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:

Business Context:

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

Our Project aims: (PPT template slide 2)

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

Key Question:
The central question we aim to answer in this project is:

- **"Can we accurately predict credit card defaults for clients in Taiwan?"**

Why is it important to solve the problem: (PPT template slide 3)

Risk Management: Accurate credit card default prediction is essential for risk management in the banking industry. It enables banks to identify high-risk clients and take proactive measures, such as lowering credit limits or, in extreme cases, suspending credit cards. These measures help mitigate the impact of defaults and reduce the bank's exposure to potential financial losses.

Cost Reduction: Defaulting clients can result in significant financial losses due to unpaid debts and administrative costs associated with collections. Accurate prediction helps reduce these costs by enabling early intervention, such as providing timely reminders or offering financial counseling services, which can help reducing the risk of default which in terms reducing the financial losses.

Lending Decisions & Resource Allocation: Accurate default prediction empowers banks to make more informed lending decisions. It helps allocate resources more efficiently:

- For high-risk clients, banks can tailor lending terms to minimize the risk of default. This may include offering lower credit limits and more stringent terms.
- For low-risk clients, banks can provide better terms and services, creating a mutually beneficial relationship.

<u>2. Data Description and Preliminary Analysis: </u>(PPT template slide 4-6)

Data Description:

This dataset contains information on demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from **April 2005 to September 2005**. and **whether they defaulted on their credit card payments on next month – Oct 2005 (TARGET).** There are in total 30,000 data points and 24 attributes.

```
RangeIndex: 30000 entries, 0 to 29999
Data columns (total 25 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   ID                          30000 non-null  int64
 1   LIMIT_BAL                   30000 non-null  float64
 2   SEX                         30000 non-null  int64
 3   EDUCATION                   30000 non-null  int64
 4   MARRIAGE                    30000 non-null  int64
 5   AGE                         30000 non-null  int64
 6   PAY_0                       30000 non-null  int64
 7   PAY_2                       30000 non-null  int64
 8   PAY_3                       30000 non-null  int64
 9   PAY_4                       30000 non-null  int64
 10  PAY_5                       30000 non-null  int64
 11  PAY_6                       30000 non-null  int64
 12  BILL_AMT1                   30000 non-null  float64
 13  BILL_AMT2                   30000 non-null  float64
 14  BILL_AMT3                   30000 non-null  float64
 15  BILL_AMT4                   30000 non-null  float64
 16  BILL_AMT5                   30000 non-null  float64
 17  BILL_AMT6                   30000 non-null  float64
 18  PAY_AMT1                    30000 non-null  float64
 19  PAY_AMT2                    30000 non-null  float64
 20  PAY_AMT3                    30000 non-null  float64
 21  PAY_AMT4                    30000 non-null  float64
 22  PAY_AMT5                    30000 non-null  float64
 23  PAY_AMT6                    30000 non-null  float64
 24  default.payment.next.month  30000 non-null  int64
```

- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **AGE**: Age in years
- **PAY_0**: Repayment status in September, 2005 (-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)
- **PAY_2**: Repayment status in August, 2005 (scale same as above)
- **PAY_3**: Repayment status in July, 2005 (scale same as above)
- **PAY_4**: Repayment status in June, 2005 (scale same as above)
- **PAY_5**: Repayment status in May, 2005 (scale same as above)
- **PAY_6**: Repayment status in April, 2005 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2005 (NT dollar)

- **Independent variable**: key variables such as 'LIMIT_BAL' (credit limit), 'SEX' (gender), 'EDUCATION' (education level), 'MARRIAGE' (marital status), 'AGE' (age of the client), 'PAY_X' (payment status for various months)
- **target variable/dependent variable**: 'default.payment.next.month' (the target variable indicating default or non-default).

Summary and Preliminary Descriptive Analysis:

In the preliminary descriptive analysis, we observed the following key points:

- We calculated summary statistics, including mean, median, standard deviation, and quartiles for categorial variable and numerical variables such as 'LIMIT_BAL' and 'AGE.'

```
# Categorical variables description
df[['SEX', 'EDUCATION', 'MARRIAGE']].describe()

# The mean of Education Lv is 1.85, meaning mostly g
# Most of the clients are either married or single
# More women than men
```

	SEX	EDUCATION	MARRIAGE
count	30000.000000	30000.000000	30000.000000
mean	1.603733	1.853133	1.551867
std	0.489129	0.790349	0.521970
min	1.000000	0.000000	0.000000
25%	1.000000	1.000000	1.000000
50%	2.000000	2.000000	2.000000
75%	2.000000	2.000000	2.000000
max	2.000000	6.000000	3.000000

```
df.LIMIT_BAL.describe()
# df.LIMIT_BAL.mean()

# average value for amount of c
```

```
count      30000.000000
mean      167484.322667
std       129747.661567
min        10000.000000
25%        50000.000000
50%       140000.000000
75%       240000.000000
max      1000000.000000
```

```
df["AGE"].describe()

# average age is 35.5 years
```

```
count    30000.000000
mean        35.485500
std          9.217904
min         21.000000
25%         28.000000
50%         34.000000
75%         41.000000
max         79.000000
```

- Further Analysis with TARGET to have a slightly deeper understanding. (example)

TARGET	Not default	Default	Default Percentage
SEX			
Male	9015	2873	0.241672
Female	14349	3763	0.207763

TARGET	Not default	Default	Default Percentage
EDUCATION			
Graduate	8549.0	2036.0	0.192348
University	10700.0	3330.0	0.237349
High Sch	3680.0	1237.0	0.251576
Other	116.0	7.0	0.056911

SEX	MARRIAGE	Not Default	Default	Default Percentage
1	0	12	2	0.142857
	1	3844	1346	0.259345
	2	5068	1485	0.226614
	3	91	40	0.305344
2	0	37	3	0.075000
	1	6609	1860	0.219625
	2	7555	1856	0.197216
	3	148	44	0.229167

- Married men have a higher probability of default,
- Single men and Married Women's default is similar to the a
- Single women have a lower probability of default

TARGET	0	1	Default Percentage
LIMIT_BAL			
10000.0	296.0	197.0	0.399594
16000.0	2.0	NaN	NaN
20000.0	1278.0	698.0	0.353239
30000.0	1042.0	568.0	0.352795
40000.0	138.0	92.0	0.400000
50000.0	2480.0	885.0	0.263001
60000.0	592.0	233.0	0.282424
70000.0	521.0	210.0	0.287278
80000.0	1204.0	363.0	0.231653
90000.0	485.0	166.0	0.254992

- We visualized data distributions using histograms, box plots, and bar charts for categorical variables.

Analysis on Target

- Total number of Not Default: 23364
- Total number of Default: 6636
- Average default rate: 23364/ (23364 + 6636) = **22.12%**

- 1-22.12% = 78% would be our baseline in evaluating model performance when evaluating in terms of model accuracy rate, ofc since it is an unbalanced dataset (22.12% vs 78%), other metrics i.e., Recall will be used as well

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: **Credit card defaults can lead to significant financial losses for banks. How can we predict defaults accurately to minimize these losses?**

- Problem 2: **How do demographic factors (e.g., gender, education, age and marital status) and credit card limit and payment history (PAY_X) impact the likelihood of default?**

Approach:
To address these problems and questions, we plan to:

- Explore the data, Preprocess the data, addressing missing values, outliers, and encoding categorical variables.
- Create and train predictive models using **logistic regression**, **ensembled decision trees (e.g., Random Forest)**, and **neural networks**. (Why? Since these kinds of model is suitable for binary classification problem) → Problem 1
- Evaluate model performance using metrics such as accuracy, precision, **recall**, F1-score, and AUC-ROC. → Problem 1
- **Analyze feature importance** to understand the factors influencing default. → Problem 2

Expected Solutions:
From this business analytics project, we expect to:

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

Our final stage: we will deploy our model and have an interface to input data for a client and the model will output the probability of a client's default on credit card.