Lead Scoring CASE STUDY

E.Tejaswini

Problem Statement

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us 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 has given a ballpark of the target lead conversion rate to be around 80%

Approach

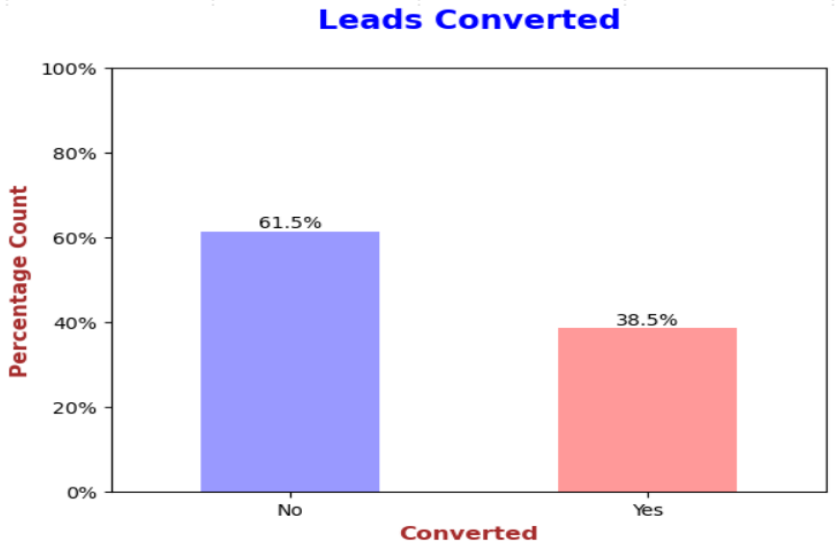
- Data Cleaning: Loading Data Set, understanding & cleaning data
- Data Analysis: Check imbalance, Univariate & Bivariate analysis Data Preparation Dummy variables, test-train split, feature scaling
- Model Building: RFE for top 15 feature, Manual Feature Reduction & finalizing model
- Model Evaluation: Confusion matrix, Cut-off Selection, assigning Lead Score
- Predictions on Test Data: Compare train vs test metrics, Assign Lead Score and get top features
- Recommendation: Suggest top 3 features to focus for higher conversion & areas for improvement

Data cleaning

- Columns with 40% nulls are dropped
- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Dropped unnecessary columns like tags,city
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Additional categories were created for some variables.
- Columns with no use for modeling (Prospect ID, Lead Number) or only one category of response were dropped.
- Numerical data was imputed with mode after checking distribution

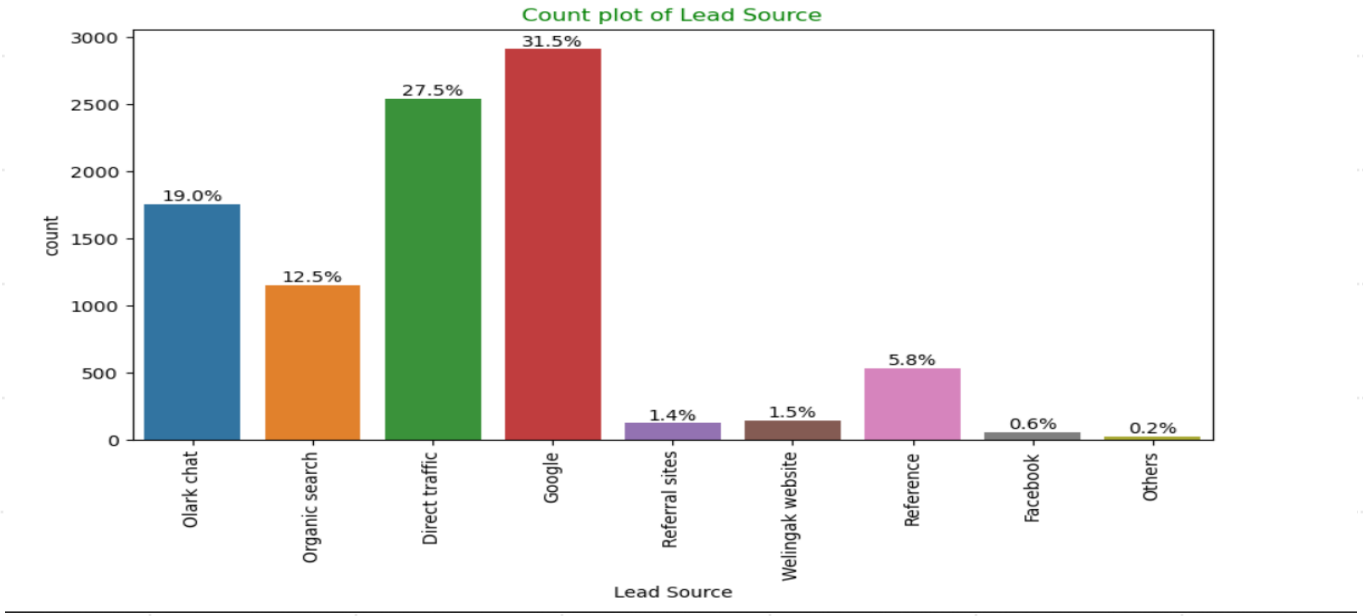
Data Analysis

- Data is imbalanced while analyzing target variable.
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)



Univariant Analysis

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.
- Most entries are 'No'. No Inference can be drawn with this parameters



Bivariant Analysis

- Lead Origin:
- Around 52% of all leads originated from "Landing Page Submission" with a lead conversion rate (LCR) of 36%.
- The "API" identified approximately 39% of customers with a lead conversion rate (LCR) of 31%.
- Lead Source:
- Google has LCR of 40% out of 31% customers, Direct Traffic contributes 32% LCR with 27% customers, which is lower than Google, Organic Search also gives 37.8% `of LCR, but the contribution is by only 12.5% of
- customers,
- Reference has LCR of 91%, but there are only around 6% of customers through this Lead Source

Model Building

- Feature Selection
- The data set has lots of dimension and large number of features.
- This will reduce model performance and might take high computation time.
- Hence it is important to perform Recursive Feature Elimination (RFE) and to select only the important columns.
- Then we can manually fine tune the model.
- RFE outcome ○ Pre RFE – 48 columns & Post RFE – 15 column

Model Evaluation

- Confusion Matrix & Metrics
- Train Data Set
- Test Data Set
- Using a cut-off value of 0.345, the model achieved a sensitivity of 80.05% in the train set and 79.82% in the test set.
- Sensitivity in this case indicates how many leads the model identify correctly out of all potential leads which are converting
- The CEO of X Education had set a target sensitivity of around 80%.
- The model also achieved an accuracy of 80.46%, which is in line with the study's objectives

Predictions on Test Data

- Train – Test
- Train Data Set: Accuracy: 80.46%
- Sensitivity: 80.05%
- Specificity: 80.71%
- Test Data Set: Accuracy: 80.34%
- Sensitivity: 79.82% ≈ 80%
- Specificity: 80.68%

Recommendations

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.

- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
  - Lead Source_Welingak Website: 5.39
  - Lead Source_Reference: 2.93
  - Current_occupation_Working Professional: 2.67