Summary

X Education, online courses selling. There are leads captured from various sources. There is other metadata for lead captured for each lead. Team for nurturing hot leads is assigned and conversion into confirmed opportunity (leads).

Solution: -

An effective way of working with the leads would be starting with hot leads i.e. those leads with more chances of getting converted. Not only would this result in a higher conversion ratio but also the effective use of time. Time spent on nurturing hot leads would increase and time spent on the leads with low scores (cold leads) could be reduced.

This can be done through a logistic regression model to determine hot and cold leads. Using the meta data provided for every lead, we will build a logistic regression model and assign a lead score to every lead.

a. Data Analysis-

a. There are columns with a 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. Numerical variables whose % missing value is less will be imputed with median. Statistical analysis shows that for the columns, there isn't much difference between the median and mean, hence imputation using median should not lead to any problems.

d. Categorical columns whose % missing value is above 70% will be dropped.

e. The remaining missing values will be considered as missing values because imputing might exaggerate data .

Data preparation-

a. Boxplot and descriptive statistics reveal the presence of outliers in dataset. If the outliers are removed, that'll lead to approximately 9% data loss.

b. We are not removing the outliers since this way we will be able to assign the lead score to all the leads. Review of final model does indicate that the metric (Accuracy, Sensitivity, Specificity) is good, hence we will not remove the 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. Data will be splitted into train and test dataset. We will train the train datataset and predict 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. Create model - 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)

d. 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. Very High VIF, it means there is multicollinearity with some other variable. iii. Model6 - Statistical Summarises indicates that is a good model

f. ROC and AUC States that we have a good model.

g. We will use following technique to finalize the optimal cutoff value

i. Plot accuracy, sensitivity, specificity

ii. Plot recall, precision.

iii. Since there is trade-off between sensitivity vs specificity it is important to determine the optimal cutoff value.

e. 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 gives any improvement in model.

b. Using PCA also helps to reduce dimensionality and solves multicollinearity problem.

c. Prediction made by model build using PCA gives good result but present with below difficulty

i. Score less than 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 data frame to and plot conversion vs cutoff.

g. Conclusion-

a. Top three features that contribute to decision

i. Tags

ii. Lead Quality

iii. Asymmetries 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 very crucial step before building the model. Important insights

derived from EDA will be useful for proper treatment of the data.