Lead Case Study Summary Report

Main goal for this case study 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

First, we performed data preparation and analysis. To prepare data, operations like data loading, missing value handling, data cleansing were performed. It was followed by EDA. In EDA on the basis of data analysis we plotted various graphs e.g., bar plots, box plots, heat maps etc. and useful insights were drawn. Also, distribution of categorical variables, univariant & bi variant analysis and correlation between attributes were analyzed.

Once the attributes were finalized, the data was sliced between train and test data sets (70-30%). On the train data feature scaling were performed to standardize the independent features present in the data in a fixed range. We have used fit_transform() method so that we can scale the training data and also learn the scaling parameters of that data.

Logistic regression model is developed using Scikit-learn and Logistic Regression libraries. To get the best variables from the dataset we used RFE technique. Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. After performing RFE top 15 features were selected. Detected Insignificant columns based on VIF (variance inflation factor, how much the variance of an estimated regression coefficient increases if your predictors are correlated.) and P value (p-value is the probability of obtaining results). According to VIF and P value we dropped some columns. This not only improves the model performance significantly but also the overall model interpretability becomes easier.

After performing all the steps, we conclude that:

1. Analysis Outcome &Final Model shows that lead sources obtained from social media platform such as Welingak Website, Olark Chat positively affects the probability of lead conversion
2. It has been observed that Occupation is a very crucial criteria to increase the lead conversion. Analysis depicts that Working Professional is the best occupation that X Education shall focus
3. It has been observed that customers who spend appreciable time on X Education Website have high chances of conversion to Lead. These customers shall be treated on priority.

Post analysis if was observed that cutoff point came out to be 0.38 which essentially will take the specificity & sensitivity and thereby increasing the model efficiency and performance

Post taking this cutoff point below are the new metrics values-

1. Model Accuracy is 78.5%
2. Sensitivity is 74%
3. Specificity is 81%

We can observe here that accuracy is almost the same but sensitivity has increased to 74% from 65% which is great whereas specificity lies almost in the same range.