Lead-Scoring-Logistic-Regression

X Education needs a machine learning model which will increase their lead conversion beyond 80%. To solve this problem logistic regression model is developed which predicts a lead conversions and also assigns a lead score to each Lead to determine the probability of it's conversion.

The steps followed in developing this model is as follows:-

1.Importing necessary libraries We have used Numpy, Pandas for dataframe, Matplotlib and Seaborn for visualization and SKLearn for TrainTest Split, Scaling and Model evaluation, Statsmodel.api for building the LRM.

2.Importing and Describing After importing the dataset, we replace all the 'Select' values in the Dataframe with NULL since they are all NUL values where the user has not selected anything from the Drop-Down menu.We drop the columns with more than 70% Null. After the imputations are complete, we remove those rows which have less than 2% null value in their columns. Finally, we have a Dataset with zero null values.

3.Handling anomalies and outliers Columns such as Page Views Per Visit and Total Visits has quite a few outliers, thus we cap the datapoints . Also, we drop those columns which are highly imbalanced and does not play as part in deciding the outcome of the converted column.

4.EDA We compare all the columns with respect to 'Converted' column to analyse which attributes play a part in deciding the outcome of a Lead Conversion. We decide which column to use in the Model building. Countplot has been used predominantly for this purpose

5.Splitting the data into Train and Test Data The final dataset is separated randomly into a 70-30% split dataset for Training and Testing purpose. This is done using SKLearn's train_test_split function.

6.Scaling the data Since the dataset has data of different dimensions, we need to scale the data in order to make it suitable for a regression model. Since there are several dummy variables created from the categorical values, we opt for MinMax scaling . For this we use SKLearn's MinMaxScaler function. We fit and transform the train dataset only.

7.Building the model Using the GLM method we build a regression model in the train dataset. We use RFE function to identify the top 15 features which we can use in building the model. We drop the rest of the columns from the train dataset and build another model using statsmodel. We check the VIF of these 14 variables and find out that all the VIF

8.Evaluating the model We determine the Confusion Matrix and the parameters like Sensitivity, Specificity, etc. We plot the ROC curve ,the Accuracy, Sensitivity and Specificity plot, we

determine the optimal cut-off at 0.2 and got an accuracy of 91.93%.

9.Making predictions on the test dataset We scale the test dataset with only transform and then predict the probabilities using the final model. On the test dataset we use the optimal cut-off of 0.39 and get an accuracy of around 90.45%

10.Generating Lead Scores for the sales team for the full dataset We provide that lead score in a range of 1-100 based on the probability determined by the final model.

Learnings

In case of X Education, the sales and marketing team must target leads who have,

Spent more time on their website

Visited their website a greater number of times

Their Last activity was SMS or Email Following these traits would increase the lead conversions to a higher percentage.