Lead scoring case study - Summary

Problem statement:

X Education company sells online courses to the industry professionals. The company getting the details of the leads through fill up forms provided in the website or through referrals. This will be given to the sales team to make calls, sending emails and SMS etc. Even though the company getting lot of leads, the lead conversion rate from the lead to paying customer is poor and the conversion rate is around 30%.

The company CEO wants to build the model to predict the hot leads and cold leads. So, the sales team will start with the hot leads to use the time effectively to improve the conversion rate. Also, expected to get 80% accuracy from the model predictions.

Solution summary:

1. Reading the dataset:

- Imported the necessary libraries and validated the given dataset.

2. Cleaning data:

- Converted the categorical labels to lower case to avoid having duplicate label due to case mismatch.
- Removed the label 'select' from the dataset which is the default value while filling the submission form. So, we can handle this value with missing values.
- Dropped the columns which are having only one label or highly skewed with the one label.
- Validated missing value by row wise and there are no rows with missing value >70%.
- Dropped the columns which are having the missing values more than 40%.
- While checking in column wise, if any of the column having missing value less than 2%, removed the rows which are having missing values.
- In the columns 'City', 'Specialization', 'Tags' and 'What is your current occupation' replaced the missing values.
- Dropped the columns 'Prospect ID' and 'Lead Number' which are having unique value for each row.

3. EDA:

- Since we are building model to make prediction to provide the leads details to sales team, removed sales team data from the dataset.
- Validated correlation between the numerical columns.
- Columns 'TotalVisits' and 'Page Views Per Visit' are having outliers and handled by removing the values which are beyond 99th percentile.
- Bivariate analysis done on the categorical columns to get the insights.
- Data retained through these cleaning and EDA process is 96.58%

4. Data preparation for modeling:

- Binary columns 'Do Not Email' and 'A free copy of Mastering The Interview' are converted to 1 and 0.
- Created dummy variables for the categorical columns.
- Input 'x' and target 'y' dataset created.

Lead scoring case study - Summary

- Training -70% and test-30% datasets are split for both input and target dataset.
- Numerical variables are rescaled in the training dataset.
- Correlation validated in the input training dataset and removed the columns which are having high correlation.

5. Model building:

- Used automated approach sklearn's RFE method to reduce the features to 20.
- Manual approach done using statsmodels GLM() function. And calculated the p-value and VIF value for each feature them removed one feature from the current selection based the p-value and VIF scores. Steps are repeated till we get the p-value and VIF scores of all features less the 0.05 and 5 respectively.

6. Model evaluation:

- Predictions are made using the final model on the training dataset and probability values are stored.
- Optimal cutoff value calculated using both sensitivity-specificity and precision-recall.
- Accuracy, sensitivity, specificity, TPR and FPR scores are calculated.
- Predictions are made with the test dataset for probability values and calculated the conversion value based on the optimal cutoff values.

Conclusion:

- We have considered the optimal cut off using both sensitivity - specificity and precision-recall tradeoff.
- We got accuracy 79%, sensitivity 77% and specificity 79% for both training and test dataset.
- The conversion percentage as per the model prediction is around 77%
- With the final model, we can find that the below features are important to predict the hot leads.
    - The total time spent on website and the total visits.
    - When the lead source is 'welingak website' and 'google'.
    - Lead origin is 'lead add form'.
    - Leads who are working professionals.