Lead Scoring Case Study Summary

Problem Statement

X Education sells online courses to industry professionals and It wants to know its most promising leads.

The company requires a model to be built wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance

Summary

1. Loading Datasets

2. Data Understanding

- Understanding Each variable using provided Data Dictionary
- Checking Structure of dataset
- Checking other attributes like using info(), describe(), etc.

3. Data Cleaning

- We checked for duplicated rows, none were found
- We checked if any rows had more than 70% fields missing, none were found.
- The dataset had few columns with more than 40% data missing while few had less than 2% data missing, Some columns had about 15 to 40% data missing
- The columns which had more than 35% data missing were dropped as imputing these columns would certainly deprecate the data quality.
- For columns with less than 2% missing data we dropped the rows having these missing values; After which 98% of total data was retained
- All the numeric columns had less than 2% data so they we handled during the above process
- The remaining categorical columns which had between 15% to 30% missing values but also highly skewed data were dropped.
- Columns not having highly skewed data were imputed using the modal value of that column.
- Columns having improper data type were changed to appropriate ones.
- Few categorical columns had a lot of categories, so categories having very less value counts as compared to other categories were clubbed together into a single category named ‘Other’

4. Handling Outliers

- 2 out of the 4 numerical variables had outliers which were in huge amount, these were capped to 99.3rd percentile value

5. Data Analysis

- A quick EDA was done on all the remaining variables in the cleaned dataset where we checked the imbalance in ‘Converted’ variable
- Segmented Univariate Analysis with respect to ‘Converted’ variable was performed on each variable

6. Data Preparation

- All the sales team generated columns were dropped as sales team generated columns come after lead generation
- All variables having only 2 categories were converted to binary numeric variables with 1’s and 0’s
- For categorical columns having more than 2 categories, Dummy Variable were created of such variables and the original variables were dropped
- Then we performed Train-Test Split on the dataset and numeric features were scaled using StandarScaler method
- Then we checked the correlation between variables using heatmap

7. Data Modelling

- Firstly, a basic model was created using all variables then using RFE, 18 features were selected
- Variables having high p-value and high VIF were dropped while creating a new model every time a feature was dropped
- Our final model had 10 variables

8.Model Evaluation

- A dataframe was created with Converted variable, probabilities of Converted variable and a ‘predicted’ column containing 1’s if probability was above 0.5 and 0’s otherwise, here 0.5 was the assumed cutoff
- Then we plotted the ROC curve with respect to our final model and the area under curve came out to be 83% which further validated the model
- Then we plotted Sensitivity-Specificity and Accuracy plot from which the optimal cutoff came out to be 0.3 with Accuracy = 78.8%, Sensitivity = 74.8%, Specificity = 81.2%
- Then using Precision-Recall tradeoff plot the optimal cutoff came out to be 0.35 with Accuracy = 79.5%, Precision = 72.8%, Recall = 72%
- A dataframe was created for both cutoff’s, 0.3 and 0.35 with Converted variable, probabilities of Converted variable and a ‘predicted’ column containing 1’s and 0’s depending on the selected cutoff
- Finally, cutoffs of 0.3 was selected as the optimal cutoff because of sensitivity being high
- Then, predictions were made on the test dataset with cutoff being 0.3
- We achieved a final accuracy of 80.2%, sensitivity = 75.8% and specificity = 82.9%