Group Project for IS 5740

Group name: TO BE DECIDED

Session: Monday

Dataset: Credit Card Defaults in Taiwan in 2005

**Project Proposal Draft**

**1. Business Context and Background:**

* The project focus the issue of **prediction of credit card defaults in Taiwan**.
* Credit card default prediction is a critical task for the banking industry, with direct implications for the financial stability of banks.

**Our Project aims to:** (PPT template slide 2)

* Assist financial institutions in achieving an **effective risk management** and **making informed lending decisions,** particularly for credit card product**,** Through **building a model that can predict credit card defaults accurately** of clients in Taiwan

**2. So Why is it important to have an accurate prediction on credit card default and How can it be applied for the business problems:** (PPT template slide 3)

**Effective Risk Management & Informed lending decision:**

Accurate credit card default prediction is essential for risk management in the banking industry. Since Defaulting clients can result in significant financial losses, due to the unpaid debts and administrative costs associated with collections.

If the banks are able to identify high-risk clients, they can reduce these potential cost by

**1. Early intervention**, such as providing timely reminders and offering financial counseling services.

**2. Take proactive measures**, such as lowering their credit limits or offering at a stricter term, in extreme cases, even suspending their credit cards. These measures help reduce the risk of default, mitigate the impact of defaults as well as lowering the bank's exposure to potential financial losses.

And for low-risk clients: offer them at a better commercial terms and better customer service.

**3. Data Description and Preliminary Analysis:** (PPT template slide 4-6)

**Data Description:**

So our dataset contains the demographic characteristics, and information about the usage of credit card for the clients in Taiwan from **April to September in 2005**. And the Target is **whether they were defaulted on their credit card payments in the next month that means October 2005**.

There are in total 30,000 data points and 24 attributes, including the target.

So let have a closer look of the independent variable, we have common demographics variable like Gender, where 1 refer to male and 2 refer to female, and also like Education level, where 1 mean Graduate, 2 mean university and so on.

Another major groups of independent variable is about the information of the usage of credit card of the clients, it includes the bill statement, client’s payment history and their repayment status for April to September in 2005 respectively.

And Eventually for our TARGET, our dependent variable, which indicates the default status of the client in October 2005, where 1 mean Default, they didn’t fully settled their bill in October, and 0 mean they didn’t Default.

**Preliminary Descriptive Analysis on independent variable:**

In the preliminary descriptive analysis, we calculated summary statistics, including mean, median, standard deviation for categorial variable and some of the numerical variables. And We visualized data distributions using histograms, bar charts for categorical variable

By just focusing the mean value, it can be seen that there is more women than men in the dataset.

The average education level is mostly centered in university and then graduate

The average age is 35

The average credit limit is around 160,000

**Preliminary Analysis on Target variable**

* Total number of Not Default: 23364
* Total number of Default: 6636
* Average default rate: **22.12%**

By dividing the total number of observed Default over the total number of observation, we have got the average default rate of 22.12%

**3. Identified Problems, Questions, and Approach:** (For PPT last slide content)

**Problems and Questions:**

We identified the following problems and questions for further investigation:

* Problem 1: **How to build a model that can predict client defaults on credit card accurately?**
* Problem 2: **How do various attributes impact the likelihood of default?**

**Approach:**

To address these problems and questions, we will begin with:

* Explore the data further, Preprocess the data including addressing missing values, outliers, and encoding categorical variables.
* Create and train various machine learning models including **logistic regression**, **ensembled decision trees (e.g., Random Forest)**, and **neural networks**. To analyse our data (Why? Since these kinds of model is suitable for binary classification problem)
* Evaluate model performance using accuracy, the baseline of our model accuracy rate is 78%, how do we get this number, recall back the average default rate of this dataset is around 22%, imagine there is a stupid AI who learn nothing but only predicting the TARGET of every client with the most common category, in our case which is predicting that “the client will Not Default”, and it is not surprising that it can probably obtain a pretty good accuracy rate of around 78% in the validation and testing dataset. Therefore, due to the presence of a slightly unbalanced dataset issue, the accuracy rate of our model would have to be higher than 78%, and other evaluation metrics such as Precision, **recall** and F1-score will also be used to evaluate our model performance, and we will keep on evaluating it through fine-tuning the model parameters and choice of attributes used to train the model
* **To address problem 2, we would Based on the above prediction model built, and analyze the feature importance** in order to understand the factors influencing the TARGET, the default

**Expected Solutions:**

* Develop accurate predictive models for credit card default prediction, and
* Provide banks with tools for risk management and improved lending decisions.

This is the end of the proposal presentation, we will share more detailed finding and result in the final presentation, and thank you so much for your patient listening.