LEAD SCORING CASE STUDY

Upgrad Assignment

SUMBMITTED BY,

Josemon Joy, Mohd Ibney Ali

Problem Statement

Leads acquisition through multiple channel

Poor lead conversion rate ~30%

Identify most potential leads i.e. “Hot Leads”

Build a model to assign lead score to each lead-

- higher score > higher conversion chance
- Lower score > lower conversion chance

Target lead conversion rate ~ 80%

Social Media Marketing
Initial Pool of Leads
Lead Nurturing
Hot leads
£ € £ $
Converted leads

Objectives

Build a logistic regression model

Assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Data Understanding and Preparation

Import libraries and read data set

Understanding the data info and description for preprocessing

Dealing with missing values

- Converting missing values to null values

Dealing with null values

- Replacing incorrect entries as null value, e.g. replacing “Select” with nan
- Dropping columns having significant null values
- Replacing null values for important feature using business understanding and mode, e.g. replacing null values in Lead Quality with “Not Sure”
- Dropping rows with low percentage null values

Forming EDA

- Dropping unnecessary columns based on univariate and bivariate analysis

Creating Dummy Variables for categorial columns and replacing the parent columns

EDA

- Univariate analysis
  - Check for value counts for categorical variables
  - Dealing with outliers in numeric variables-
    - In case of 'Total Visits' and 'Total Time Spent on Website' and 'Page Views Per Visit’, all the three columns have outliers
    - Dropping rows with “total visits” >10
    - Dropping rows with “total time spent on website” >1800
    - Dropping rows with “page views per visit” >7
- Bivariate analysis
  - Checking feature relationships with target variable “Converted” using pair plots (continuous variables) and count plots (categorical variables)
    - Understanding important features and unimportant features
      - Important- e.g. Last Activity, What matters of choosing, Tags, Lead Quality
      - Unimportant- e.g. Lead Origin, Specialization
    - Clustering origin of leads as “Indian”, “Non_Indian_Asian”, “Non_Asian”

Model Approach

Test train split

- Scaling continuous variables
- Checking conversion rate
- Identifying and dropping high correlated variables

Model building

- Using RFE for feature selection- no. of features =15
- Checking for p values and VIF Scores
- Deleting features with high p values i.e. >0.05, e.g. tags_wrong number given, tags_invalid number
- Checking for prediction accuracy, precision by creating confusion matrix
- And assigning lead score based on conversion probability
- Plotting the ROC curve to see the trade b/w sensitivity and specificity- curve follows thelft hand border and then the top boarder signifying the accuracy of the model
- Finding the optimal cut-off by plotting accuracy, sensitivity and specificity –
  - The optimal cut off is decided based on the sensitivity and specificity values where it maximizes simultaneously- the final cut off is 0.27

Model Evaluation

Test set results

- Running the model on test data
- Checking for prediction accuracy, precision by creating confusion matrix

Model is capable of predicting 87% customers out of all the converted customers

- The model has an accuracy of 89.8%
- The final model has Precision of 0.8586, this means 85.86% of predicted hot leads are True Hot Leads
- Also we built a reusable code to find the optimum cut off to find out the best precision score.

And assigning lead score based on conversion probability @cut off conversion probability at 0.27

Test set specificity- 0.91

Key Variables & Business Insights

Following are the key variables that should be focused the most on in order to increase the probability of lead conversion

(decreasing order of impact on target variable)

- Tags_Lost to EINS
- Tags_Closed by Horizon
- Lead Source_Welingak Website
- Tags_Busy
- Tags_Will revert after reading the email
- Last Activity_SMS Sent
- What is your current occupation_Working Professional

Following are the key variables that contribute most towards decrease in the probability of lead conversion

(decreasing order of impact on target variable)

- Lead Quality_Not Sure
- Lead Quality_Worst
- Last Notable Activity_Olark Chat Conversation
- Last Notable Activity_Modified
- Tags_switched off
- Tags_Ringing

Recommendations for immediate implementation

Recommendations to Increase Lead Conversion:

Focus on High-Impact Tags:

- Prioritize leads tagged as **"Lost to EINS"** and **"Closed by Horizon"** since they have the highest positive impact on lead conversion.
- Actively engage with leads tagged as **"Busy"** or those who will **"Revert after reading the email"** to improve conversion rates.

Optimize Lead Source:

- Leverage the **Welingak Website** as a primary lead source since it significantly impacts lead conversion positively. Enhance marketing efforts and user experience on this platform.

Enhance Communication Strategies:

- Use **SMS as the last activity** when reaching out to leads, as it correlates positively with conversions.
- Develop tailored strategies for **working professionals**, as this occupation group is more likely to convert.

Recommendations to Mitigate Negative Impact:

Address Lead Quality Issues:

- Reduce the proportion of leads categorized under **"Lead Quality - Not Sure"** and **"Lead Quality - Worst"** by refining the lead qualification process.

Improve Follow-Up Activities:

- Avoid **"Last Notable Activity - Modified"** and **"Olark Chat Conversation"** as these are associated with lower conversion rates. Instead, focus on activities with proven positive impacts.

Handle Unresponsive Leads Effectively:

- Implement strategies to re-engage leads tagged as **"Switched off"** or **"Ringing"**, such as alternative contact methods or tailored messaging.

Operational Adjustments:

Refine Lead Scoring Model:

- Ensure continuous monitoring and recalibration of the lead scoring model based on new data to maintain high accuracy (current accuracy: 89%).

Track Key Metrics:

- Regularly evaluate precision, sensitivity, and specificity to ensure the model performs optimally at the identified cutoff of 0.27