Lead Scoring Case Study Summary Report

The Lead Scoring Case Study is about an Education Company named X Education sells online courses to industry professionals and wanting to increase the lead conversion rate. In order to achieve that the company wants to build a model that assigns a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance. This basically sums up the Problem Statement by the company. The dataset for the Leads and the dictionary that contains the information on each data column are proved.

The main goal by the company is 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. The leads with higher scores are then targeted to improve the conversion rate.

The initial steps of importing and examining the data into a data-frame is performed with the help of python libraries. Then we perform Data Cleaning and Wrangling by handling the missing values, treating the outliers wither by dropping or imputation based on the data descriptive statistics and visualization.

We performed the Exploratory Data Analysis (EDA) on the cleaned data both Univariate and Bi-Variate and the correlation matrix. Several insights are drawn from the analysis such as conversion rate is high for customers who are working professionals or when the last activity after a phone conversation, SMS sent ,Lead Source for conversion rate is pretty high among the Reference and Wellingak Website. These insights are important as they contribute as actionable items on what is going well and then in turn contribute to the final model features.

We then moved to the data preparation to build the logistic regression model. We import the necessary packages from python libraries like sklearn, stats models for performing feature scaling, feature selection with RFE and looking at the model metrics. We scale the numerical features using the StandardScaler and create dummies on the categorical features. We split the dataset into Train and Test with 70:30 ratio.

We select the Training dataset and using Recursive Feature Elimination we cut down the 20 features and build our first model. Looking at the p-values significance we drop the features with higher significance are dropped and this step is repeated until we have all features p-value < 0.05.
The we calculated the Variance Inflation Factor (VIF) on the features to check if there is any multi collinearity. Once this is ok we proceed by selecting the final model to predict the target variable on the Training data. Then the Converted probabilities are stored in a final result data frame along with the actual target values. We randomly choose 0.5 as the cutoff and assign the predicted values on the probabilities calculated by the model.

We plotted the ROC cube to look if the Model is optimal and perform the probability calculation on several cutoffs ranging from 0-0.9 in steps of 0.1. The confusion matrix for every scenario is taken and plotted for three metrics Sensitivity, Specificity and Accuracy. The point where these three metrics converge is the ideal value for the model prediction on calculated probability values. It comes out as 0.3. Then we use this cut-off to calculate the final prediction of converted values. The Lead Scoring is obtained by just multiplying the calculated probabilities with 100. Then we proceed to check the precision and Recall trade-off. The final model that runs on the Training dataset has the Accuracy of 79.61% and a Sensitivity of 83.42%.

In the final step we proceed to test our model on the Test dataset. The model predictions are stored in a result data-frame along with the actual test values. The Lead Scores are assigned to each record based on the probability score calculated by the model. The Final Logistic Regression Model has the Model Accuracy (%) : 80.04 and Model Sensitivity (%) : 84.74 on the Test dataset. The F-score for the model is 0.76 which is > 0.7.

The KPI's from the logistic regression model are extracted and are documented for the company to derive actionable items that can contribute for improving the Conversion Rate.