Lead Scoring Assignments

Summary Report

Goal of building this model:

- To build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads.
- A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Steps followed:

1. **Understanding the data**
    1. Total of 9230 rows and 37 columns
    2. There were missing values in the data
2. **Data Cleaning**
    1. Remove null values
    2. Dropping the columns where the missing values are greater than 40%. Used bar chart to find the missing values
3. **Exploratory data analysis**
    1. Check for duplicate values
    2. Findings from Univariate analysis and Bivariate analysis
        1. API and landing page submission has 35% conversion rate but count of lead originated were less
        2. Lead add form had more than 90% conversion
        3. Google and Direct tariff generates maximum number of leads
        4. Conversion rate from references and welingak website were high

5. Following columns were not had much to used
    1. Do not call
    2. Country
    3. Search
    4. Magazine
    5. X Education forums etc
3. Outlier analysis. Following column had outliers
    1. Total visits
    2. Total timespent on website
    3. Page view per visit
4. Based on Univariante analysis, dropped the columns which are not adding any information. After dropping I had 9074 rows and 14 columns

4. Data Preparation for Modeling

1. Converting to binary variables like for the column - Do not Email
2. Creating dummy variables for categorical columns:
    1. Lead Origin
    2. Lead Source
    3. Last activity
    4. Specialisation
    5. What is your current occupation
    6. City
    7. Last Notable Activity
3. Drop the columns for which dummies were created

5. **Modeling**

    1. Splitting the data into train and test data. Kept 70% for training and 30% for test data
    2. Next is to scale the feature
        1. Used StandardScaler: this is preprocessing technique used for standardising features in the dataset.
        2. It helps in features comparable and centred around Zero.
        3. Standard scaler calculates mean and Standard deviation of each feature and then subtracts the mean , then divides by Standard deviation.
        4. After the scaling, the lead conversion rate was 38%

    3. **Feature Selection using RFE**
        1. Keep building Model by adding constant
        2. Dropped the columns which had high P Value
        3. When the number of columns having less P value (less than 0.005), calculate the Variance Inflation factor (VIF)
        4. Repeated the mode till P value of all variables are zero and VIF has low value

    4. **Making Predictions**
        1. Created a data frame with actual converted flag and predicted probabilities
        2. Took 50% as probability cut off point
        3. Calculated the following coefficients
            1. Accuarcy
            2. Confusion matrix
            3. Sensitivity

    4. Specificity
    5. False Positive Rate
    6. Positive Predictive value
    7. Negative Predictive value
4. Plotted ROC Cuve
    1. This gave chance to find the optimal cutoff point
    2. Created columns with different probability cutoff
    3. Again calculated all the coefficients like accuracy, Sensitivity etc
    4. Arrived new cutoff of 0.35 as optimum point
5. Assigned lead score for training data
6. Calculated Precision and Recall
    1. Precision gave the percentage of result which are relevant
    2. Recall gave the percentage of total relevant result correctly classified by algorithm.
7. Next, Scale the test data and repeated the point (3) and (5)