**Machine Learning Analysis for Loan Default**

In that analysis, we will discuss how our project uses historical data to train and evaluate a model to show how banks collect loans, how they view people's loans, whether they default or do not default on the loans, and how several factors make people default on loans. That project looks at the customer's standard of living, whether male or female, income level, education level, and amount of loan they default on and the lessons we learned as a team from that project.

The data is divided into training and testing data. We chose a large dataset with about 25 columns and about 30,000 rows for our project. These columns show the borrower's loan size, sex, income level, number of opened accounts, marital status, education level, and total debt. They also have a label showing whether the loan defaults or not.

The Machine Learning Process has features: the size of the loan, sex, the borrower's income level, the number of opened accounts, marital status, education level, and total debt. Also, it has a label that shows the loan's status and whether people default on it. Using that data, we can do a first model and oversampling data. These two models can calculate the precision, accuracy, and recall level.

We used several packages within that machine learning project, such as Pandas, Matplotlib, SkLearn, random forest, and logistic regression models. We started by importing the data from CSV, changing the data type, normalizing the data, and resampling the data if required; this step is the most critical part. Then, we reviewed the data and started the cleaning process by dropping the first row of data. Our column included the ID, limit balance, sex, education, marriage age, bill amounts, payment amounts, and the amount of the loan that was defaulted for the coming month that the loan was due.

Furthermore, we tested different models for our machine learning. One of these models was the Logistic regression model, which tested about 500 decision trees. The accuracy for that model was approximately 71.88%. As for the precision, it reached 71.93%, and the recall score was 99.57%. When we used the random forest model, its accuracy level reached 73.26%. The precision level reached 75.96%, and the recall level reached 91.65%. When we compared both models, we found out that the random forest model has a higher accuracy level than the logistic regression model, it is the best classification model for machine learning. We tried SVC, but the accuracy score did not increase.

In addition, we used Grid Search CV and random search CV to optimize and improve the result of our random forest classifier and find the best performance for a given dataset. To achieve a higher accuracy level, we set parameters for the grid space regarding the maximum depth, estimators, max features, and minimum samples. After using the Grid Search CV and random search CV to find the best hyperparameter, especially for n\_estimator and max\_depth, each search takes around 90 minutes because the dataset is huge. The dataset has 630,000 entries. The best parameter we found was (n\_estimator = 162, max\_depth= 15), the accuracy level reached 74.43%.

We stored our clean data in an SQLite database within our machine learning. We could not use our initial data because our data at the beginning was not clean and was not up to the standards of SQL. Since we are using a large dataset, SQL can employ SQLite3 to analyze large datasets. By using SQLite3, we use one single file instead of worrying about reducing the number of queries. Within SQLite, we created a table with many columns and rows.

As a team, we learned that data is essential to any project. Data makes us make better decisions and analyze the data better. While conducting that project, we discovered the importance of data gathering, cleaning, and experimenting with different models to ensure the project's success. Machine Learning is more accurate than humans, and businesses use it to solve problems and predict better accuracy levels.