Leads Scoring Case Study

Harsh Pandya

IIITB, DS40 || Nov-22

A brief summary report in 500 words explaining how we proceeded with the case study and the learnings that we gathered.

Answer:

Below is the report describing the steps and description about the proceeding and results obtained:

Data loading, inspection and cleaning

- Starting with the loading of the data set to the file and inspected file by checking the shape, information and description.
- Then to clean the dataset, we choose to remove the redundant variables/features.
- After removing the redundant columns, we found that some columns are labeled as 'Select' which means the customer has chosen not to answer this question. The ideal value to replace this label would be replacing them as null value as the customer has not opted for any option. Hence, we changed those labels from 'Select' to null values.
- Now, as we calculated the null value percentage, we removed columns having more than 30% null values.
- For remaining missing values, we have imputed values with maximum number of occurrences for a column.
- We found that a few columns have two identical label names in different formats (capital letter and small letter). We fixed this issue by changing the labels names into one format.

Data Transformation:

- By changing the multicategory labels into dummy variables and binary variables into '0' and '1'.
- And after that by creating a boxplot, we checked the outliers and created bins for them.
- Removed all the redundant and repeated columns.
- Data Visualization
- A quick EDA is done to check the condition of the data. It was found that a lot of elements in the categorical variable are irrelevant. The numeric values seem good and no outliers were found.
- Univariate Analysis of Numerical variables
- Analysis of Data imbalance in TARGET column i.e. 'Converted'
- Bivariate Analysis - This includes analysis of categorical variables against target variables.
- Multivariate Analysis
- Dropping the variables which are irrelevant.
- After this, we plot a heatmap to check the correlations among the variables.

Data Preparation:

- Split the dataset into train and test dataset
- Scaling the dataset.

Model Building:

- We created our model with RFE count 15 and 12 and compared the model evaluation score like AUC and chose our final model with RFE 12 variables as it has more stability and accuracy than the other.
- For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
- We found one convergent point and we chose that point for the cutoff and predicted our final outcomes.
- We checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs.
- Prediction made now in the test set and predicted value was recorded.
- We did model evaluation on the test set like checking the accuracy, recall/sensitivity to find how the model is
- We found the score of accuracy and sensitivity from our final test model is in acceptable range.
- We have given the lead score to the test dataset for indication that high lead scores are hot leads and low lead scores are not hot leads.

Conclusion:

Learning gathered are below:

- Test set has accuracy, recall/sensitivity all in an acceptable range.
- In business terms, our model is having stability and accuracy with adaptive environment skills which means it will adjust easily with the company's requirement changes made in the coming future.

Top features for good conversion rate:

- Tags_Closed by Horizon,
- Tags_Lost to EINS,
- Tags_Will revert after reading the email

The top 3 categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion:

- Tags,
- Lead Source,
- Lead Origin