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: –

Effective way of working on the leads is to start with 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 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.

a. Data Analysis-
   a. There are columns with higher missing % in the data. Also, there are column where default value "Select" is populated. We will be initially considering this as missing values and apply same missing value treatment for such values.
   b. Categorical columns where % missing value is less (<5) will be imputed with mode.
   c. Quantitative variables where % missing value is less will be imputed with median. 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.
   d. Categorical columns where % missing value is greater than 70% will be dropped.
   e. Other missing values will be treated as missing values since imputing might exaggerate the data.
b. Data preparation-
   a. Boxplot and descriptive statistics indicate that there are outliers in the dataset. Removing the outliers results in approximately 9% data loss.

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

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

d. Since logistic regression uses numerical data, we will convert categorical data using below technique.

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

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

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

f. Quick heatmap indicates some correlation between variables. VIF will be further used during model building.

c. Model Building-

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

b. Split data into train and test dataset. We will train the train dataset and make predictions on test dataseet.

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

d. 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, specificity)

e. 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.

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

    g. We will use below technique to final optimal cutoff value

        i. Plot accuracy, sensitivity, specificity

        ii. Plot recall, precision.

        iii. Since there is a tradeoff between sensitivity vs specificity it is important to find optimal cutoff.

d. Making prediction-

    a. Use model6 and optimal cutoff to made predictions on test dataset.

e. Use PCA-

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

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

    c. Making predictions using model build using PCA gives decent results but presents below challenge.

        i. Score less that model build without using PCA.

        ii. Identify original variables/factors leading to high score.

f. Model Selection and Lead Score-

    a. Use the model build using the rfe technique for final prediction.

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

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

    d. Create a dataframe to and plot conversion vs cutoff.

g. Conclusion-

    a. Top three features that contribute to decision

        i. Tags

        ii. Lead Quality

        iii. Asymmetrique Profile Index

    b. Top three categories that contribute to decision

        i. Lead Origin ==> Landing Page Submission

        ii. Lead Origin ==> Lead Add Form

        iii. Lead Source ==> Olark Chat

h. Learnings-

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

b. 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 value 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 should bee scaled to ensure that they are on same scale.
c. 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.
d. Functions to perform repetitive steps can help in building a modular code. This also help in reusability of the code.
e. Understanding trade-off between sensitivity and specificity is key in determining ideal optimal cutoff for the mode.
f. Confusion metrics is good indicator to determine how model performs. Accuracy, sensitivity, specificity can be derived from confusion metrics.