Univariate data analysis: value count, distribution of variable etc
    2. Bivariate data analysis: correlation coefficients
- Feature Scaling & Dummy Variables and encoding of the data
- Classification technique: logistic regression used for the model making and prediction
- Train-Test split
- Model Building
- Model evaluation
- Prediction
- Conclusions and summary

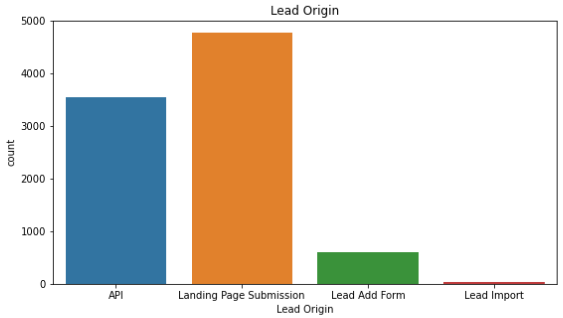
Data Cleaning

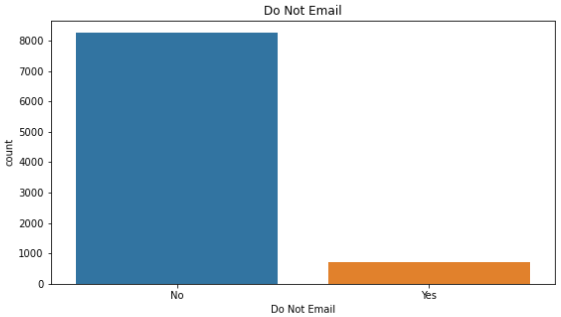
- Total Number of Rows was 37 and Total Number of Columns was 9240
- Dropped the columns having more than 45% as missing value
- Unique value features have been dropped
- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped the features as well.
- Clubbing has been done for features so that it was easy to understand

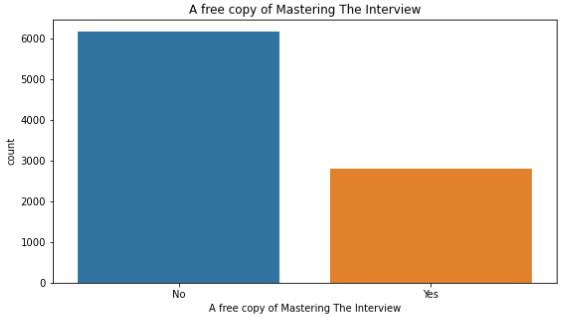
Exploratory Data Analysis

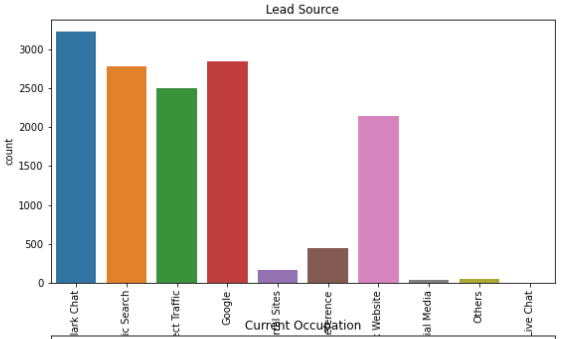
Numerical variables analysis

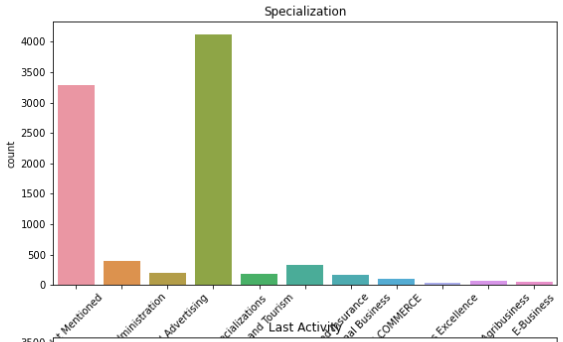


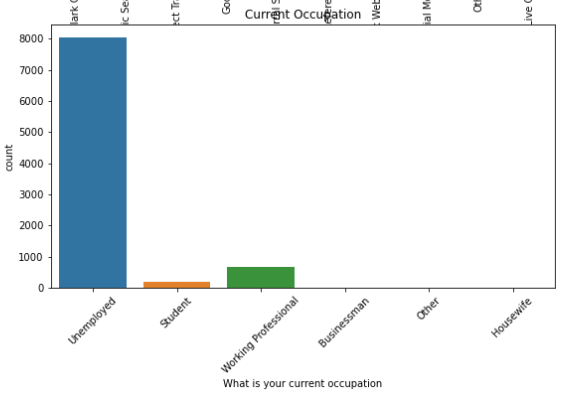


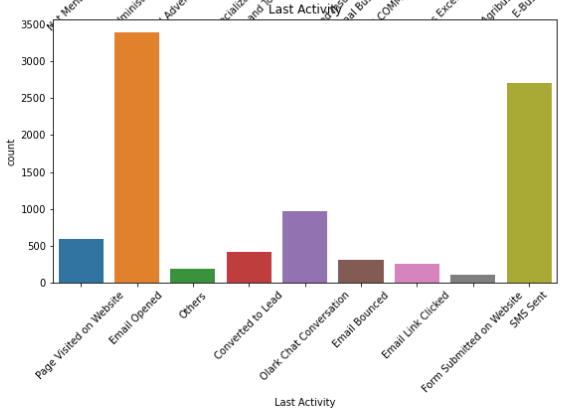












Categorical variable analysis

Data conversion

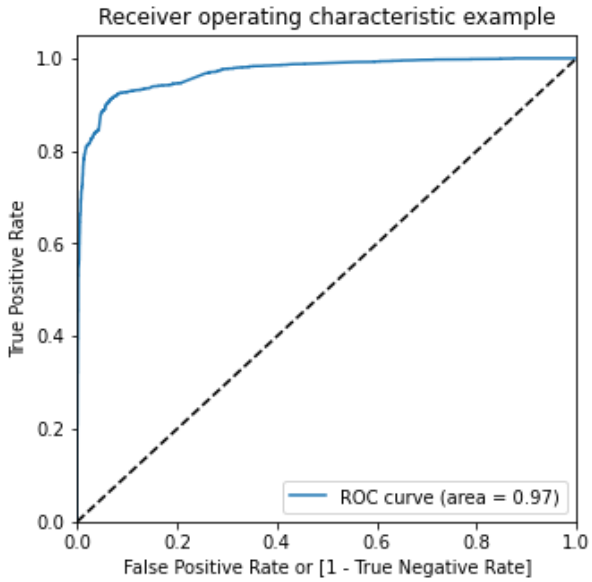
- Dummy Variables are created for object type variables
- Total Rows for Analysis: 8953
- Total Columns for Analysis: 57

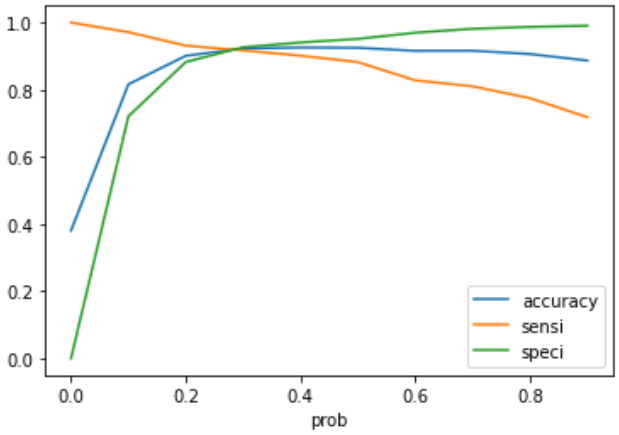
Model building

- 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
- Building Model by removing the variable whose p-value is greater than 0.05 and vif value is greater than 5
- Predictions on test data set
- Overall accuracy 92%

ROC Curve

- Finding Optimal Cut off Point where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.3





Final Observation

Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

Test Data:

- Accuracy : 92.70%
- Sensitivity : 91.68%
- Specificity : 93.32%

Conclusion

The factors that affected potential purchasers the most were discovered to be (in descending order):

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

Thank you