Lead Score Case Study

Studied business problem statements and data dictionaries. Please find the summarized problem statement below.

Identify the most potential leads, also known as 'Hot Leads'. Meaning assign a lead score to each of the leads. 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%.

Note : Data shared has both form initial lead data and sales team updated data

Approach

Drafted approach to explore and to build an acceptable model with good accuracy(80%). Please find the approach followed below.

1. Data Loading
2. Checking
    - 2.1 Size
    - 2.2 Shape
    - 2.3 Data Attributes
    - 2.4 Data type of each Attribute
    - 2.5 Description and Distribution of Data
    - 2.6 Identifying Categorical Data
    - 2.7 Understanding Categorical Data Provided
    - 2.8 Observations 1
    - 2.9 Identifying Continuous Data
    - 2.10 Understanding Continuous Data Provided
    - 2.11 Observations 2
3. Data Selection and Data Correction
    - 3.1 Data Correction
    - 3.2 Data Type Correction
    - 3.3 Data Section and Data Elimination
4. Missing Values Analysis and Outlier Analysis and Handling
5. Data Analysis and Patterns Identification
    - 5.1 Target Data Balanced or Imbalanced?
    - 5.2 Data Isolation Based on Target Variable
    - 5.3 Univariate Analysis on Categorical Data with respective to Target
    - 5.4 Bivariate analysis on Categorical-Categorical Data with respective to Target
    - 5.5 Finding Hidden Correlation among Continuous Data
    - 5.6 Observations
    - 5.7 Univariate Analysis on Continuous Data with respective to Target

Data Exploration

Then, proceeded with data exploration and visualization. Please find data exploration result below

- Number of Leads provided are 9240
- Number of given possible predictors provided are 36
- Number of Target Label in dataset 1 (converted)

Categorical Variables

- Lead Origin
- Do Not Email
- Converted, Lead Source
- Do Not Call
- Last Activity (**Sales Team Generated**)
- Country
- Specialization
- How did you hear about X Education
- What is your current occupation
- What matters most to you in choosing a course
- Search
- Magazine
- Newspaper Article
- X Education Forums,
- Newspaper
- Digital Advertisement
- Through Recommendations
- Receive More Updates About Our Courses
- Tags (**Sales Team Generated**)
- Lead Quality (**Sales Team Generated**)
- Update me on Supply Chain Content

- Get updates on DM Content
- Lead Profile (**Sales Team Generated**)
- City
- I agree to pay the amount through cheque,
- A free copy of Mastering The Interview
- Asymmetrique Activity Index (**Sales Team Generated**)
- Asymmetrique Profile Index (**Sales Team Generated**)
- Last Notable Activity (**Sales Team Generated**)

Continuous Variables

- Total Time Spent on Website
- Asymmetrique Activity Score (**Sales Team Generated**)
- Asymmetrique Profile Score (**Sales Team Generated**)
- TotalVisits
- Page Views Per Visit

Few Observations

- Conversion Rate is ~38 %
- ~ 92% learners chosen Do Not Email
- Lead Quality seems to have lot of has null values
- Many learners from Mumbai and also many not provided by the city. 'Select' has to be imputed with NaN
- Many people not given information about how they hear about X Education
- Lead Source gives the impression that many people hear about X Education through Google, Direct Traffic or Organix Search.
- Profile Score, Index has many None Values
- Many learners not given Profile Info
- 100% Chosen Do not Call(May be default option)
- Specialization many learners not provided the info

- Note 1: Data Has few Duplicate columns which represents the same meaning. It is because of data set containing initial lead data and sales team updated leads
- Note 2: Data has many null values and duplicate categories and Not available information is represented as 'Select' for few attributes
- Less Numerical Data
- Data is positively skewed
- Has outliers in TotalVisits, Page Views Per Page, Activity Score
- Many people spent on website 8 to 30 min

Data Cleaning

With Above knowledge on data proceeded with data cleaning. Please find data cleaning information below.

As per the above observations 'Select' values are replaced with NaN.
Removed data variables with no variance.
Corrected redundant data. Ex: google, Google

After correcting data, proceed with missing values and outliers analysis. Please find actions taken in this process below.

- Dropped variable with NaN percentage 40 or above
- Dropped records with NaN percentage 70 or above
- Imputed NaN values in categorical variables with NA(Not Available)
- Imputed continuous variable NaN with Median

- Extreme Outliers records were dropped
- Data after above modification: Records 9044, Variables 16

EDA

In the next step we performed EDA on data. Please find the results below.

- Concentrate on working professionals
- Reference and Welingak Website conversion rate is high
- Google 50% conversion rate
- Mumbai, Thane & outskirts conversion rate is higher
- the leads which has specialization chances of conversion is bit higher

Data Preprocessing

Before proceeding with preprocessing, the sales team generated data ws had dropped as per the problem statement.

As part of preprocessing , Dummies were created for categorical data , Created Train Test data with split ratio 7:3 and applied MinMax scaling on continuous data.

Model Building

Later proceed with model building.

Recursive Feature Elimination(RFE) is used to select the best 20 features. Also used class as balanced.
Then followed the manual feature elimination method . Dropped P predictors with p value 0.05 and VIF > 5.
At the end of this process we got below predictors.

TotalVisits, Total Time Spent on Website, Lead Source_Other, Lead Source_Welingak Website, Country_NA,What matters most to you in choosing a course_NA,Lead Origin_Landing Page Submission, Lead Origin_Other,
Do Not Email_Yes, Specialization_Hospitality Management, Specialization_NA,
What is your current occupation_Working Professional

Trained model with above predictors and identified accuracy , sensitivity , Specificity scores.
Got cut off as 0.43 after comparing all scores at different thresholds.

Tested model on the test data and assigned lead score to each lead.
Class assigned to leads with cut off 0.43.
Test set sensitivity calculated.
Recorded sensitivity score(0.79).
Verified Specificity(0.8) , Accuracy(0.79), F1 Score(0.78) a score also in accepted rage or not.

Conclusion

- Concentrate on working professionals
- Reference and Welingak Website conversion rate is high
- Google 50% conversion rate
- Mumbai, Thane & outskirts conversion rate is higher
- the leads which has specialization chances of conversion is bit higher
- Model Achieved testing and training sensitivity score 79 and added lead scores to test leads
- Top 3 predictors
    - Total Time Spent on Website (Score : 4.5777)
    - Lead Source_Other (Score : -2.6994)
    - Lead Origin_Other (Score : 2.5672