Lead Scoring Case Study - Summary Reports

Problem Statement

A Company sells online courses to working professionals. These courses are advertised on different channels - print and electronic media adverts. The company collects the data of all individuals who land on their page. Employee of the company calls all those individuals to sell the course. The current conversion rate is 30% which the company wants to improve.

Analysis Approach

Our approach to analysis are as follows:-

- Load and observe the data.

  We have loaded the data in pandas dataframe and have observed different columns present in it. The columns are - Numerical and Categorical columns.

  Also, we have observed that the data for each customer is collected at different times. They are :-
  - Web Collected Data : Those data which are collected when the customer lands on the company web page
  - Sales Team Collected Data : Those data which are collected when the company employee makes a call to the customer.

- Exploratory Data Analysis

  We have performed the exploratory data analysis on the given dataset. We have performed.
  - Null Value treatment
  - Outlier Detection

  We have then analysed the column for data distribution and skewness

- Data Preparation

  For data preparation, we have performed,
  - Bucketing of the numerical variables
  - Encoding of the categorical variables
  - Train Test Split

- Feature Selection

  We have done following:
  - Statsmodel Api - to find the importance of the feature
  - VIF calculation - to check and remove multicollinearity
  - RFE - to find which all features need to be selected

- Model Building and Testing

  We have used sklearn logistic regression to build the model by adding one feature at a time. We have tested it on test data.

- Finding Suitable Threshold

  We have to check the value of precision, recall at different values of threshold to find the suitable threshold of the given scenario.

Result

Without the use of the columns which are gathered by the sales team, we are able to predict the outcome with 79% accuracy. The important predictors in decreasing order of importance are as follows :-

- leadOrgn
- leadSrcEncoded
- totalTimeSpent
- specialization
- masteringInterview
- pageView
- matterMost
- Email

Many of the columns are categorical columns. The selection with different impacts are as follows:-

Positive Impact	**Negative Impact**
NC_EDM Live Chat WeLearn Flexibility & Convenience Reference Olark Chat Welingak Website Click2call Social Media Google Facebook Direct Traffic Lead Add Form Healthcare Management Banking, Investment And Insurance Travel and Tourism E-COMMERCE Media and Advertising Business Administration Finance Management Organic Search masteringInterview Marketing Management	Lead Import International Business pageView bing Services Excellence API No selection Rural and Agribusiness Landing Page Submission Better Career Prospects Human Resource Management email Retail Management Other welearnblog_Home google youtubechannel testone blog Pay per Click Ads Press_Release

Operations Management Supply Chain Management totalTimeSpent Referral Sites Hospitality Management IT Projects Management E-Business	

Learnings

During this exercise, the learnings which we have gathered are as below:-

- Analyse the source of data and at which point in the process they are collected.This will have impact on the feature selection.
- Value present in the data and their meaning. Null value can be obscured with other values like ‘select’
- Bucketing/Binning of the continuous variables as this will add stability to the model.
- One hot encoding of the categorical variable.
- We have to analyse each predictors to see the distribution of data in them. Highly skewed predictors can be dropped as they will not add any value.
- We have to look into the distribution of the target variables and check for class imbalance in them. If class imbalance is present we need to choose the suitable class weightage while building the model.
- Process to choose the threshold value so that the business need can be satisfied by achieving proper tradeoff between the precision and recall score.