Contents

[1. Introduction 2](#_Toc155175833)

[2. Machine Learning 2](#_Toc155175834)

[3. Big Data 3](#_Toc155175835)

[4. Blockchain 4](#_Toc155175836)

[5. Credit Scoring Implementation in Python 4](#_Toc155175837)

[5.1 Supervised Learning 5](#_Toc155175838)

[5.2 Unsupervised Learning 9](#_Toc155175839)

[6. Ethical Considerations 10](#_Toc155175840)

[7. Future Challenges 11](#_Toc155175841)

[8. Conclusion 11](#_Toc155175842)

[References 11](#_Toc155175843)

AI-Based Credit Scoring Assessment

# Introduction

Credit risk assessment is one of the most challenging issues in the field of financial analysis. This process can significantly mitigate the macro-damages resulted from making poor decisions when granting loans to borrowers. However, when they make an accurate decision, it will increase their income as part of paid amenities. Providing services to genuine customers is one of the primary ways financial institutions generate their revenue. In order to determine the most beneficial decisions, it is crucial to assess the customer’s ability to pay the loan together with the benefits of the amenities that have been granted. In other words, it is beneficial to minimise the risk of credit (Torabian & Azizi, 2014). Developing the credit scoring assessment system for financial institutions is one of the strategies to lower the risk indicated. In order to understand what customer wants while granting loans, financial institutions should determine the characteristics of their consumers. These operations have decreased risks, particularly credit risk, through validation. The validation refers to an action in which the creditworthiness of the potential borrowers is being assessed by using the data they have provided. This allows for a more thorough comprehension of people's financial situations and their capacity to pay back the loans that they have received on time and obtain extra amenities (Standard Chartered, 2022). Lending has been highlighted as a high-risk industry for discriminatory algorithms due to the use of historical data, but algorithmic credit scoring can dramatically enhance banks' evaluation of customers and credit risk (Sargeant, 2022). The use of AI in credit scoring also makes it possible to evaluate an individual's creditworthiness even if they don't have a standard credit history because their online activities such as social media interactions, browsing patterns, or the use of mobile applications leave records that can be reviewed.

# Machine Learning

Machine learning is a subfield of AI which focuses on the development of algorithms and models that will be used to make predictions, recommendations, and decisions by using data. Machine learning algorithms use historical data as input to predict new output values. Some of ML algorithms are neural network, linear and logistic regression, random forest, and many others. Machine learning algorithms can do a data analysis and interpretation, then act based on the result of that analysis. The banking and finance industry could enhance decision-making processes, automate jobs, improve risk management, and offer specific services to clients by utilising machine learning techniques and algorithms.

Financial institutions are now leveraging machine learning to make more accurate credit evaluations, as well as to reduce manual processes and manual errors (Frąckiewicz, 2023). Machine learning helps lenders to analyze a large amount of data faster and more accurately than traditional methods. This can help lenders to make better decisions about capturing potential borrowers, helping to reduce risk and improve loan approval rates. In addition, machine learning is being used to provide more accurate credit scores. Supervised and unsupervised machine learning models can be employed based on research by Tsai & Chen (2010). Supervised is for classification which is training model using labeled dataset to predict the output based on the input and in this project, the objective is to predict the credit score based on various factors namely age, payment history, etc. Unsupervised used for clustering which is training model with unlabeled dataset to find new pattern such as finding cluster of individuals based on transaction data or payment behaviour. By leveraging data from a many sources, lenders can make a prediction regarding the risk of a loan and create an accurate credit score. This can help lenders to minimise the amount of risk associated with a loan, as well as provide more accurate information to potential borrowers. Machine learning for credit risk prediction has a wide range of advantages. First, financial institutions may utilise it to determine the safest borrowers with the lowest default risk. Second, it could lead to higher profits as well as a reduction in losses caused by giving bad loans to customers. In addition, financial institutions can also use machine learning to spot unusual activities and possible scammers.

# Big Data

Big data is an extremely large dataset that can be analysed to gain some insights, to reveal some patterns and trends (Pence, 2014). In financial sectors, these insights are used for creating strategies for banks and financial institutions. There are 2 types of data which are structured data and unstructured data (Smallcombe, 2023). On the one hand, structured data is a well organised data that follows an established structure or format, and usually set in rows and columns. Data is normally stored inside relational databases or a spreadsheet to be analysed more easily. In structured data, relationships between data points are clearly defined, and every piece of information has a specific data type. Structured data usually stored numerical data, classification data, and dates. The most commonly used formats for storing structured data are Excel files, SQL (Structured Query Language) databases, and CSV (Comma-Separated Values) files. Unstructured data, on the other hand, are those that do not have a specific file format. Unstructured data is harder to be processed and analysed. Because it usually is not sorted in rows and columns like in structured data. Pictures, video files, audio, email, social media post, and text document are usually considered as unstructured data. The majority of information is stored in NoSQL databases or a variety of file formats. The algorithms that mostly used unstructured data are Natural language processing (NLP) and computer vision. The information for big data in financial came from various sources, for instance, stock market, client transactions, and social media.

Big Data analytics plays a critical role when analysing credit risk. The majority of the challenges that banks or lenders encounter in the sector can be resolved with this technology. In order to assess credit risk accurately, data must be sufficient and big data can provide all that information for analysation (Smith, 2021). Financial institutions used to reject providing loans to people that do not have any credit history. However, in the present day, social media or even mobile data can be utilised for assessing creditworthiness. These solutions can help financial institutions to minimise the risk of their potential borrower.

# Blockchain

Blockchain is a distributed ledger that keeps a record of all transactions in a distributed peer-to-peer network of multiple nodes. Blockchain excels in providing secured and transparent data storage and transaction. Transactions are stored in blocks, and when the blocks are completed, this new block will be linked to the previous existing blocks forming a chain, which is why it is called blockchain. Data in blockchain is immutable, meaning that every transaction that has been stored in the blocks cannot be changed, this is why blockchain is secured (Consensys, 2019). In the case of credit scoring, blockchain is not necessarily used for these solutions because credit scoring mostly relies on data processing, machine learning, and analysis of historical financial activity in order to assess creditworthiness, and blockchain features does not align with this. However, blockchain will be very useful for storing information related to credit scoring especially because financial institutions use that data to assess people’s credit score. With the help of blockchain, information used for profiling customers will be more reliable and accurate. Customer data, for example, their salary, assets, payment history should be stored inside a blockchain to make sure it is trustworthy. Because blockchain is immutable, even if a customer tries to fake the information to get loan, they cannot change the data and financial institutions can feel safer from fraud. Machine learning algorithms can also benefit since accurate training data can also improve the model performance and provide better predictions.

# Credit Scoring Implementation in Python

The solution is made in a simple code in python using machine learning and big data. Machine learning is applied here as a tool to make a prediction on whether the borrowers are worthy of getting the loan based on some factors and to find a new pattern based on clustering.

## 5.1 Supervised Learning

The factors or inputs for supervised learning are:

* Age
* Occupation
* Monthly Salary in Hand
* Annual Salary
* Credit Utilisation Ratio
* Outstanding Debt
* Credit Mix
* Number of Loan
* Type of Loan
* Number of Bank Accounts
* Total EMI per month
* Interest Rate
* Number of Delayed Payment
* Credit History Age
* Total EMI per month
* Payment Behaviour
* Amount Invested Monthly
* Monthly Balance

The machine will predict the credit score based on those 18 features above with *Good*, *Poor*, or *Standard* asthe output*.* The data used in this project is categorised as structured data since the format of the data is on CSV file and it is well structured in rows and columns, and it is numerical data. However, for this project, even though the size of the dataset is 100.00 and it is not considered as big data yet, but it will use more than that to make more accurate prediction in the future. Creating the solution will be divided into several steps which are data understanding and preprocessing, modelling, and evaluation.

First, the dataset came with more than 18 features, but not all of them are relevant for example, customer id, month, name, etc. Next is to find whether the data have a null value and find the types of the data for pre-processing method later. The code used below is to remove the null value and duplicate in the dataset, also removing an underscore (‘\_’) because most of the features categorised as string although it is a numerical value, so all of them need to be converted.

A screen shot of a computer screen

Description automatically generated

After removing null value on numerical data, the string data also needs to be checked for any unrelated value.

A screen shot of a computer code

Description automatically generated

A computer screen shot of a computer program

Description automatically generated

Based on the result, it showed that the data need further cleaning because it has value that can disrupt training process like ‘\_\_\_\_’, ‘Not specified’, etc. Next, the string value will be converted into integer value by assigning the string in a number. Here is the code.

A screen shot of a computer screen

Description automatically generated

Before making a prediction, the dataset will be split into input and output, and train and test. The train will be used to train the model, and the test will be used to validate the performance. The model that will be used in this project is Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree Classifiers (CART), Gaussian NB (NB), And Random Forest Classifiers (RF). After training the model, here is the result:

A black background with white numbers

Description automatically generated

A diagram of a number of boxes

Description automatically generated with medium confidence

Based on the result above, Random Forest has the highest score which is 77 percent, has no outliers, and the lowest score still higher than all the models. Because Random Forest is the best model, this model will be validated using the test data and try to predict. Below is the result which displayed with the accuracy, confusion matrix, and classification report.

A screenshot of a computer screen

Description automatically generated

The score is even higher than the training. Based on the confusion matrix, the true positive of each class which are good, poor, and standard are relatively good and based on the classification report, the model can predict 1 (Standard) more easily and having a hard time predicting 0 (Good) based on the F1-Score.

## 5.2 Unsupervised Learning

Next is the clustering algorithm and the data used in this code is for customer segmentation and it is different from the classification.

A screenshot of a black and white table

Description automatically generated

Clustering is useful for customer segmentation because in credit risk assessment, financial institutions will be able to differentiate between various risk profiles. Different payment behaviours can be identified to adjust the lending standards or criteria. Credit risk could vary depending on the type of customer. Financial institutions can assign different lending terms to each segment by grouping customers based on their risk profiles. This helps in reducing default rates and improving lending decisions. KMeans is used for this project because it can perform well with numerical data, and it works well with large dataset and for segmenting customer for credit assessment required a lot of data. Result from KMeans also easy to interpret and analysed. Below is the code and result of clustering using KMeans.

A computer code on a black background

Description automatically generated

A diagram of a number of dots

Description automatically generated with medium confidence

The result above segmented 3 types of customers based on income and score. However, credit risk assessment can find patterns to make better decisions, for example, based on occupation & monthly salary, annual income & payment behaviour, etc.

# Ethical Considerations

* Machine Learning Bias: Biases created by the training data may unintentionally be carried on or amplified by machine learning algorithms. Historical data that related to race, gender, or others may make the model to have a discriminatory decision. Eliminating those factors that are not relevant to assessing someone’s credit score is the solution for machine learning to make a fair decision (COINTELEGRAPH, 2023).
* Financial Inclusion Impact: The use of machine learning and big data for credit scoring maybe bring financial inclusion to some people, but some other people may be excluded due to bad credit histories and lack of digital footprints. Therefore, other factors, for example, employment history, utility bill, rental bill, or maybe educational background can be considered as one of the features to assess credit history.
* Data Quality: The data used in the solution above is not viable since it did not show where the data originated from, but it is used for making the algorithm run. In the future, data volume must be increase and quality must be ensured because it impacts the decision-making process made by the model and to avoid machine learning making an unfair credit decisions and inaccurate prediction, data must be validated and cleaned regularly to maintain data integrity.

# Future Challenges

Even though AI-based credit scoring is innovative and has the potential to increase accuracy in assessing creditworthiness, it also introduces some challenges that must be managed. It can be difficult to understand the logic behind credit score due to the complexity of AI models. The "black box" issue can make it more challenging to comply with regulations, especially in jurisdictions where lenders must clarify credit decisions to borrowers.

# Conclusion

In conclusion, credit risk assessment is an important financial activity, which helps lenders make wise lending decisions and minimise risks while generating revenue. As a powerful strategy to improve credit scoring accuracy, efficiency, and fairness, machine learning and big data is used, and even though blockchain is not implemented for the solution, it is very useful for storing information, so financial institutions can reduce fraud before giving loans. Both algorithms implemented in python on chapter 5 above are a simple implementation for the credit scoring assessment, more complex features can be added. This code is relevant because credit scoring is used to decide whether people can get loans, and there are many deciding factors that can impact the score, having the algorithm is very useful to reduce time and human errors by automation for more accurate score.

# References

Choudhary, V. (2018). Mall Customer Segmentation Data[Data Set]. https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

COINTELEGRAPH. (2023, July 9). *AI-based credit scoring: Benefits and risks*. Retrieved from COINTELEGRAPH: https://cointelegraph.com/learn/ai-based-credit-scoring

Consensys. (2019, October 10). *Blockchain in Financial Services*. Retrieved from Consensys: https://consensys.net/blockchain-use-cases/finance/

Frąckiewicz, M. (2023, March 15). *How Machine Learning Helps in Credit Scoring and Loan Approval*. Retrieved from TS2: https://ts2.space/en/how-machine-learning-helps-in-credit-scoring-and-loan-approval/

Paris, R.(2022). Credit score classification[Data Set]. https://www.kaggle.com/datasets/parisrohan/credit score-classification

Pence, Harry E. "What is big data and why is it important?." Journal of Educational Technology Systems 43.2 (2014): 159-171.

Sargeant, H. (2022, December 6). *Economic and Normative Implications of Algorithmic Credit Scoring*. Retrieved from Oxford Business Law Blog: https://blogs.law.ox.ac.uk/oblb/blog-post/2022/12/economic-and-normative-implications-algorithmic-credit-scoring

Smallcombe, M. (2023, February 16). *Structured vs Unstructured Data: 5 Key Differences*. Retrieved from Integrate.io: https://www.integrate.io/blog/structured-vs-unstructured-data-key-differences/

Smith, S. (2021, December 21). *How Big Data Analytics can make credit risk assessment more efficient and accurate*. Retrieved from spyrosoft: https://spyro-soft.com/blog/credit-risk-assessment-with-big-data-analytics

Standard Chartered. (2022). *Role and Importance of Credit Score*. Retrieved from Standard Chartered: https://www.sc.com/in/stories/role-and-importance-of-credit-score/

Torabian, A., & Azizi, K. (2014). Credit scoring of real customers: A case study in Saderat bank of Iran. European Online Journal of Natural and Social Sciences: Proceedings, 2(3 (s)), pp-2725.

Tsai, C. F., & Chen, M. L. (2010). Credit rating by hybrid machine learning techniques. Applied soft computing, 10(2), 374-380.