Lead Scoring Case Study Summary

Went through the case study and understand the business problem and Business objects to improve the lead score and generate more number of hot leads.

Performed Following steps to generate the end result.

1. **Required Library**: Import the required library to support and perform the required process.
2. **Load the Data & Data Understanding**: Load the data Leads.csv and go through the details of data structure.
    - Check the data types of cols
    - Numerical variable distribution etc..
    - Missing value analysis
    - Duplicate row details

3. **Data Cleaning**: Data cleaning play critical role for ML model building and need to remove the unwanted data to make stable model.
   Below are the points covered.
    - Clean the data and find the unique value present into the dataset (i.e. duplicate rows of data if any need to drop. But for this use cases we don't have any duplicate rows) and drop them during the process of data cleaning
    - Convert sting data into lower case
    - Replace 'select' with NaN
    - Drop the columns having null values grater equal to 45%
    - Remove all the columns have near zero variance or imbalance data
    - Merge the highly skewed categorical variables to single category.
    - Impute the missing values with 'not given' for categorical variables.
    - Impute the missing values with median for numerical variables, as we have right skewed data

4. **EDA**: Univariate, Bi-variate and multivariate analysis for the numerical and categorical variable
5. **Outliers Handling**: Outliers detection for numerical variables and fix using capping upper and lower range.
6. **Data Transformation**: Created dummy variable for categorical, for binary categorical variable used binary_maping to change 'yes', 'no' with '1', '0' and removed all the redundant variables.
7. **Train Test Split:** Data is divided into two parts 70% for the training and 30% for test dataset
8. **Feature Rescaling:** StandardScaler is used to numerical variable feature rescaling. And correlation matrix is created to find the highly correlated variables.
9. **Model Building:** Following points are covered under model building.

- Used statsmodels.api for GLM model building
- There are total 72 features initially and used Recursive Feature Elimination (REF) and selected only top 15 features for model building.
- Train the model using the statsmodels.api as sm for GLM model and recursively removed the column using p-values to select the most significant features and which are contributing more on prediction.
- Used Variance Inflation Factor (VIF) to recursively removed the columns which are insignificant and having higher value of VIF score.
- So total 12 significant features are used to predicting the outcome
- The equation for ln(odds) is below
  ln(odds) = -1.2434 * const -0.9643 * Do Not Email + 0.7926 * Total Time Spent on Website + 2.1849 * Lead Origin_others + 2.2037 * Last Activity_sms sent -1.1646 * Specialization_travel and tourism -0.8361 * What matters most to you in choosing a course_not given + 6.9507 * 'Tags_closed by horizzon' + 6.4053 * Tags_lost to eins -3.7579* Tags_ringing -3.8393* Tags_switched off + 4.2868 * Tags_will revert after reading the email -1.6657* Last Notable Activity_modified

- The cutoff probability for the optimal point is .3, so chose the cutoff probability threshold value is .3
- Below the are details of matrix
  Training DataSet
    - confusion matrix = [[3812 190]
      [340 2126]]
    - accuracy = 91.8%
    - sensitivity = 90.9%
    - specificity = 91.8%

  Test DataSet
    - confusion matrix = [[1512, 165],
      [81, 1014]]
    - accuracy = 91%
    - sensitivity = 92.6%
    - specificity = 90%
    - recall = 92.6%
    - precision = 86%

10. Conclusion: From the model output we can conclude below points

- The Lead score for test dataset is 92% on the final model which indicates very high converting the lead into Hot leads.
- The accuracy of test dataset is very high i.e., 91% which makes this model very high to converge the leads into Hot leads accurately.
- The three variables in the model which contribute most is
  - Tags
  - Last Activity

    - Lead Origin
- 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 are below:
    - Tags_closed by horizzon
    - Tags_lost to eins
    - Tags_will revert after reading the email

11. Recommendations: Below are the points need to consider for the lead conversion.

- Focus more on the 'Total Time Spent on Website' and chances are more to get the lead can be converted as Hot leads
- Focus more on 'sms sent' for last noticeable activity the change is more to get more Hot leads converted.
- Tags (i.e. Tags_will revert after reading the email & Tags_closed by horizzon) is very important factors to get Hot lead converted
- Target the user with sending email as well as making the calls.