CREDIT CARD LEAD PREDICTION

Problem Statement

Happy Customer Bank is a mid-sized private bank that deals in all kinds of banking products, like Savings accounts, Current accounts, investment products, credit products, among other offerings.

The bank also cross-sells products to its existing customers and to do so they use different kinds of communication like tele-calling, e-mails, recommendations on net banking, mobile banking, etc.

In this case, the Happy Customer Bank wants to cross sell its credit cards to its existing customers. The bank has identified a set of customers that are eligible for taking these credit cards.

Objective

Based on the above problem statement task was to create a prediction model that will help the Happy Customer Bank to evaluate that if a credit card deal is provided to the customer then the customer will accept or reject the deal.

Approach:

To create a Prediction Model that will help the Happy Customer Bank to increase the cross sell of the credit cards, I have used the following approach:

- **Identify the problem statement: Thoroughly understood the data and the problem statement to use the required ML algorithm for prediction.**
- **Based on the problem statement and objective the ML algorithm best identified for the prediction of the credit card leads is supervised (Logistic Regression) Machine Learning Algorithm, because the target variable that we are targeting is categorical.**

Data Preprocessing:

For Data Preprocessing I have followed the following approach:

- **The dataset was given as train and test dataset. Both the datasets are imported, visualized, cleaned accordingly.**
- **Anatomy of each column of both the datasets is identified that whether the columns are Quantitative, Categorical or Qualitative.**
- **The columns which were qualitative were dropped from the datasets as they would create high dimensionality.**
- **Exploratory Data Analysis of both the datasets was done which helped to evaluate the outliers present in the datasets.**
- **Quantitative variables were identified using histogram**
- **Categorical Variables were identified using Bar Graph.**
- **Treatment of outliers was done on both the datasets based on the values which were very far away from the average value.**
- **Outliers treatment was not required in the Quantitative variables but it was done in the Categorical Variables, Values that very less in number were replaced by the another lowest category in the column to balance the data**
- **After the treatment of outliers, Missing Value Treatment was done in both the datasets, in which the missing values which were less than the 30% of the total number of rows were replaced by the average value of that column.**
- **Bivariate analysis was done on the Train dataset with the Target Variable(Is_Lead) to visualize that whether the predictors were correlated with the target variable or not.**
- **To visualize the Categorical vs Quantitative predictors Box Plots were plotted.**
- **To visualize the Categorical vs Categorical predictors Grouped Bar charts were plotted.**
- **The bivariate analysis of the target variable and predictors was done on the basis of the average value, that how much the average value of one category is different from the other category's average value.**

Model Development:

- **To create a final prediction model, those predictors were required that were correlated with the target variable.**
- **The correlation of the predictor variables with the target variables was statistically evaluated on the basis of the statistics tests.**
- **For categorical vs quantitative predictors ANOVA TEST was performed.**
- **For categorical vs categorical predictors CHISQ Test was performed.**
- **On the basis of the p-value predictors were considered to be correlated with the target variable.**
- **P-value was evaluated on the basis of the hypothesis test by accepting and rejecting the null hypothesis test, if the p-value was less than 5% than the null hypothesis was rejected and vice-versa.**
- **After selecting the final predictors, dummies of the categorical variables were created.**
- **The data was then splitted into70% train and 30% test datasets, in which the data was to be trained on the train dataset and the accuracy was to be calculated on the train dataset.**
- **Then the train dataset values were passed through the Logistic Regression Model and the model was created to learn the train data that has been passed through it.**
- **The predictons were made on the test dataset on the basis of the model has learned from the train dataset.**
- **The roc_auc_score for logistic regression was 0.72, it means that learning rate of the model is good to predict the predicton values as the original values.**
- **The accuracy of the logistic regression model is 77%**
- **Another ML model is used evaluate the learning accuracy .**
- **Second model I have used is Decision Tree Algorithm.**
- **Through Decision Tree Algorithm the roc_auc_score is 0.77 and accuracy of the model is 78%**
- **As the roc_auc_score and accuracy of the model is higher of the Decision Tree Algorithm, Decision tree is chosen as the final model for predicting Is_Lead values on the new dataset.**