A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. MFS becomes very useful when targeting especially the unbanked poor families living in remote areas with not much sources of income. The Microfinance services (MFS) provided by MFI are Group Loans, Agricultural Loans, Individual Business Loans and so on.

Many microfinance institutions (MFI), experts and donors are supporting the idea of using mobile financial services (MFS) which they feel are more convenient and efficient, and cost saving, than the traditional high-touch model used since long for the purpose of delivering microfinance services. Though, the MFI industry is primarily focusing on low income families and are very useful in such areas, the implementation of MFS has been uneven with both significant challenges and successes.

Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

We are working with one such client that is in Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at Lower Prices to all value conscious customers through a strategy of disruptive innovation that focuses on the subscriber.

They understand the importance of communication and how it affects a person's life, thus, focusing on providing their services and products to low income families and poor customers that can help them in the need of hour.

They are collaborating with an MFI to provide micro-credit on mobile balances to be paid back in 5 days. The Consumer is believed to be defaulter if he deviates from the path of paying back the loaned amount within the time duration of 5 days. For the loan amount of 5 (in Indonesian Rupiah), payback amount should be 6 (in Indonesian Rupiah), while, for the loan amount of 10 (in Indonesian Rupiah), the payback amount should be 12 (in Indonesian Rupiah).

The sample data is provided to us from our client database. It is hereby given to you for this exercise. In order to improve the selection of customers for the credit, the client wants some predictions that could help them in further investment and improvement in selection of customers.

PROBLEM STATEMENT:

Build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of insurance of loan. In this case, Label '1' indicates that the loan has been payed i.e. Non-defaulter, while, Label '0' indicates that the loan has not been payed i.e. defaulter

EXPLORATIVE DATA ANALYSIS:

Data contains 36 features including the Label column

Data contains nearly two lakh records in it…so we are dealing with the larger data

- There are no null values in the dataset.
- There may be some customers with no loan history.
- The dataset is imbalanced. Label ‘1’ has approximately 87.5% records, while, label ‘0’ has approximately 12.5% records.
- For some features, there may be values which might not be realistic. You may have to observe them and treat them with a suitable explanation.
- You might come across outliers in some features which you need to handle as per your understanding. Keep in mind that data is expensive and we cannot lose more than 7-8% of the data.
-
- **SAMPLING TECHNIQUE**

Since the dataset contains label 1 with 87.5% records,we use undersampling technique.

This makes the lablels 50% each..so we that we don’t overfit the data..

For this techmique we import the package ‘imblearn..from this package, we use the Nearmiss() function.This makes the Label distributed equally in both classes.

Visualization

We use visualization techniques for finding the distribution of the data in different columns.

We plot different plots for univariate and multivariate analysis of the columns in the data.

Checking for outliers

There are lot of columns which have the outliers.we remove the outliers from the data as some machine learning models are very sensitive to the outliers.

We remove the outliers using the method called interquantile range method..

Encoding method

In this dataset we only have 3 columns which are of categorical data

Since machine learning models can understand the machine readable codes we convert them into the numerical variables..

For this process,we used the techniques called one hot encoding

So the data gets converted into the numerical variables.

At last we concatenate the numerical and categorical variables after encoding.

STANDARDIZATION METHOD

This method is used to standardization the data i.e.zero mean and unit variance..so all the data will be centred around the mean..

At last we use the data for training and testing

We use the train_test_split function

Modelling phase

In this we import 6 machine learning models

We train the data on this 6 models.The model which has the lowest difference between the accuracy_score and the cross_val_score

So in our case the Random forest model works better compared to rest of the models..

HYPERPARAMETER TUNING

We use the hyper parameter training technique to improve the accuracy of the better working model with its respective parameters..

Random forest will have the training parameters such as min_sample_split,n_estimators,min_sample_leaf,max_depth,learning rate..etc

We use the gridsearchcv function to train this parameters..Best parameters from the random forest model will be trained on the data…so we get the better accuracy

TESTING THE MODEL

We trained the data on the random forest model on the training data..so its time to test it with the test data so that we can conclude how our model working on the unseen data.

It seems our model is working the accuracy of 0.86 ..so our model is good to go.