Data Description:

Topic: To build a classification methodology to determine whether a person defaults the credit card payment for the next month.

What does default means?

Default means if the customer is unable to pay its credit card payment bill in the particular period of dates provided by the bank so that the customer is said to be as defaulter.

In simple term any customer not able to pay their dues on time considered as defaulters by the bank.

Why it is necessary?

There are several ways bank earns money. Mostly the bank earns money by interest charged by the bank towards the credit given to a customer.

So here with this problem statement bank is trying to get the forecast of what the inflow of the money bank will get from the customers.

Based on this clustering models bank will know which customers will pay their due and which customer will not pay their dues.

So accordingly, bank will decide its upcoming strategies based upon the inflow of the money from the customers.

In order to solve this problem, we will be given with the dataset by the bank.

To test this project, we have collected the data from the source:

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

Source:

Name: I-Cheng Yeh

email addresses: (1) icyeh '@' chu.edu.tw (2) 140910 '@' mail.tku.edu.tw

institutions: (1) Department of Information Management, Chung Hua University, Taiwan. (2) Department of Civil Engineering, Tamkang University, Taiwan.

other contact information: 886-2-26215656 ext. 3181

It consists of 30000 rows and 24 columns.

1. **LIMIT\_BAL**: continuous.Credit Limit of the person.
2. **SEX**: Categorical: 1 = male; 2 = female
3. **EDUCATION**: Categorical: 1 = graduate school; 2 = university; 3 = high school; 4 = others
4. **MARRIAGE**: 1 = married; 2 = single; 3 = others
5. **AGE**-num: continuous.
6. **PAY\_0 to PAY\_6**: History of past payment. We tracked the past monthly payment records (from April to September, 2005) The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
7. **BILL\_AMT1 to BILL\_AMT6**: Amount of bill statements.
8. **PAY\_AMT1 to PAY\_AMT6**: Amount of previous payments.

**Target Label:**

Whether a person shall default in the credit card payment or not.

1. default payment next month: Yes = 1, No = 0.

Apart from training files, we also require a "schema" file from the client, which contains all the relevant information about the training files such as:

Name of the files, Length of Date value in FileName, Length of Time value in FileName, Number of Columns, Name of the Columns, and their datatype.

Architecture:

The problem which we have taken we will divide it into sub smaller problems.

Like pre-processing, validation, modelling and then again, we will merge it into one file.

Various dataflow activities will take place:

A. DATA INGESTION: Various steps of data ingestion.

1. Data sourcing: We will be getting the data from the client in the batch folder or a network shared location.
2. Data Validation: We will read these files and perform certain validations of the data format, filename formats, this will be the number of columns and rows etc. So accordingly, we will put the data into good and bad data folder after the validation.
3. Data transformation: There are certain things that we will be doing before inserting it into a database.
4. Missing values replaced by NULL
5. Categorical values in single quotes should be replaced by double quotes

So that our internal database understands the data.

Why local database?

Client is placing data into shared file location.

Why again we are placing same kind of information into a database?

Client is going to place n different files so for n different files we are not going to create separate model for separate file.

We will aggregate whatever valid data client has sent we will validate all rows and we will aggregate them here in the database (common database) and using this good data we will do the model training we will export the data from the database to csv file and do the model training. This csv file will act as input for our model training.

B. DATA PREPROCESSING: Various steps in data pre- processing are :

1. Impute missing values.

2. Categorical values into numerical values

3. Imbalance dataset to balance data set.

4. Data not normal to normal dataset.

C. CLUSTERING:

The reason for the clustering is the performance of the models.

If the data is clustered and model is applied on each cluster then the sum of the performance of the clusters model is more as compared to one model applied on the complete data set.

D. DATA MODEL SELECTION AND TUNING

Now we will train individual models for individual clusters. And after that we will perform hyper parameter tuning. With this we will measure the performance of the models for individuals’ clusters by measuring the ROC or AUC score and finally we will get the model which perform best for the individual clusters.

And after that we will save that file into local system.

E.