Micro Credit Defaulter prediction:

Problem Description:

Here in this particular dataset we are required to predict whether the person in the dataset will be able to pay back the borrowed(loaned) amount within 5 days of insurance of the loan or not. Since the target here is to predict only yes or no(1 or 0) this here is a classification problem to predict the target(label).

First we are required to import all the necessary libraries from python to import , analyze and solve the problem at hand which in this case is classification(whether the person is defaulter or not) . After importing all the libraries, we are required to load the dataset into the notebook using pandas.

Exploratory data analysis (EDA):

After importing the dataset comes the analysis part. Detailed Exploratory data analysis must be performed so that we can understand the problem and choose the necessary steps required for model training. Here analysis was done about dataset

- No null values were observed
- Highly imbalanced data(label) had high ratio of ‘1’ as compared to ‘0’
- Msisdn was string , had no importance in model building whatsoever

- ‘pdate’ was string , had to be preprocessed into ‘int’ or ‘float’
- Highly skewed data was observed
- ‘1’ represents yes whereas ‘0’ represents no
- Outliers were observed
- ‘year’ which was created had only one value ‘2016’ when it was converted into integer type.
- ‘pcircle’ column had only one value ‘UPW’ was dropped
- Ratio of ‘0’ to ‘1’ was very high in every aspect
- High correlation between the independent variables were observed

Data preprocessing(Feature Engineering):

Feature engineering plays a vital role in proper functioning of a model. So it is absolutely necessary that the features must be treated precisely. From EDA we concluded many things and how we must act on it. Some of the features had no importance and had to be dropped. They were ‘unnamed:0’ , ‘msisdn’. ‘pdate’ was in string format so it was first converted into integer format with the help of ‘to_datetime’ into 3 columns and after that ‘pdate’ was dropped and ‘year’ was also dropped as it has only one value. Now data was split into x and y . ‘x’ as the independent variables and ‘y’ as the dependent variable(label). After splitting there were way to many columns to deal with some of them had high correlation between the dependent variables and some of the columns had unreal values and were treated with the help of PCA . Now the columns were intelligently reduced to 16 columns, dropping

them separately would have been very difficult. Now standardization of the data was required. After standardization now it was time for fitting the model into multiple algorithms.

Model Training:

Now as we have defined our dependent and independent variables or features as y and x respectively. ‘x’ is all the features excluding ‘label’ which is out dependent variable ‘y’. Now using train_test_split method we split the data into 70% for training and 30% with random state as 42. Hyperparameter tuning was done on KNNeighborsClassifier and RandomForestClassifier with RandomizedSearchCV and best_params_ was applied to give the best parameter for approach. The algorithm we applied gave pretty descent scores but the data was highly imbalanced so it gave very low precision, recall and f1 scores so we cannot use the particular model. Now it was time to address the problem of imbalanced data . So SMOTE and ADASYN function from imblearn and applied now we train the data in the same algorithms.

Report :

We trained the model and found the results for various algorithm . All the models performed really well except for Gaussian NB . But DecisionTreeClassifier gave the best result when ADASYN was applied to it. It is wise to save the model with the same parameters and conditions.