

# Credit Worthiness Classification for Taiwanese Banks



# Table of Contents

Introduction: .....	2
Background: .....	2
Research Aim.....	4
Research Questions .....	4
Financial Industry Overview.....	4
Money and credit markets: .....	5
Investments:.....	5
Financial management:.....	6
Literature Review .....	7
Data Description.....	10
Data Collection Process .....	11
Data collection activity: .....	12
Data Analysis.....	13
Data Understanding and Preparation .....	13
Multicollinearity .....	15
Predictive Model .....	16
Interpretation: .....	18
Recommendation: .....	19
References .....	20
Annex.....	21

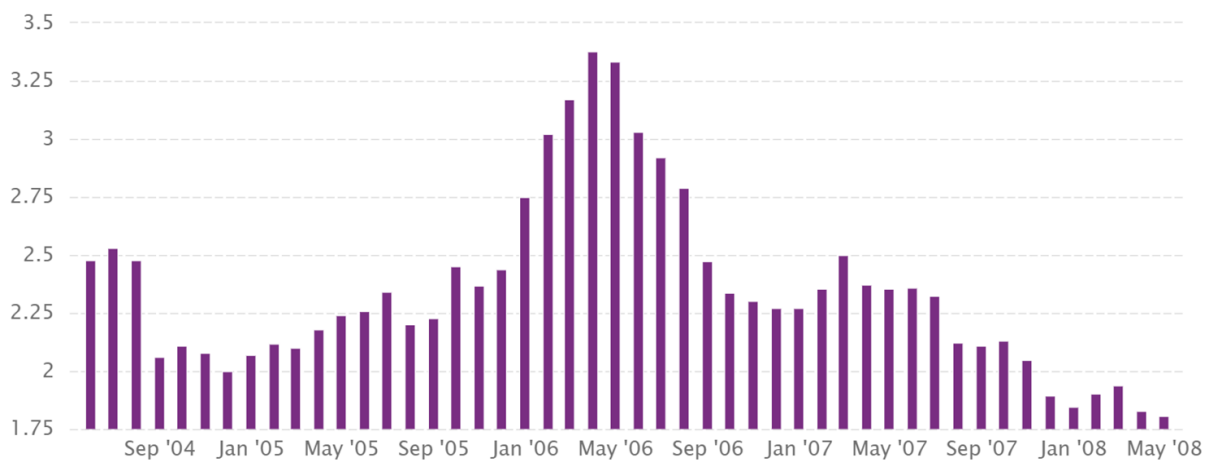
## **Introduction:**

The credit card industry of the banking domain has always been a major concern for the banks in terms of identifying legitimate customers. There is a strong need for risk prediction especially in the financial industry to help manage uncertainty. Banking operations are something that we all come across in our daily lives. In the recent years the use of credit card has become very popular as it is one of the most convenient payment options for everyone. However, this convenience does come with its own risk for the banks. As the number of customers using credit cards increase, more efforts need to be taken to consider managing the risk involved in terms of delinquency. The overall objective of our project is to utilize the past behavioral information of the customers- financial, demographic, personal information, and understand the patterns to make sound decisions for optimizing their profit i.e., to find out whether the customer is good or bad and that the bank can lend them money or not.

## **Background:**

Banks in Taiwan spent money on advertisements urging individuals to apply for credit cards to consume, seemingly without penalties, to increase their market share. In order to attract more clients, some banks eased the standards for credit card approvals. Taiwanese consumers took to credit cards like the proverbial duck to water, as they did not have a legacy cheque culture to relinquish. (Wang) In 2005, the consumer financial crisis broke out in Taiwan, sometimes known as the double-card crisis since the issue was mostly caused by two types of credit card debt (credit cards and cash cards). The issue included the following aspects: first, as of September 2005, there

were 133 cards for every 100 Taiwanese over the age of 15, and each cardholder normally carried more than four cards in his or her wallet. The lack of a risk management infrastructure by Taiwanese issuers, however, led to a consumer debt crisis by the middle of 2005. By the first half of 2006, the media was having a field day covering the 520,000 problematic borrowers (out of a population of 23 million), whose bad debts averaged \$9,300 per individual. The inappropriate collection of payments results in social crimes like theft and robbery and even drives some debtors to commit suicide. Secondly, financial institutions' late and default rates rapidly increased, and some financial institutions almost went out of business because of significant losses. Thirdly, the debt crisis caused a slowdown in consumer expenditure and economic expansion. (Sunshine Research Center for Financial Innovation)



## **Research Aim**

The main goal of this project is to analyse the past data and classifying the credit card holders' profile as good or bad based on factors such as Age, income, annual expense estimates, previous loan repayment behaviour, existing loans and payment schedule, net worth using Machine Learning model.

## **Research Questions**

1. What information is the best criteria to determine credit worthiness?
2. What are the strategies for risk management?
3. How can banks best manage bad credit?

## **Financial Industry Overview**

Financial industry is one of the major sectors of business in a country that plays a pivotal role in the development of the country. It consists of the firms and organizations that provide financial services to the commercial and retail client. The primary focus or business of the companies falling under this industry is to generate commissions by offering services like banking, personal loans, insurance, mutual funds, stock market, mortgages and many other financial services. It encompasses a broad range of businesses that manage money, including credit unions, banks, credit cards, collateral based loans, insurance companies, accountancy companies, consumer-finance companies, stock brokerages, investment funds, individual asset managers, and some

government-sponsored enterprises. In other words, this industry is the funding window for any industry.

The banks providing financial services have various categories under it like the commercial banking services, investment banking services, foreign exchange, insurance etc. The economy of a country totally relies on how well this industry is performing. The financial industry consists of three interlinked areas. They are discussed below:

- Money and credit markets
- Investments
- Financial management

#### **Money and credit markets:**

The money and credit markets deal with the security markets and financial institutions. They are the central to the allocation of capital, the efficient distribution of liquidity and the hedging of short term risks.











#### **Investments:**

The investments area usually focusses on the approvals or decisions taken by the two sides – individuals and the institutional investors.

## Financial management:

The financial management area is generally to oversee the financial health of an organisation and supervise ideal functions like determining profitability, managing expenses etc. It involves decisions made within the organization regarding the use of funds.

The top 10 performers of financial industry are mostly from China and the United States. A statistical table of the same is below:

Rank ^	Company ⇅	Industry ⇅	Revenue (USD millions) ⇅	Net Income (USD millions) ⇅	Total Assets (USD billions) ⇅	Headquarters ⇅
1	<a href="#">Transamerica Corporation</a>	Conglomerate	245,510	42,521	873	 <a href="#">United States</a>
2	<a href="#">Ping An Insurance Group</a>	Insurance	191,509	20,738	1,460	 <a href="#">China</a>
3	<a href="#">ICBC</a>	Banking	182,794	45,783	5,110	 <a href="#">China</a>
4	<a href="#">China Construction Bank</a>	Banking	172,000	39,282	4,311	 <a href="#">China</a>
5	<a href="#">Agricultural Bank of China</a>	Banking	153,884	31,293	4,169	 <a href="#">China</a>
6	<a href="#">China Life Insurance</a>	Insurance	144,589	4,648	776	 <a href="#">China</a>
7	<a href="#">Allianz</a>	Insurance	136,173	7,756	1,297	 <a href="#">Germany</a>
8	<a href="#">Bank of China</a>	Banking	134,045	27,952	3,739	 <a href="#">China</a>
9	<a href="#">JP Morgan Chase</a>	Banking	129,503	29,131	3,386	 <a href="#">United States</a>
10	<a href="#">AXA</a>	Insurance	128,011	3,605	984	 <a href="#">France</a>

U.S based Transamerica Corporation from the Conglomerate department of the financial industry tops the table with a Net income of 42,521 US million dollars followed by China based Ping An Insurance Group with Net income of 20,738 US million dollars, and many more companies.

## **Literature Review**

Credit scoring techniques date back to the 1940s and has since then evolved into much more sophisticated methods which uses data analytics to determine credit worthiness. With the introduction of credit cards in the 1960s, banking institutions saw the need of determining credit worthiness of their clients. A typical credit scoring model would input a number of measurable attributes of a potential lender to determine the probability that the loan will be paid back as. (Yap et al.) A credit rating system in principle measures the likelihood that a debtor may not meet the requirements under the debt/loan contract. A low score means higher risk for the lender. (Chen and Cheng) Many studies and research has been conducted in the field of client credit worthiness since its inception and different ideas and models have been introduced to measure credit worthiness. Most of the research claims that there is no best model for all types of classifications but rather depends on its purpose (Yap et al.).

Previous studies and research conducted on the subject stress on the importance of data quality when it comes to predicting credit worthiness. Two of the biggest limitations are unavailability of accurate data due to errors in recording, missing entries, etc. and in determining a sample (Yap et al.). A study done on the use of machine learning to improve noise related problems in credit risk analysis uses a three step strategy. First it measures the noise level attributes of the data. Then it sets a threshold to identify high and medium noise levels in attributes and high noise attributes are deleted while medium ones are de-noised using machine learning algorithms. The aim of the step is to preserve as much attributes as possible to that the prediction will be more data based. Next



and last, the CART method is used to classify the credit worthiness of the consumer. This method is targeted to improve accuracy and reduce error statistics. (Yu et al.) It was noted that most of the studies conducted in this field used sample datasets without any missing values, like in the case of determining credit worthiness of farmers who applied for loans in a Chinese commercial bank. It used a data subset of containing information of 2044 farmers without any missing data. (Bai et al.)

Another study drew attention to a social matter which may take place as a result of such classifications such as credit worthiness. It suggested that some characteristics such religion, sex and race are illegal to be used in some countries for credit scoring models although they may have an considerable indirect impact on the likeliness of loan repayment by individuals. The same study further highlighted that through this modelling, society will be categorized into two main segments: those who can borrow from all and those who cannot borrow from any. This may result in unfavorable conditions for the latter. However, it also highlights that accounting for economic factors in the prediction model has proven to extend the lifetime of the model itself. (Mwesigwa et al.) This goes on to show that changing economic factors carry a considerable weight in determining credit worthiness. Another study confirmed the above findings where it showed that even though gender and ethnicity are not considered to a great extent, they do carry a potentially significant say in both application and outcome of loans. The same study, done in Trinidad and Tobago goes on to say that the loan application by females is comparatively much lesser to males. (Storey)

It was seen from the literature review that most of the studies and models used predicted credit worthiness using a scoring system where an individual or a company was given a certain score depending on the nature of the criteria that was input. This approach could seem rather subjective when it comes to determining whether the customer is worth the credit. The purpose of this study is to do away with the concept of scoring and introduce a direct binary solution to state whether the customer is worthy or not. The worth of the customer will depend on factor such as gender, age, marital status, outstanding credit, nature of past payments, etc. The model is expected to consider many factors among these in order to derive at the final decision whether the customer is eligible for the loan or not.

## Data Description

This dataset consists of a total of 23 fields and 30,000 observations. By analyzing and performing data audit, we observed that all the 23 fields can be used to establish pattern and then classify a new customer with the similar profile as a defaulter or not.

The data shared by the business consists of information like: -

1. Current outstanding credit of the applicant and immediate family.
2. Gender.
3. Education.
4. Marital status.
5. Age.
6. Past payments YTD
7. Past usage trends of last 5 months.
8. Past payments of last 5 months.
9. **Customer defaulter Report.**

Please refer Annex – Table 01 for data description.

The chosen dataset is cleaned with minimal missing values and already converted to number format as per the table below. This is ready to be fed into a machine learning model as an input. This helps us understand how people of different backgrounds manage credit and can help us do a binary classification of a 'Good' and 'Bad' credit for customer. The goal of the project is to only do a binary classification and not rank or score the customers.

We will be using all 23 attributes because for our initial analysis we think that all the factors will be contributing towards the predictor variable i.e., whether a customer will be defaulter or not. Using this historical data, we will use machine learning algorithms to train the model and then classify a new record as defaulter or not.

Link to Dataset: - <https://www.kaggle.com/datasets/srijithl/taiwan-credit>

## **Data Collection Process**

Data regarding this may include qualitative data and a little data in other formats, however we did not consider collecting those data but focused on numbers to generate values and to use for analysis. So, the nature of the data will be structured.

The data is collected across different platforms and may not be cleaned. We experienced few problems with the quality of data like the Errors, inconsistencies, and few other difficulties that are common with raw data. Usually, the data collecting methods would be intended to avoid to reduce such issues. However, in many cases, this is not fool proof. Hence, the collected data

typically requires data profiling and data cleansing to discover flaws and resolve them. Hence, the need to merge, sort, filter etc. may be needed on the data.

Gathering valuable data: With so many systems to navigate, acquiring data for analysis can be a difficult task for data scientists and other related users inside a company. The application of data curation strategies aids in the discovery and accessibility of data. This could include developing a data catalogue and searchable indexes.

### **Data collection activity:**

We searched available datasets on numerous websites, however we were only able to find data that was more broadly applicable than data that was specifically relevant to our research topic. Because Kaggle provided us with more relevant data than other sources, we decided to use it to collect and finalise our dataset. Additionally, the data descriptions were clear, intelligible, and pertinent to our research. A dataset that was better suited to our study questions was discovered. Our database is based on information acquired from 30000 clients by a credit card company [1]. Based on a few factors, we want to determine whether the consumer will end up being a credit payment defaulter. The information gathered has been structured and is available in the right format for analysis using machine learning models.

## **Data Analysis**

Before diving in to the analysis section, it is prudent to recapture the project aims and research questions. The main goal set out at the beginning of the project was to analyze past data of a Taiwanese bank and use it to predict a classification of a credit card holder to be either creditworthy or not so that the model can be applied to future customers. For this purpose a dataset of 23 fields and 30,000 observations were selected as stated in the data description section of the project. This data set was used to analyze patterns and classify the default possibility of a customer.

The research questions were to determine the variables that mostly affected possibility of defaulting, discuss risk management strategies and provide recommendations for Taiwanese banks on determining credit worthiness.

Data analysis for the project was performed using Python programming language. Both descriptive and predictive approaches were used during the analysis process. Descriptive was used to analyze what had already happened and a predictive model was used to determine expected outcome of default possibility.

## **Data Understanding and Preparation**

A key part of data analysis is data preparation. Most datasets are not analysis ready in their raw formats. There may be missing values, improper data types, mistakes in the data, etc. An initial analysis of the raw data indicated that all variables were of data type, 'object'. This data type needed to be changed to numerical format so that any relations between variables could be analyzed. The dataset did not have any missing values and therefore there was no requirement for

any fixes. By comparing the raw data with the description of variables, it appeared that all values were in fact relevant and within the descriptions provided. For example, there were no values other than 1 and 2 for male and female respectively and for variable Age, there were no surprisingly high or low values (ranged from 21 to 79).

The dataset had clearly defined independent and dependent variables. Independent variables were identified to be fields such as limit balance, age, sex, marital status, payment history and billed history. “Default payment next month” variable was identified as the dependent variable in order to predict the credit worthiness outcome using the predictive model. Different approaches were made in order to prepare the dataset for analysis. Given below are some of the steps taken in Python to prepare the dataset:

1. Data type of all variables by default were ‘object’. For analysis purpose, the datatypes were converted to ‘float’.
2. The data set presented billed amounts and paid amounts across a period of 6 months.
3. In order to reduce dimensionality, 2 new columns titled ‘Total\_Credit\_Used’ and ‘Total\_Paid\_Amounts’ were created to sum up the information spread across 6 months.
4. A third new column was added to the dataset to determine the total percentage of credit paid during the 6 month period.
5. Columns pertaining to individual months and the totals were subsequently dropped, reducing the total number of variables to 14.

## Multicollinearity

Due to the presence of many independent variables, it seemed apt to conduct multi-collinearity to determine the relations between independent variables. The below output indicated multicollinearity:

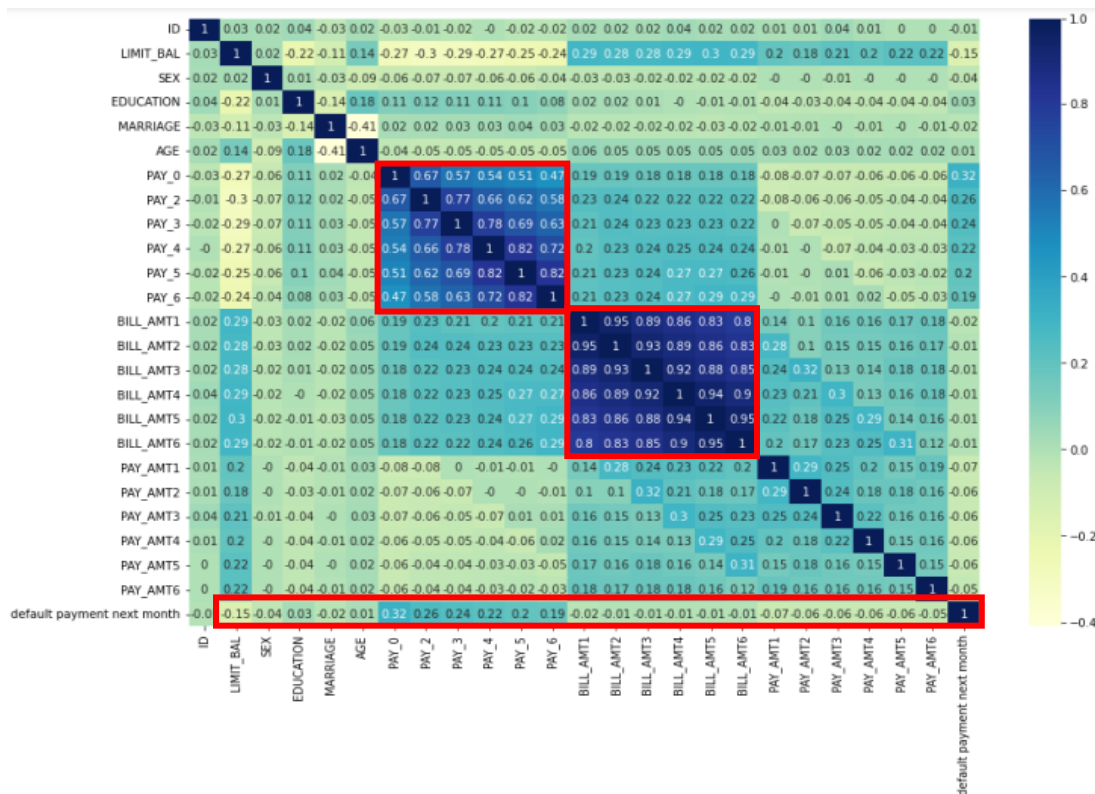


Chart 01: Multicollinearity Analysis

Multi-collinearity analysis indicated a high positive correlation between paid amounts and bill amounts. It was interesting to note that all variables pertaining to billed and paid amounts had more than 50% positive correlation while some even went as high as 95%. This correlation was expected as naturally when the billed amounts are high, customers tend to settle them off by paying more. There were no other significant correlations noticed among the independent variables.



Another key observation was in the bottom most line where the correlation between independent variables and the dependent variable were measured. It showed that none of the independent variables, individually, had any strong direct or indirect relationship with the dependent variable. This goes on to show that any factor alone does not have an impact on whether a given customer will default payment next month or not. This is further more reason as to why a predictive model that captures all variables is required for a predictive classification.

### **Predictive Model**

The set-up of the data with multiple independent variables and one clearly defined dependent variable demanded the use of multi-linear regression as the base for the predictive model. The ultimate aim of the model was to predict the likelihood that a given customer will default will default in the next month payment. Multi-linear regression model is as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \cdots + b_{23}X_{23}$$

Y = Credit Worthiness

$b_0$  = Value of Y when all independent variables are zero

$X_1$  = Amount of given credit

$X_2$  = Gender

$X_3$  = Education

$X_4$  = Marital Status

$X_5$  = Age

$X_6 - X_{11}$  = History of past payments (April to September 2005)

$X_{12} - X_{17}$  = Amounts of bill statement (April to September 2005)

$X_{18} - X_{23}$  = Amounts of previous payments (April to September 2005)

$b_1 - b_{23}$  = Estimated regression coefficients

In order to achieve a realistic prediction model, the given dataset was divided in to two sets, namely, the training set and the test set. The division was done on random basis with 75% of entries allocated to the training set while the remaining 25% was allocated to the test set. Subsequently, predicted results of the test set were compared with the actual results to determine the accuracy of the model and it's practicality in real life.

	<b>Training Set (75%)</b>	<b>Test Set (25%)</b>
No of entries	22,500	7,500
Accuracy	0.999	0.815
Match %	100%	81.5%
Matched number of entries	22,500	6113

The results of the analysis indicated that the prediction model was accurate almost 82% of the time. As per industry standards, an accuracy level of 80% is accepted. Hence we conclude that the model is in fact usable in real life. It must be mentioned that additional information has the possibility to improve accuracy of the model.

## **Interpretation:**

The main aim of the project was to find a solution to predict the credit worthiness of the customers who are applying for a credit card in a funding company. Using our algorithm, the companies will easily be able to find if the customer is eligible for a credit card or not. Using the default payment next month as the dependent variable, the credit worthiness of the customer is predicted. The steps involved data visualization, data preparation etc.

To determine the relationship between independent variables, the multi-collinearity was performed while multi-linear regression was performed for predictive model as there were more than one independent variable. The multi-collinearity indicated a high positive correlation between paid and bill amounts.

Test set had 7500 records of which, the training set and test set had a random selection of data entries at 75% and 25% respectively. There is a decent match between the data however, other details like earlier loans taken, and customer annual income etc. is helpful. Random Forest classifier was performed which determined the accuracy of 99.95% while the accuracy of test data is 81.5%.

As the variables in the dataset predict the possibility of defaulting upcoming payments, it is beneficial for banks and funding organizations to decide better in terms of categorizing who should be given loans.

## **Recommendation:**

Based on the predictive model, the variables included in the dataset plays an important role in predicting the possibility of a defaulted payment next month, and therefore determining a client's credit worthiness. The banks will be able to make more informed decisions about who to lend to because of this data. Even though the model indicated a decent match between the training data and the test data, it is believed that the model could be improved further if more information (variables) was available, such as the annual income of the borrower, the number of other loans he or she has taken (or a method to calculate the Debt-to-Income Ratio (DTI), savings, investments, collateral, etc.

Data-warehousing is another way to move forward where information pertaining to clients are organized and stored to be retrieved in a timely manner. A more interconnected set of systems and processes will enable predictive analysis to be conducted with supplementary related information.

According to this outcome objectively from the perspective of the customer, financial institutions (in Taiwan) need to educate their customers on how to improve their credit worthiness and handle their loans better. As a result, both parties would benefit from this arrangement. It is imperative for banks to increase the income and job requirements of applicants in the future, as well as to prohibit improper credit card commercials so that customers do not get a bad impression of credit cards.

## References

1. Wang, Eric. "The Taiwan Credit Card Crisis - Financial Ethics." Seven Pillars Institute, Seven Pillars Institute, <https://sevenpillarsinstitute.org/case-studies/taiwans-credit-card-crisis/>.
2. Sunshine Research Center For Financial Innovation. "Research on Taiwan Consumer Financial Crisis and Its Management." Institute for Fintech Research Tsinghua University, Tsinghua PBCSF, 23 Apr. 2021, <https://thuifr.pbcsf.tsinghua.edu.cn/1697.html>.
3. CEIC. "Taiwan Credit Card Statistics." CEIC, <https://www.ceicdata.com/en/taiwan/credit-card-statistics>.
4. "List of Largest Financial Services Companies by Revenue." Wikipedia, Wikimedia Foundation, 1 Sept. 2022, [https://en.wikipedia.org/wiki/List\\_of\\_largest\\_financial\\_services\\_companies\\_by\\_revenue](https://en.wikipedia.org/wiki/List_of_largest_financial_services_companies_by_revenue).
5. Yap, Bee Wah, et al. "Using Data Mining to Improve Assessment of Credit Worthiness via Credit Scoring Models." WWW, ScienceDirect, 15 Sept. 2011, <https://www.sciencedirect-com.libaccess.senecacollege.ca/science/article/pii/S0957417411006749#s0010>.
6. Chen, You-Shyang, and Ching-Hsue Cheng. "Hybrid Models Based on Rough Set Classifiers for Setting Credit Rating Decision Rules in the Global Banking Industry." WWW, ScienceDirect, Feb. 2013, <https://www.sciencedirect-com.libaccess.senecacollege.ca/science/article/pii/S0950705112003139#s0010>.
7. Yu, Lean, et al. "Can Machine Learning Paradigm Improve Attribute Noise Problem in Credit Risk Classification?" WWW, ScienceDirect, Nov. 2020, <https://www.sciencedirect-com.libaccess.senecacollege.ca/science/article/pii/S1059056020301969>.

8. Bai, Chunguang, et al. “Banking Credit Worthiness: Evaluating the Complex Relationships.” WWW, ScienceDirect, Mar. 2019, <https://www-sciencedirect-com.libaccess.senecacollege.ca/science/article/pii/S0305048317307648>.
9. Mwesigwa, Rogers, et al. “Credit Allocation, Risk Management and Loan Portfolio Performance of MFIs—A Case of Ugandan Firms.” WWW, 3 Sept. 2017, <https://www-tandfonline-com.libaccess.senecacollege.ca/doi/full/10.1080/23311975.2017.1374921>.
10. Storey, D. J. “Racial and Gender Discrimination in the Micro Firms Credit Market?: Evidence from Trinidad and Tobago.” ProQuest, ProQuest, Dec. 2004, <https://www-proquest-com.libaccess.senecacollege.ca/docview/220958975?pq-origsite=primo>

## Annex

### 1. Table 01: Data Description

A	B	C
Variable Name	Description	Data Type
ID	Unique Identifier of the customer	Numeric
outstanding credit	Total amount of outstanding credit of applicant and their immediate family	Numeric
Gender	1 = Male 2 = Female	Numeric
Education	1 = graduate school; 2 = university; 3 = high school; 4 = others	Numeric
Marital Status	1 = married; 2 = single; 3 = others	Numeric
Age	age during application	Numeric
Past Payments YTD	-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	Numeric
Past usage trends of last 5 months	Total credit usage in dollars for the last 5 months	Numeric
Past payments of last 5 months	Total credit repayments in the last 5 months	Numeric
Customer defaulter Report	Classification as defaulter or regular 1= Defaulter 0=Regular	Numeric