X Education Lead Scoring Case Study Report

Introduction

X Education which is a online Industry Professional courses seller markets courses on several websites like Google. When a person lands on the website to browse courses they fill up a form providing email and phone numbers, These are classified as leads. The Company also gets leads from referrals as well.

Once the leads are gathered the sales team steps in to contact the lead and tries to convert them. The Sales team has a typical conversion rate of 30%.

Problem Statement

X Education would like to know the possible leads that are sure to convert and most of the efforts are provided to those leads with the high lead score.

Solution

Build a Logistic regression model that would analyse the given data and provide which leads to be concentrated on more by providing a lead score to them.

Approach

The process is started with importing the required libraries and loading the data into a data frame for analysis. The Data was then analyzed for Null checks and NaN value checks. In some of the columns there were more then 30% of data missing. Therefore those columns were removed from the data frame. Some of the columns where having less than 30% of missing data, therefore different techniques were used to fill the missing data with Median and Mode. Some of the Columns data from transform into 1 and 0 for better processing.

Once the Data cleaning steps are completed EDA is performed to find correlations between the columns.

For the categorical columns dummies were created and appended to the data frame. The columns which were used for dummy creation is dropped to avoid redundancy.

Scaling

The Data is scaled using MinMaxScaler and Feature selection is done. After this Model building process is started

Model building

Using the statsmodel library the model is build and the training data is used. The model result is obtained and analyzed. To finetune the model the columns with **P>|z| > 0.05** is dropped. After droping the column the model is built again. This process is followed until all the columns/ feature has **P>|z| <= 0.05**

In the model built there were 8 iterations until the required result are visible.

VIF is calculated for the remaining features and once all results are ready. Model is evaluated.

Evaluation

The results from the last model is use for evaluation by trying the predict the results and the accuracy , recall and precision is calculated using the metrics in sklearn library.