Case study by:
Sanakausar Kazi
Laxit Goenka

Summary report for X Lead Scoring case study

Aim of the case study:

The case study is to help the X education Company to build up a strategy to find/predict the leads they get as

a. a completely potential lead and
b. not so positive lead

thus enhancing the working efficiency of its Sales team.

For this case study we were provided with

1. CSV file containing past data
2. Data dictionary for columns in CSV

EDA & Modelling:

For this we chose the following methodology for regression analysis:

1. Loading of CSV file into a data frame
2. Exploratory Data Analysis:
    1. Data Cleaning:
    - First check for null values.
        - We have deleted rows having <30% missing values
        - We have deleted columns having >45% missing values
    - Some columns had value as Select in it same have been replaced with appropriate values.
    - Check for datatypes of the columns - All were proper
    - Check for any outliers - No outliers found
    - Check for splitting/aggregation of columns - Not required for any columns
    - Some columns had a single value as No those columns were dropped.
    - Unique columns such as lead number and prospect ID were dropped.
- Final list of columns is
    - Do Not Email
    - Do Not Call
    - Converted
    - TotalVisits
    - Total Time Spent on Website
    - Page Views Per Visit
    - Last Activity

- 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
- Lead Quality
- Update me on Supply Chain Content
- Get updates on DM Content
- Lead Profile
- City
- I agree to pay the amount through cheque
- A free copy of Mastering The Interview
- Last Notable Activity

2. Visualization using Univariate Bivariate and Multivariate analysis
- We have used Heatmap to analyse the correlation between different variables.

4. Dummy Variable Creation
- Binomial category columns were transformed to 1 for a Yes, 0 for a No value.
- Dummy variables were created for all categorical columns. Drop first was passed as true. Total number of columns now is 152.

5. Model building:
- Splitting dataset Scaling of Values Logistic Regression Using Feature Elimination. -
  - we split the data into 70-30 ratio for training and test data.
  - Converted is our y variables
  - Rest of the columns are for now X variables.
  - Lets start working on training dataset:
    - Scaling of variables 'TotalVisits','Total Time Spent on Website','Page Views Per Visit' using standard scaling
    - RFE is used first to select top 30 features.
    - We used GLE model with Binomial family for building model using logistic regression for the above 30 columns.
    - VIF and p-value was used on each step for feature elimination.

- Final list of variables after model building is:

  Do Not Email
  Total Time Spent on Website
  Last Activity_SMS Sent
  Last Activity_Unreachable
  Specialization_Hospitality Management
  Current Occupation_Working Professional
  Tags_Busy
  Tags_Closed by Horizzon
  Tags_Lost to EINS
  Tags_Ringing
  Tags_Will revert after reading the email
  Tags_in touch with EINS
  Tags_switched off
  Lead Quality_Might be
  Lead Quality_Not Sure
  Lead Quality_Worst
  Last Notable Activity_Modified
  Last Notable Activity_Olark Chat Conversation

6. Deciding Threshold for cut-off:
   - Using ROC curve and meeting point of Sensitivity and Specificity, we decided **0.5** as our threshold for cutoff.

7. Predictive Analysis -
   - all probabilities > 0.5 are converted to 1, rest are 0

     Confusion Matrix now is:

     [1812, 122],
     [83, 1430]
   - Accuracy is 95%
   - Sensitivity is 94%

8. Model Evaluation:
   - We evaluated our model on test data and accuracy came to be 95%

Conclusion:

- The strategy found by analyzing the Lead data to find more potential leads are as following:
  1. Focus first on leads having current profession as working professionals first who have spent time on website > 972 (which is mean of total time spent by leads that are predicted by model to be
  2. positive ones). This group has highest probability of lead conversion.
  3. When a lead asks to send an email consider it to be a potential ones and follow up with these
  4. leads.

5. If lead says he/she is unsure there is little chance of conversion hence push these leads to
6. bottom of your list.
7. Leads who have given incorrect number or calls are unanswered or are unreachable more often
8. put these leads at bottom of your targets.
9. Potential leads whose last activity is marked as SMS