**COMP5434 Group Project Report**

**Introduction**

As a virtual bank in Hong Kong, X-bank provides deposit and loans to local residents. In order to avoid possible financial risks, the bank’s auditor has to evaluate the repayment ability of a customer according to their basic information, including job, income level, balance, etc. Hence, a system is developed for predicting customers’ credit level. Furthermore, X-bank also prefer to get credit evaluation of customers without sharing raw data by establishing a credit federation with other banks.

In this project, a model is designed for the customers’ credit level prediction. A federated learning framework is also built for the model training and aggregation. The superiority of collaborative training would be shown over individual local training.

**Data Processing/Analytics**

The dataset of BankChurners.csv that contains basic information of 9000 bank’s customers and the target variable is the credit level between 1(bad) to 10(excellent) is described as follows:

● CustomerId – unique Ids for bank customer identification.

● Geography – the country from which the customer belongs.

● Tenure – number of years for which the customer has been with the bank.

● Balance – bank balance of the customer.

● NumOfProducts – number of bank products the customer is utilizing.

● HasCrCard – binary flag for whether the customer holds a credit card or not.

● IsActiveMember – binary flag for whether the customer is an active member or not.

● EstimatedSalary – estimated salary of the customer in Dollars.

● Exited – binary flag 1 if the customer closed an account with the bank and 0 if the customer is retained.

● CreditLevel – credit level of the customer

By creating the histogram and scatterplot, it’s difficult to summarize the relationship between each variable and the target variable. In this situation, we create new features by combining different variables together to test the correlation between new features and credit level of customers. Additionally, the value for the variable Balance that equals to zero influence the result for the testing data, then it is replaced by new values calculating with linear regression model.

**Model Design and Implementation**

Machine Learning

Model is designed by using XGBoost, which is efficient in gradient boosting decision tree. The training data and testing data are divided from the whole data for the machine learning and result prediction. By using XGB classifier, parameters including n\_estimators, max\_depth, and learning\_rate can be adjusted to get the best model.

Deep Learning

Build a neural network model to predict the credit level. We define our neural network by subclassing nn.Module, and initialize the neural network layers. Every subclass of nn.Module could operate on input data in the forward function.

**Framework of Federated Learning**

**Performance Evaluation and Discussions**

The test accuracy for our initial model is around 20%. In order to enhance the accuracy and improve the model, we have discussed the following methods:

1. By using the linear regression model, we get rid of the Balance value which is zero, and calculated the new value for the better model prediction.
2. Create new features according to different variables because of the weak correlation between the independent variable and the target variable-credit level.
3. Classify the target variable into three categories, which is 1-3, 5-8, 9-10 respectively for improving the model accuracy for the training data.

**Summary and Future Work**

**Reference**