Lead Score Case Study

By :Sweta G. Mohature

Problem Statement

An education company named X Education sells online courses to industry professionals.

X Education seeks to improve its 30% lead conversion rate by identifying 'Hot Leads'—those with the highest likelihood of conversion—to enhance sales efficiency and focus efforts on the most promising prospects.

Key factors influencing conversion include time spent on the website, number of visits, lead source, last activity, lead origin, and occupation. By focusing on these factors, X Education aims to enhance its lead conversion efficiency and increase overall sales.

Goals of the Case Study

There are quite a few goals for this case study:

Build a logistic regression model to 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.

There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well.

Business Objective

The business objective for X Education is to increase the lead conversion rate by identifying and focusing on the most promising leads, termed as "Hot Leads“.

This will be achieved by developing a model to assign lead scores, prioritizing leads with higher conversion potential to improve efficiency and raise the conversion rate from the current 30% to a target of around 80%.

Solution Methodology

Data Collection and Understanding :

Dataset Overview: Obtain the provided dataset with around 9000 data points, including various attributes such as Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc.

Target Variable: Identify the target variable 'Converted', where 1 indicates the lead was converted and 0 indicates it was not.

Data Preprocessing :

Handling Null Values: Identify and handle missing values appropriately. This may involve filling missing values with appropriate measures (mean, median, mode) or removing records with excessive missing values.

Feature Engineering: Create new features if necessary, based on domain knowledge and exploratory data analysis (e.g., combining related features or creating interaction terms).

Exploratory Data Analysis (EDA):

Descriptive Statistics: Generate summary statistics for numerical and categorical variables to understand their distributions.

Visualization: Use visual tools (e.g., histograms, box plots, correlation heatmaps) to identify patterns and relationships between variables and the target variable.

Correlation Analysis: Analyze the correlation between independent variables and the target variable to identify potential predictors for the model.

Model Building:

Train-Test Split: Split the dataset into training and testing sets to evaluate model performance.

Logistic Regression Model:

Develop a logistic regression model using the training dataset to predict the likelihood of lead conversion. Ensure the model outputs a probability score between 0 and 1 for each lead, which can be scaled to a lead score between 0 and 100.

Model Fine-Tuning:

Hyperparameter Tuning: Optimize model hyperparameters using techniques like grid search or random search to improve model performance.
Cross-Validation: Perform cross-validation to ensure the model's robustness and generalizability.

Scoring and Lead Prioritization:

Lead Scoring: Assign a lead score to each lead based on the model's predicted probability of conversion.
Lead Categorization: Categorize leads into 'Hot Leads' (high score) and 'Cold Leads' (low score) based on predefined thresholds.

Data manipulation

Handling Missing Values:

Identify Missing Values: Use methods such as .isnull() or .isna() to detect missing values in the dataset.

Impute or Remove:

Numerical Data: Impute missing values using mean, median, or mode, or remove records with excessive missing values.
Categorical Data: Handle missing values in categorical data by imputing with the most frequent category or creating a new category like 'Unknown'.

Handling Categorical Variables:

Identifying 'Select' Values: Recognize 'Select' values in categorical variables as they are equivalent to null values and replace them appropriately.

Feature Engineering:

Creating New Features: Derive new features from existing ones if they add value to the model. For instance, creating interaction terms or combining related features.
Normalizing Numerical Features: Scale numerical features to ensure they have similar ranges, which can help improve the performance of the logistic regression model. Techniques like Min-Max scaling or Standard scaling can be used.

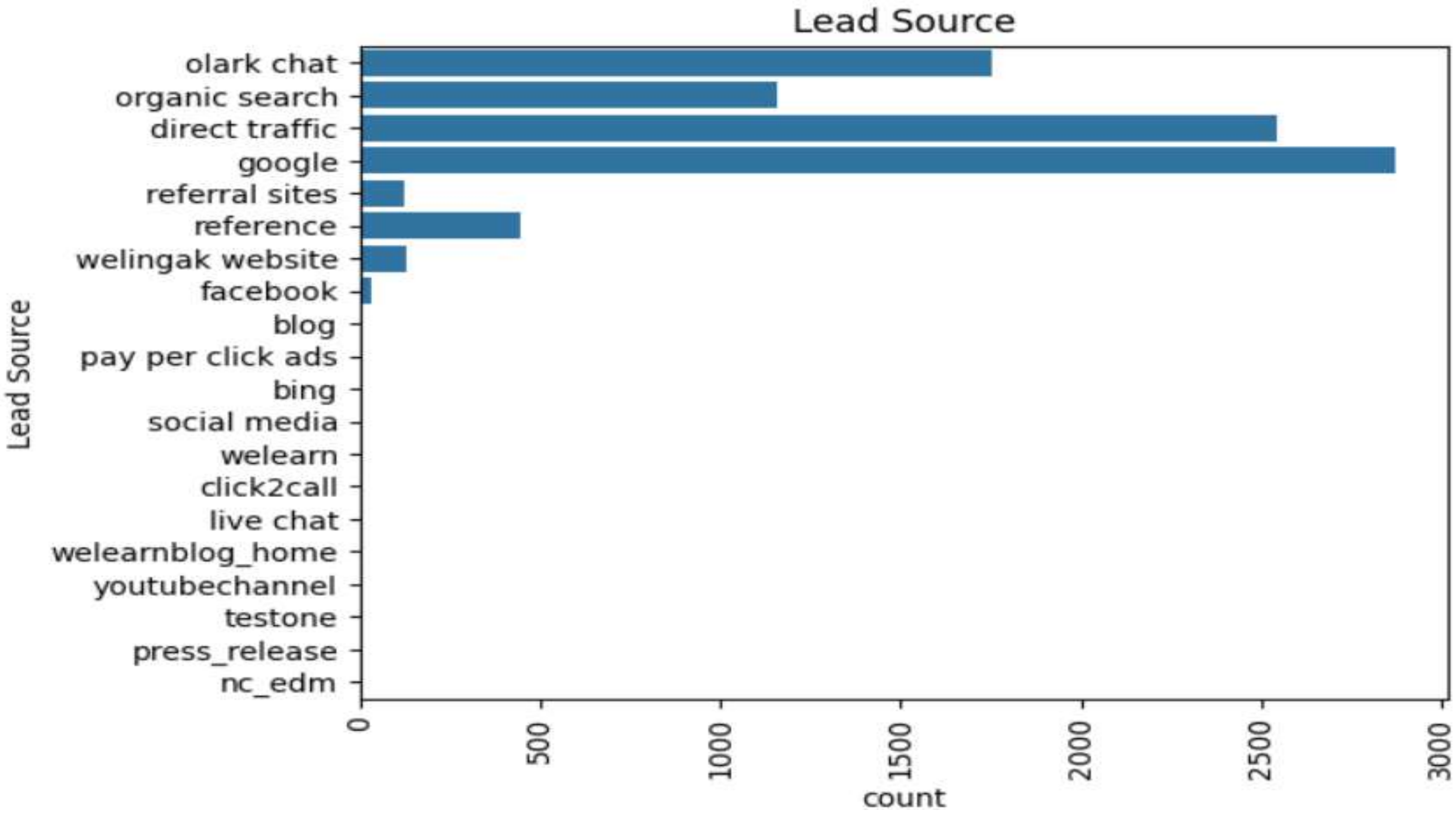
Dealing with Outliers:

Detecting Outliers: Use statistical methods such as the Z-score or IQR (Interquartile Range) method to identify outliers in the numerical data.
Handling Outliers: Decide whether to remove or cap outliers to reduce their impact on the model.

Feature Selection:

- **Correlation Analysis**: Analyze the correlation between features to identify and remove highly correlated features that do not add significant value individually.
- **Recursive Feature Elimination (RFE)**: Use RFE to iteratively remove less important features based on model performance.
- **Regularization Techniques**: Apply L1 or L2 regularization to select features that contribute most to the prediction.

EDA


Lead Source
olark chat
organic search
direct traffic
google
referral sites
reference
welingak website
facebook
blog
pay per click ads
bing
social media
welearn
click2call
live chat
welearnblog_home
youtubechannel
testone
press_release
nc_edm
Lead Source
0
500
1000
1500
2000
2500
3000
count