Lead Score Case Study

Prepared by Mukul Dutt Bharadwaj and MS Alexander Sooraj, DSC59-Aug2023 Batch. Prepared and Submitted on 25 Feb 2024

Case Study Summary

Problem Statement

X Education offers online courses for industry professionals and promotes them on various websites and search engines like Google. Despite attracting many potential customers, only a small percentage end up purchasing courses. X Education aims to increase the conversion rate of potential customers to paying customers. Currently, they convert about 30% of leads into customers by targeting those interested in their courses. However, their lead generation strategies are not very effective in converting leads into customers.

Steps Followed

1. Data Cleaning and Exploratory Data Analysis
2. Data Preparation
3. Model Building
4. Conclusion and Recommendation

Data Cleaning and Exploratory Data Analysis

1. The dataset was loaded into the python notebook and explored statistically and visually to get an idea of outliers, distribution of data, feature redundancies, and so on.
2. Correlations between the variables were also identified using heatmap and boxplots.
3. The duplicate variables were removed.
4. Some of the columns had string values “Select” which meant that the user did not select a particular value in the form. This means it is as good as null values.
5. Columns with more than 45% null values were removed from the model building.
6. Outliers were removed statistically from the data from two numerical variables to avoid noise in the model building.
7. A few rows with missing values were removed. 98% of the dataset was remaining for the data modelling out of 9240 rows in the original dataset provided

Data Preparation

1. Categorical variables were converted into numerical data of 0 or 1 using dummy variables.
2. The dataset was split into training and testing datasets in a ratio of 7:3
3. Data points were standardized to similar scale using Standard scaler for numerical features to avoid bias of some variables in higher scales in the model.

Model Building

1. Initial model was created by selecting 15 variables using the RFE technique.
2. Insignificant variables that were identified were removed to optimize the model for better model. VIFs were also checked to see the multicollinearity between the model features.
3. An important step in logistic regression is to find the optimal cutoff for the probability to fit the business needs and to have good accuracy, sensitivity and specificity. We obtained 0.186 as the cutoff from the plot between accuracy, sensitivity and specificity and probability range.
4. Recall and Precision view was then considered to finalize the cutoffs for the prediction in test dataset. The cutoff in the Recall-Precision graph was obtained as 0.38
5. Model evaluation was completed by checking the values of all the performance measures namely ROC curve, between accuracy, sensitivity and specificity, Recall and precision. All of them were in acceptable range.
6. Predicted values were calculated using the model. Lead score is 100 multiplied by the log odd which is predicted, to have a value between 0 and 100.

Conclusion and Recommendation

1. People spending higher than average time are promising leads, So, targeting them and approaching them can be helpful in conversion
2. SMS messages can have a high impact on lead conversion
3. Landing page submissions can help find out more leads
4. Management resources has high conversion rates. People from these specializations can be promising leads
5. References and offers for referring lead can be a good source for higher conversions
6. An alert messages or information has seen to have high lead conversion rate
7. The model shows high close to 78% accuracy
8. The threshold has been selected from accuracy, sensitivity, specificity measures and precision, recall curves
9. The model shows close to 96% sensitivity and close to 67% specificity
10. The model finds correct promising leads and leads that have less chances of getting converted
11. Over-all this model proves to be accurate.