Lead Scoring Assignment Summary

In this Assignment Objective was to Analyze the data and build the regression model to get the higher and lower chances of lead conversion and also given target Lead conversion rate around 80%.
We have build the model where we have analyze each features and extracted the higher and lower lead conversion client and able to meet the target of 80% conversion rate. We have used the given methods.

Step 1 – Imported all required Library, Read / Understand the data set.

We have Imported all the required libraries for analyzing, visualizing and model building. Read and and understand the data set. We have found total 9240 Rows and 37 Columns. Out that there was 6 variable which was 30% above null values.

Step 2 – Data Cleansing.

Select - In 4 Columns Client was not selected category hence considered this also as Nan and imputed values by creating new category or replaced with existing as per requirements.
Removed 30% above missing values columns also removed unused columns.
Imputed all Missing values which was Important features by creating new category or replaced with mean, mode methods.
Outlier – By plotting the Boxplot we have found outlier in 2 columns, We have treated both the 2 columns as upper outliers at u_bound and the lower outliers at l_bound while removing the top and bottom 1% of the both columns values
Total 12 important Columns and 9240 rows left after dropping all unused and missing values columns.

Step 3 – Exploratory Data Analysis.

Data Imbalance – Checked the Data Imbalance we have found 61.6 % Data Imbalance while plotting pie chart on conversion column.
Univariate Analysis – Plotted count plot, Boxplot and Pair plot each Numerical and Categorical Chart for Analysis and found the below:-

- **Do Not Email –** 68.7% client has asked for not to email wherein 31.3% client is interested to send email
- **Free Copy –** Approx. 8% client is interested to share Free copy
- **Occupation –** Compare to Student Unemployed
- **Last Activity –** Email Opened and SMS received % client are very high compare to Others
- **Lead Origin –** Landing Page Submission clients are high
- **Lead Source –** Most Numbers of client visited through Google and the very less from others source
  Specialization – Mostly client visited for Finance, HR, Marketing, however large customer has not selected any category.

Bivariate Analysis - Plotted chart for 2 features and as per analysis we have found the given below

- **Do Not Email –** Lead conversion for the Client who has asked for mail is highest conversion rate wherein not to email has very less conversion
- L**ead Source-**Mostly client wo has visited through online source such as google has highest conversion rate. Company should focus on advertising in online portal
- **Specialization –**Finance, HR and Marketing Specialization conversion is high chances of lead conversion hence we should focus on this area
- **Occupation –**Unemployed client has highest conversion rate where in Business occupation no conversion rate. Company should also target on Student and Working professional as working professional have very high conversion chances companion to others

Step 4 – Data Preparation for Modeling

- **Data Preparation for Modeling** - Created dummy variable for more than 2 category columns
- **Dropped Columns** - After Creating Dummy variable dropped the dummy variable columns
- **Train Test Split** - Splitted Data into Train and Test by takin 70 % and 30% size
- **Feature Scaling** - Scale the numeric variable using MinMax scaler methods

Step 5 – Model Building.

- **REF** - Used RFE method to get the top 15 important feature selection as the data was having more that 70 columns it was very difficult to choose important feature
- **Fitting Logistic Regression** - After added a constant into train data fitted a logistic regression model
- **Checked P Value** - Checked the P value in which features was having more that 0.05 P Value
- **Dropped Columns** - Dropped Features which was having > 0.05 P Value and build the model till all the features P Value converted to less then 0.05.

Step 6 – Model Evaluation & Prediction on the Test data.

- Calculated accuracy sensitivity and specificity for both plot and get the final cut-off as 0.35
- The area under the ROC curve is 0.89 which is indicating that we have a good model.
- Sensitivity of Test set is 81.00% and Train set is 80.17% using cutoff 0.35
- X Education company set target of lead conversion rate to be around 80% and the model is also giving 81.38% which is meeting the objectiv**e**

Final model is reflecting very close to each others. This indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.

-Sensitivity of Test set is 81.00% and Train set is 80.17% using cutoff 0.35

-X Education company set target of lead conversion rate to be around 80% and the model is also giving 81.38% which is meeting the objective.

****

Thanks & Regards

Mohd Sami Uzzaman / Agney Ravi O / Aditi Mishra