Leads Scoring Case Study - Summary Report

-By
Rahul Biswas
Suman Prabhakar
Sonal Padole

Below is the approach/ summary report describing the steps of how we approach building our model

1. **Data loading, inspection and cleaning**
   - Initiated the process by loading the data set to the file and inspected data by checking the shape, information and description.
   - Cleaned the dataset - Removed redundant variables/features.
   - Some columns were labeled as 'Select' which meant that the customer did not answer the question.These were changed from 'Select' to null values.
   - Removed all columns that contained >40% NULL values
   - For remaining missing values, we have imputed values with the maximum number of occurrences for a column.
   - We grouped the missing values in Specialization to Others
   - Few columns have two identical label names in different cases, so we converted them to the same case.
2. **Data Transformation:**
   - By changing the multicategory labels into dummy variables and binary variables into '0' and '1'. b.
   - Checked and treated the outliers.
   - Removed all the redundant and repeated columns.
3. **Data Visualization:** Conducted an EDA to check the data..
   - 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.
4. **Data Preparation:**
   - Split the dataset into train and test dataset
   - Scaling the dataset.
5. **Model Building:**
   - We created our model with RFE count 20
   - 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.

6. 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