The analysis of Credit Risk

# Introduction

Financial institutions use various machine learning models to ascertain whether a candidate for a loan is either “good” or “bad”. In other words, to confirm whether an individual will default on their loan repayments or not. The general method of evaluating a person’s credit capability is via their credit score. The credit score considers an individual’s financial history which takes into account their habits, payment history for loans, and other various factors such as age, income demographic, etc. All these elements direct decisions in financial institutions when someone applies for a loan [1]. Models such as decision trees, random forests, and logistical regression can be used to fit a data pool of candidates. These methods however can cause an error in accuracy for their prediction of who will or will not default on a loan repayment [2].

Before any kind of machine learning model can be applied to a given data pool of candidates, the dataset must be pre-processed and cleaned. In this study, a simplified version of a vintage analysis was applied to analyze this given set. The vintage analysis is somewhat similar to cohort analysis in that it looks at the historical data of a customer. This data comprises a record of whether a customer was late on repayment for a month in the observation period of their loan. Further details include up to how many days they are late in their repayment in a specific month of their observation period. The observation period for the case of this study will refer to how long they have been generally overdue on repayments. The vintage analysis is a favorable way of analyzing the behavior of loan holders, as it paves a strong path for predicting the future behavior of a candidate [3].

In this study, a dataset was taken from Kaggle [4] and it comprised a credit record and an application record. The credit record contained the customer ID of the credit card holder and their record of payment for their credit card. Each month they held a credit card was recorded with the information of whether they were overdue and by up to how many days. The application record contained information about applicants for a credit card. It possessed information about that candidate’s general demographic such as age, number of family members, and financial data (income annually, employment period, etc). within this dataset, the age and period of employment are denoted with a ‘-‘ to depict these periods as historical. If a ‘+’ was in front of the period of employment it signified unemployment. This data was processed for the application of various machine learning models whose metrics signified how individually effective they were.

# Method

The data sets were predominantly read using the pandas' module and saved under the variables, ‘Credit\_record’ and ‘Application\_record’. The datasets were cleaned and filtered in stages and then had machine learning models applied to them. In the credit records, the number of days a credit card holder was overdue was denoted by a key. This key was substituted for their actual values (see Appendix 1). Both data sets were swept for missing values after removing columns with significant missing cells in Excel. A function was then created to convert days to years and applied to the application record’s ‘BIRTHDAY’ column and ‘DAYS EMPLOYED’ column. For further cleaning of the credit record, ‘X’ and ‘C’ were removed from the ‘STATUS’ column to show only the months credit card holders were overdue. The ‘STATUS’ column was converted to a string type value and letters were replaced with empty strings, and subsequently replaced with 0s. The ‘STATUS’ column was then converted to integer-type values. A replacement function was composed to denote the true values of up to how long each month's payment was overdue (refer to Appendix 1 for key).

To calculate the observation window and to determine when each account holder had their account opened, the credit record was read into a pivot table. The pivot table was indexed by the IDs and the columns ranged from 0 to 60 which was the maximum number of months an account had been opened. The values populating these columns for each applicant were the number of approximate days they had been overdue on their repayments. The observation window for each candidate was calculated by taking the difference between the maximum and minimum number of months. The observation window reflected the period the account holder had been overdue within rows containing ‘X’ and ‘C’ that had been originally removed. A pivot table summarising the observation windows was created with the values of a count of how many IDs pertained to each window. A plot was then constructed to visualize this (see Appendix 2).

The further summarise the data about the number of days overdue by a candidate, the average of these days over their observation period was calculated. Between the application record and credit record, the shared unique values were identified. The final resulting credit record only contained information for these shared values. The resulting dataset that was constructed contained the following data: Customer ID, Average Overdue Length in days, Observation Period of Overdue Payments in months, Annual income, Age in years, Period of employment in months (if “-“ employed, if “+” then unemployed and the number of family members. A variety of plots were constructed to visually understand this data. Finally, the independent variable chosen was the Observation Period of Overdue Payment with the other variables considered as dependent. Four different machine learning modules were imported and applied and then evaluated for their effectiveness.

# A picture containing colorfulness, screenshot, purple Description automatically generatedA picture containing screenshot, text, diagram, rectangle Description automatically generatedA picture containing screenshot, pixel Description automatically generatedResults/Analysis

D

C

B

A

Figure 1 – a) This figure shows a box plot denoting the distribution of number of months of employment of each credit card holder in the dataset. b) The above figure is a histogram signifying the range of age for the given dataset of credit card holders. c) This figure depicts a box plot that ranges the annual income of the credit card holders. d) The above figure shows a scatter plot of the Average overdue length in days within a given period of observation of lateness in repayment with an associated colour representing the age of credit card holders.

Figures 1a, 1b, and 1c provide a visual representation of the range of periods of employment, ages of credit card holders, and annual income of account holders respectively. The data in Figure 1a depicts that the employment period for most applicants was between 3000 to 4000 months. There were a few outliers that had their period of employment as positive indicating they had been unemployed. These were excluded from the graph as the factor of employment is relevant to repayment for loans. In the past few decades, institutions have increased their financial lending to individuals, this has left many credit holders with uncertain payment gaps. One of the reasons that can be attributed to this is the lack of security in the job market in today’s age [5]. Figure 1b highlights the age diversity of credit card holders indicating a vast difference between the youngest account holders and older ones. However, a peak can be observed in this particular dataset between -40 to -50 years approximately. Before and after this peak a rise and run can be identified indicating that debt seems to increase as age does and then decrease as age increases. This seems to be substantiated by data (see Appendix 3) from Depietro, A. and G. Lapera (2023) detailing the average credit debt for each generation. It showed that generation z (18 to 26 years) had the least amount of credit debt and generation x had the most [6]. Comparing such data to the data in Figure 1b would strongly suggest that there is also a relationship between age and the number of credit accounts. However, starts to recede as the age of the candidate increases.

From figure 1c it can be observed that most candidates earn between 40000 to 80000 per annum. There does exist a certain pool of candidates within this dataset that earn more than the 80000 benchmark. Borrowing has seen a major increase in the past two decades following the trend that salaries have also been increasing. There does seem to be a trend that can be observed between increasing income and debt as temptation amongst high earners increases as does borrowing capacity [7]. Finally, in Figure 1d there seems to be a strong indication that most credit card holders will default in their observation period for an average of up to 40 days. The outliers that default for longer periods are significantly fewer. There are many reasons why individuals may default for longer periods on credit card debt repayments. Since there is no knowing from what absolute period this dataset is from, many factors such as cost of living, currency, country of residence, etc are unknown. There are hence many limitations to this study that hinder its accuracy and veracity. If we considered this study in a simulation of conditions where it is known that some may have children or own cars. This would indicate that perhaps certain conditions associated with these factors may affect one’s ability to not default on debt repayments.

Machine learning model application to prediction of a candidate’s potential to default on a loan repayment has become ever-increasing since the 2008 global financial crisis. One model will not satisfy the complexity of predicting an individual’s financial behavior as contributing factors change from one domain to another [8]. Out of the various machine learning models applied to this dataset (please see associated code), k-means clustering provided the most accuracy. The silhouette coefficient is the metric used to evaluate the effectivity of clustering and it represents how compact clustering is. With a silhouette constant of approximately 0.54 with the number of clusters set to two. Clustering is a popular unsupervised machine learning method with many advantages such as strong pattern identification. Despite this, a major drawback of this method is the possession of many dimensions making distance measurements less meaningful.

# **Conclusion**

Though credit debt has increased in the past two decades, the average debt and amount of individuals with credit debt follow similar trends. That trend is that an increase is first observed as age increases and then after the age bracket of generation x, a decrease is observed. This can be attributed to a variety of factors, such as age, income, period of employment, and other unique factors like number of children. Unfortunately, due to a lack of further contextual data, it is unclear whether the results from this study can be applied to real-life situations

# Bibliography

1. Arya, S., et al. (2013). "Anatomy of the credit score." Journal of Economic Behavior & Organization **95**.
2. Radović, O., et al. (2021). "CREDIT SCORING WITH AN ENSEMBLE DEEP LEARNING CLASSIFICATION METHODS –COMPARISON WITH TRADITIONAL METHODS." Economics and Organization **18**.
3. Siarka, P. (2011). "VINTAGE ANALYSIS AS A BASIC TOOL FOR MONITORING CREDIT RISK " Mathematical Economics **7**.
4. . "Credit Card Approval Prediction." from <https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction?resource=download>.
5. Hodson, R., et al. (2014). "Credit Card Blues: The Middle Class and the Hidden Costs of Easy Credit." Sociol Q **55**.
6. Depietro, A. and G. Lapera (2023). "Average American debt by age and generation in 2023." from <https://www.creditkarma.com/insights/i/average-debt-by-age#average-total-debt-by-age-and-generation>.
7. Fernández‑López1, S., et al. (2023). "Self‑control and debt decisions relationship: evidence for different credit options." Current Psychology.
8. Bhatore, S., et al. (2020). "Machine learning techniques for credit risk evaluation: a systematic literature review." Journal of Banking and Financial Technology **4**.
9. Jain, A. K., et al. (1999). "Data Clustering: A Review." ACM Computing Surveys **31**.

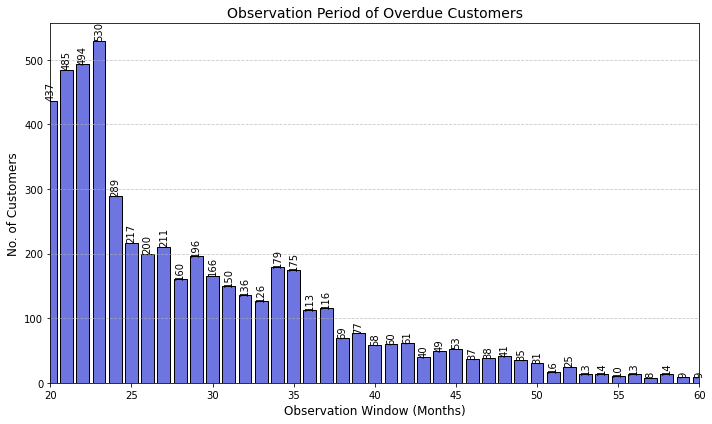
# Appendicies

Appendix 1

A picture containing text, screenshot, font, algebra

Description automatically generated

Appendix 2



Appendix 3 [6]

