Lead Scoring Case Study Summary

Problem statement :

X Education sells online courses. Leads generated from various sources are captured. There is other metadata around lead that is captured for each lead. Team is assigned to nurture hot leads and convert such potential leads to confirmed opportunity (leads).

Solution :

First effective way of working on leads is to start with the hot leads i.e. leads that have higher probability of getting converted. This will not only result in higher conversion ratio but also effective use of time. Time spent on nurturing hot leads can be increased but whereas time spent on leads with low score (cold leads) can be minimized.

Determining hot and cold leads can be done by using a logistic regression model. Using the meta data provided for each lead, we will build a logistic regression model and assign lead score to each lead.

I. Data Analysis : Step 1

-There are columns with higher missing percentage in the data. Also, there are columns where default value “Select” is populated. We will first consider this as missing values and then apply same missing value treatment for such values.

-Categorical columns where percentage missing value is less than 5 will be imputed with mode value.

-Quantitative variables where % missing value is less will be imputed with median value. Statistical analysis indicated that there isn’t significant difference between median and mean for these columns and hence imputing with median should not create issues.

-Categorical columns where percentage missing value is greater than 70% will be dropped.

-Other missing values will be treated as missing values since imputing can exaggerate the data.

II. Data preparation : Step 2

- By using Boxplot and descriptive statistics we can indicate that there are outliers in the dataset. Removing the outliers will result approximately in 9% data loss.

- We will not be removing the outliers as this will help us in assigning lead score to all leads. Final review of model does indicate that metric (Accuracy, Sensitivity and Specificity) is good and hence we will not remove outliers.

-Quick bivariate analysis indicates few categorical variables/levels are critical for lead conversion. We will use this for conclusion.

-Since logistic regression uses numerical data, we will convert categorical data using below 2 techniques.

i. Dummy Variables – Categorical variable with low or moderate level will be treated using the dummy variables.

ii. Label Encoding – We will use label encoding for variables with higher level. This is to avoid drastic increase in the dataframe size.

-Columns with no variance i.e. the columns with single constant value will be dropped since they add no information or dimension for model building.

-Plotting a heatmap indicates some correlation between variables. VIF will be further used during model building.

III. Model Building : Step 3

-Since dataframe is huge we will use both RFE and PCA to determine which techniques gives us a better model.

-Now we will split the data into train dataset and test dataset. We will train the train dataset and make predictions on test dataset.

-Numerical data will be scaled using standard scaler to ensure data is on same scale and computationally efficient.

-Below functions are created to perform repetitive tasks :

i. Createmodel – Takes dataframe as input, prints model summary, VIF and return model.

ii. Confscores – Takes confusion matrix as input and return accuracy, sensitivity and specificity.

iii. Calctrainseult – Takes cutoff as input and return confusion metrics and scores accuracy, sensitivity and specificity

- RFE –

i. We will use RFE to identify top 20 variables to start modelling.

ii. Criteria used for tuning the model i.e. dropping variables.

1. High p-value -> variable not significant
2. High VIF -> high collinearity with another variable

iii. Model6 is a good model based on statistical summary.

-ROC and AUC confirms that we have a decent model.

-We will use below technique to final optimal cutoff value

i. Plot accuracy, sensitivity and specificity
ii. Plot recall and precision.
iii. Since there is a trade-off between sensitivity vs specificity it is important to find optimal cutoff.

IV) Making prediction : Step 4

-Use model6 and optimal cutoff to make predictions on test dataset.

Use PCA :

-We will use PCA to check if this yields a better model.

-Using PCA helps in dimensionality reduction and solves for multicollinearity issue.

-Making predictions using model build using PCA gives decent results but presents below challenges.

-Scoreless that model build without using PCA.

-Identify original variables/factors leading to high score.

Model Selection and Lead Score :

-Use the model build using the RFE technique for final prediction.

-This model gives best score and is easy to make suggestion and interpretations.

-We will assign Lead score to each lead using probability predicted by
model (Lead Score = Predicted Probability * 100)

-Create a dataframe to and plot conversion vs cutoff.

Conclusion : Step 5

There are top 3 features that contribute to decision are:

- Tags
- Lead Quality
- Asymmetrique Profile Index

The Top 3 categories that contribute to decision are :

- Lead Origin -> Landing Page Submission
- Lead Origin -> Lead Add Form
- Lead Source -> Olark Chat

Learnings :

EDA is extremely important step prior to building model. Some of the Key insights from EDA helps in treating the data correctly.

Data cleaning helps in building efficient model. Steps like missing value imputation, scaling, outlier treatment must be performed at minimum to ensure quality of data is not compromised.

i. Missing value : Columns with higher percentage of missing values can be dropped whereas columns with lesser percentage can be imputed.

ii. Outlier treatment : Outlier can impact model and result in less effective model. Hence care should be taken to treat outlier effectively. Care should also be taken to ensure that this does not result in significant loss of data.

iii. Scaling : Quantitative columns is scaled so as to ensure that they are on same scale. RFE is efficient technique to identify key features to start building model. On other hand PCA is helpful in dimensionality reduction by building new principal components. Functions to perform repetitive steps can help in building a modular code. This also help in reusability of the code. Understanding trade-off between sensitivity and specificity is key in

determining ideal optimal cutoff for the mode. Confusion metrics is good indicator to determine how model performs. Accuracy, sensitivity, and specificity can be derived from confusion metrics.