# UB NO: 21023380

# Literature Review

## Introduction

Technology has been the reason for some thriving businesses and financial institutions, as most of the firms have employed financial analytics as well as artificial intelligence. Businesses aspire to combine machine learning and artificial intelligence to get a faster as well as more accurate outcome. Furthermore, artificial intelligence and machine learning increased productivity by decreasing repetitive work (Ramachandran, et al., 2022). The availability of data, combined with advances in data analytics techniques and computational processing, has worked as a catalyst, allowing networks to incorporate machine learning (ML) capabilities (Irena, Kwang-Cheng, Ahmed, & Malamati, 2020). In (Shourya & Abhijit, 2020) described the concept of data science and machine learning also artificial intelligence operations inform of brain, where brain captures various data (including images, text, figures, even sounds and so on), interpret data and apply it to make decision. This is in tandem with data collection from the world to create model by data science and machine learning to aid decision by artificial intelligence.

## Carbon (formerly known as Carbon Loans and Paylater)

Carbon Loans is a Nigerian financial technology firm that renders lending service basically on the principle of automation, this is to improve the level at which individual with under-served means of accessing credit loan (Startcredits, 2022).

Chijioke Dozie and Ngozi Dozie created the Carbon Loan Company in 2012 with a $15.8 million VC investment. The brothers founded the business in a niche digital lending sector, but it has now expanded to include a variety of services ranging from savings to payments and investments (Kene-Okafor, 2019). As illustrated in Figure 1, the lending company has developed to become one of Africa's largest fintechs, winning numerous honors and accolades.



Figure 1: Awards of Carbons proving it is one of the biggest fintechs in africa

## Impart of Artificial intelligence and Data Science Application in Credit Loans

As mentioned earlier, artificial intelligence is a vital concept in the technology world as it is one of the ways of solving difficult task, which some time might be difficult to understand or too large (big data) to treat within the timeframe. Artificial intelligence is represented by a systematic way of smart agent machines, which senses the environment in order to achieve its aim. Artificial intelligence, according to (Russell & Norvig, 2016), refers to machines (computers) that mimic the cognitive and affective functions of the human mind. Artificial intelligence has advanced tremendously in recent decades, and experts have worked relentlessly to expand AI principles (Verma, Sharma, Deb, & Maitra, 2021).

AI in field of finance has witness rapid momentum, as it has been applied in so many ways including determining of borrower’s creditworthiness, default predictions, loan frauds and so on. According to (Mhlanga, 2021), who used conceptual techniques alongside review of literature to elicit the impact of artificial intelligence on borrower’s creditworthiness and loan risk assessment. In his findings, he asserted that artificial intelligence has strong influence on risk involved in accessing credit loan, by applying secondary data sources to cater for issues of adverse selection, moral threat as well as asymmetry of information. Thus, allow lending institution to make proper loan risk assessment of the borrowers.

(Eletter, Yaseen, & Elrefae, 2010) proposed a multilayer neural network along with feed forward and backward propagation machine learning algorithm to construct loan decision model for “Jordanian commercial bank”, they made use of some important attributes usually adopted by the bank’s strategies. Therefore, they came to the conclusion that the model's high accuracy shows that neural networks can be effective for classifying and evaluating loan applications for proper credit score.

Over the year, for almost a century, accurate forecast of loan default risk in lending or lending has been a critical theme for banks and other lenders. Large datasets and open source data are now widely available, and developments in computational model and algorithmic data analytics approaches have reignited interest in risk prediction problem. Furthermore, loan approval processes that are automated provide new funding options for small firms and individuals. Due to the significant cost of human engagement in the process, they traditionally had restricted access to finance (Turiel & Aste, 2020). They made use artificial intelligence techniques, linear and nonlinear deep neural networks (DNNs), as well as logistic regression (LR) and support vector machine methods. The authors suggest a two-phase model in which the first predicts loan rejection and the second predicts the likelihood of default on sanctioned loans. With a test set recall macro score of 77.4 percent, LR was judged to be the outperform in the first phase. DNNs were only used in the second phase, and they performed best for defaults, with a test set recall score of 72 percent. This demonstrates that artificial intelligence has major impact in improving current credit risk models, lowering the chance of default on provided loans by up to 70% accuracy (Turiel & Aste, 2020).

In the sphere of finance and credit risk, the advancement of artificial intelligence and machine learning is becoming increasingly essential. The goal of artificial intelligence is to use mathematical modeling approaches to imitate human intelligence and thinking. Machine learning, one of the disciplines of artificial intelligence, is altering the realm of finance and credit risk by developing new models and algorithms. In the field of credit risk, new machine learning techniques are being developed and implemented. Because credit risk requires the acquisition of data that must be carefully assessed, tested, and processed, machine learning in context artificial intelligence offers a hug help (Mhlanga & Varaidzo, 2021).

(Mogaji, Soetan, & Kieu, 2020) in their study discussed the issues combatting businesses and attempt to incorporate artificial intelligence to digital marketing and their financial services. Also, they investigated the relationships that exist among artificial intelligence, digital marketing and financial services; taking into consideration the vulnerability of the customers. The study identified a new thorough framework to map the stream from consumer input data through information processing and delivery to AI-enabled digital marketing, which creates offerings and experiences that meet and exceed customers' needs. This offers data that helps financial services providers and corporate leaders understand the ramifications of artificial intelligence and digital marketing data.

**Challenges of Artificial intelligence and Data Science Application in Credit Loans**

There is no doubt as regards the influence of artificial intelligence on forecasting and analyzing data related to credit loan. On the other hand, there exist some challenges experienced in application of artificial intelligence and data science. This may include supply of falsify data by the borrowers, fraud and so on.

As presented by (Tfaily, 2017), the current state of affairs concerning information imbalance and credit risk management. Credit loan in financial or lending institution come with huge risk, according to (Tfaily, 2017), this is a factor that is always present in banking firms, and it usually entails the uncertainty of achieving particular levels of profit or even the possibility of a loss.

(Saito & Daisuke, 2018) In their research carried out, looked into whether there was any perverse incentive or disposition effect in credit guarantee systems for small and medium-sized businesses (SMEs). Most credit guarantee organizations have major issues distinguishing low risk from hazardous borrowers, according to Saito and Tsuruta (2018). As a result, they tend to attract a large proportion of dangerous borrowers, resulting in inefficient resource allocation. The study discovered that adverse selection and moral hazard were consistent with the data when using bank-level data to determine whether the default rate is positively associated with the ratio of guaranteed loans to the total loans.

In the study carried out by (Chang & Jing, 2018), they highlighted some possible challenges that may be experienced in analyzing and working with big data most especially in credit loan institution. This includes distinction between loan big data (such as online and offline) report and traditional report; disordered, unreasonable and falsified data; non-transparency of dataset, and discrimination. Mobile location data can easily lead to inaccuracies. Individual credit activities can be severely harmed by inaccuracies in data. While big data might enhance speed and connectivity, it can also make it more difficult to double-check data integrity and authenticity, making the situation worse. Individuals are unable to identify unfair credit scores due to non-transparent issues, and hence fail to improve or prevent ratings from sliding deeper (Hurley & Adebayo, 2016). According to (Guzelian, Stein, & Akiskal, 2015), loan servicers treat their data as commercial secrets. As a result, the data collection process would be opaque; if a rival discovered data resources, customers would be unable to access the data, and they would also be unable to determine whether the data was correct.

## Conclusion

Determining the creditworthiness of borrowers is a vital area in financial as well as lending institution, integrating artificial intelligence or data science to it help to reduce the default risk attached to credit loan approval. Prediction of loan default have carried out by numerous researcher which affirmed the fact about impact of artificial intelligence. It is no doubt that artificial intelligence can also be used to track fraud among the borrower.

# 

# Section B: Data Analysis

The Paylater dataset was used to perform detailed analysis, with the results, interpretations and recommendations for financial performance presented in this section.

## Dataset Description

The dataset consists of 159596 rows and 28 columns. The columns of the dataset were described below:

|  |  |  |  |
| --- | --- | --- | --- |
| **S/N** | **Column Name** | **Column Description** | **Data type** |
| 1 | loanId | ID of the loan provided to a client | Character |
| 2 | clientIncome | Income of the client of who took the loan | Number |
| 3 | incomeVerifed | Whether the income was verified or not | Integer |
| 4 | clientAge | Age of the client | Integer |
| 5 | clientGender | Gender of client | Character |
| 6 | clientMaritalStatus | Marital status of the client | Character |
| 7 | clientLoanPurpose | Purpose of applying for the loan | Character |
| 8 | clientResidentialStatus | Residential status of the client | Character |
| 9 | clientTimeAtEmployer | How long the client has been employed for their new job | Character |
| 10 | clientNumberPhoneContacts | Phone contact of the client | Number |
| 11 | clientAvgCallsPerDay | Average calls made by client per day | Number |
| 12 | loanNumber | Number of the loans taken | Integer |
| 13 | loanAmount | The amount of the loan | Integer |
| 14 | interestRate | The amount paid as interest | Number |
| 15 | loanTerm | The term of the loan | Integer |
| 16 | maxAmountTaken | The maximum amount taken by a client | Integer |
| 17 | maxTenorTaken | The maximum tenor of the loan | Integer |
| 18 | settleDays | Number of days to settle the loans | Integer |
| 19 | paymentRatio | Ratio of payment | Number |
| 20 | firstPaymentDefault | First default at payment | Integer |
| 21 | loanDefault | Whether the loan was defaulted or not | Integer |
| 22 | LoanIncomeRatio | Ratio of loan to income | Number |
| 23 | ApplicationDateMonth | Month number of application date | Integer |
| 24 | ApprovalDateMonth | Date of approval | Integer |
| 25 | DisbursementDateMonth | Date the loan was disbursed | Integer |
| 26 | repaidDateMonth | The date loan was repaid | Character |
| 27 | FirstPaymentDueDate | The date first payment is to be made | Integer |
| 28 | DueDateMonth | Date the loan was due for repayment | Integer |

## Rationale for Choosing the Dataset

The Paylater dataset was chosen because it contains the necessary features and attributes that make it suitable for providing answers to our research questions. Important analysis was carried out on the data. This was made possible with the use of the data analytics software, R and its interactive development environment, RStudio.

## Research Questions

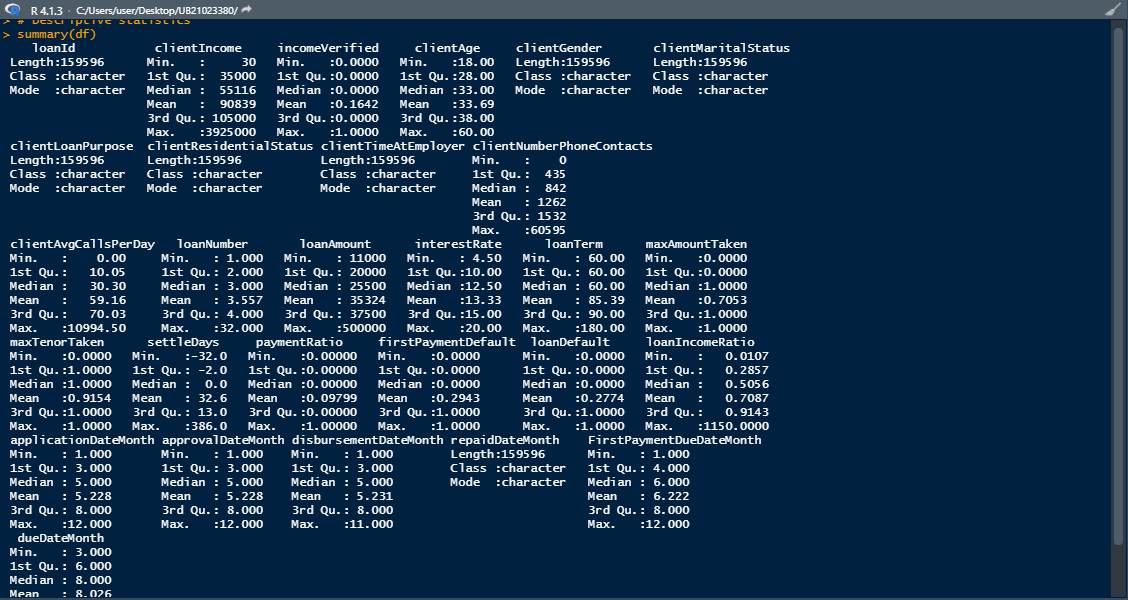
The important research question this analysis seeks to provide answers to include:

* What genders appled for most loans from Paylater?
* What is the maximum loan amount offered to a client by Paylater?
* For what purpose clients apply for loan the most?
* What is the average income of the clients to which Paylater provide loans?
* What impact do income of the clients, age of the clients, number of loan taken, and interest rate have on loan amount?
* How do the income of the clients, number of loan taken, loan amount, age of the clients, interest rate, and loan term affect the chances of loan default?
* How can Paylater increase its client base?
* How can Paylater increase it profit?

## Descriptive Analysis

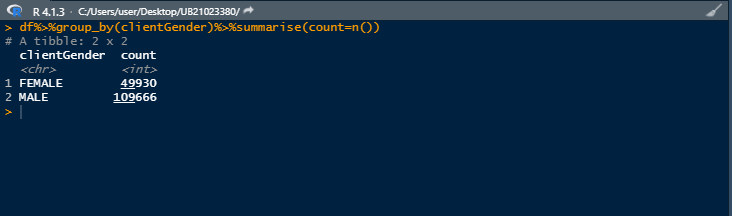
Descriptive analysis enables us to obtain relevant statistical properties of the variables included in our analysis. It also provides answers to the research questions raised earlier. The statistical properties of the individual variable used for this analysis was obtained by using the **summary** function in base R and providing it with the data. The results are presented in Figure 1. Variables that are of character data type do not have statistical properties. The numerical variables have their measures of central tendencies such as median and median, and measures of dispersion such as minimum, 1st Quartile presented in the figure.

The descriptive statistics revealed for instance that the average income of the clients who applied for loans was 90,839.

****

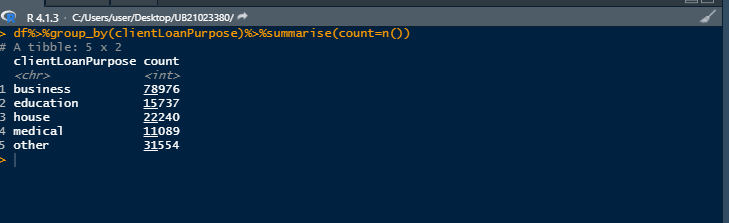
**Figure 1: Descriptive Analysis Result**

It was revealed by the analysis conducted that male applied for loan more than their female counterpart. According to the result presented in Figure 2, 109666 male clients have applied for loans from Paylater, while 49930 female clients have applied for loans from the loan company.

****

**Figure 2: Count of Gender that Applied for Paylater Loans**

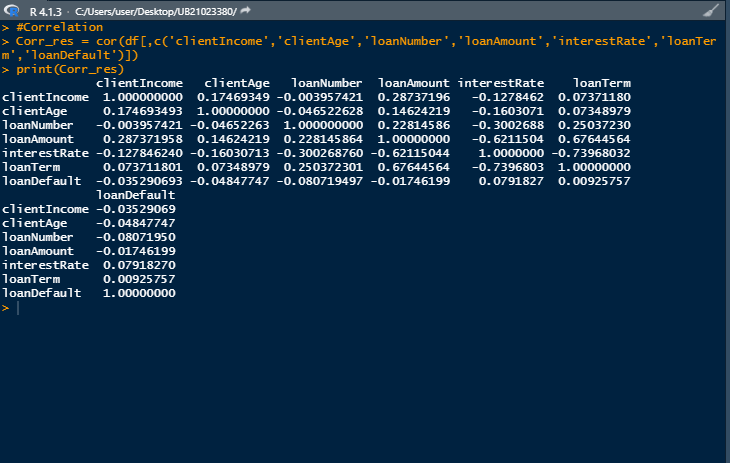
As regards the purpose for which clients apply for loans from the company the most, it was revealed that most clients apply for loan for the purpose of using it in their business.

****

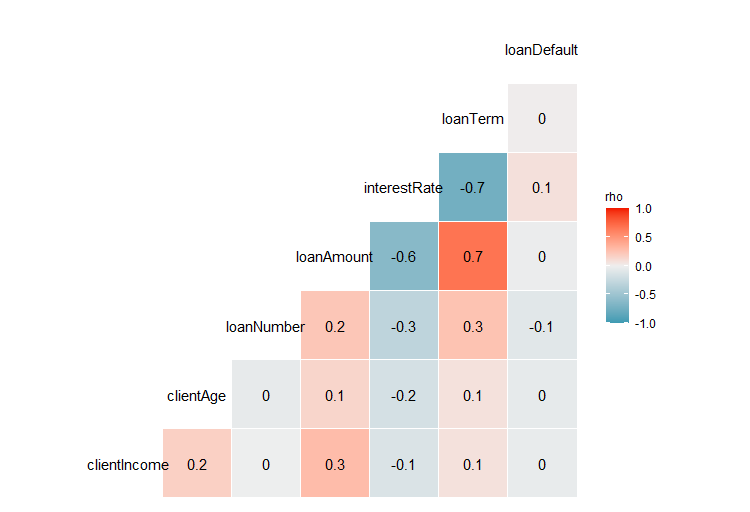
**Figure 3: Purpose for Which Clients Apply for Loan**

## Correlation Analysis

When the aim is to know what association exists among the variables of study, correlation analysis becomes essential. The **corr** function in base R was employed to perform correlation analysis on the data of the loan company, with the result presented in Figure 4. To visually present the correlation result, a correlation matrix was used. The GGally R library which provides option to build a correlation matrix was utilized in presenting the correlation result in a graphical way. The correlation matrix was presented in Figure 4.5. According to the correlation matrix, the deeper the red color, the stronger the positive correlation among the variables, and the deeper the blue color, the stronger the negative correlation among the variables. Boxes with grey color signified absence of correlation or low correlations among the variables.

****

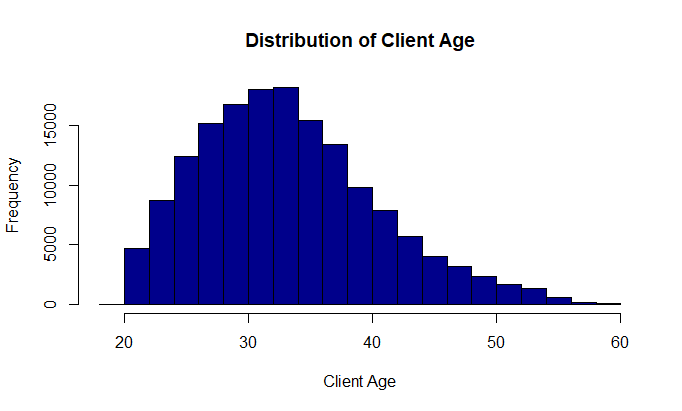
**Figure 4: Correlation Analysis Output**

****

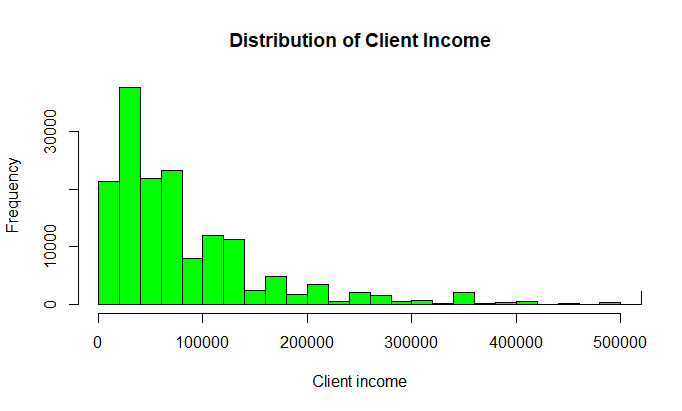
**Figure 5: Visual Display of Correlation Analysis**

## Data Visualisation

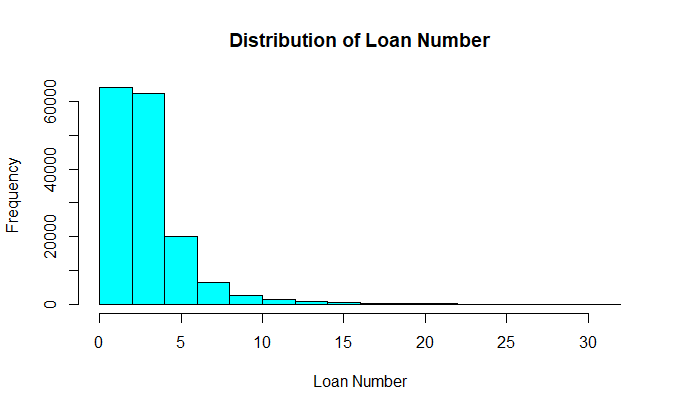
The Histogram was constructed using **hist** function the base R. It was revealed by the visuals that client age and loan amount were as close to normal distribution as possible. It revealed also that client income and loan number were skewed to the left, indicating that they were not normally distributed.

****

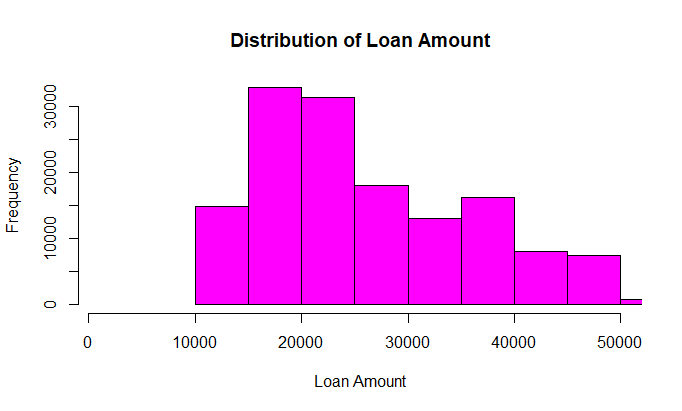
**Figure 6: Histogram of Age Distribution of Clients**

****

**Figure 5: Distribution of Client Income**

****

**Figure 7: Distribution of Loan Numbers**

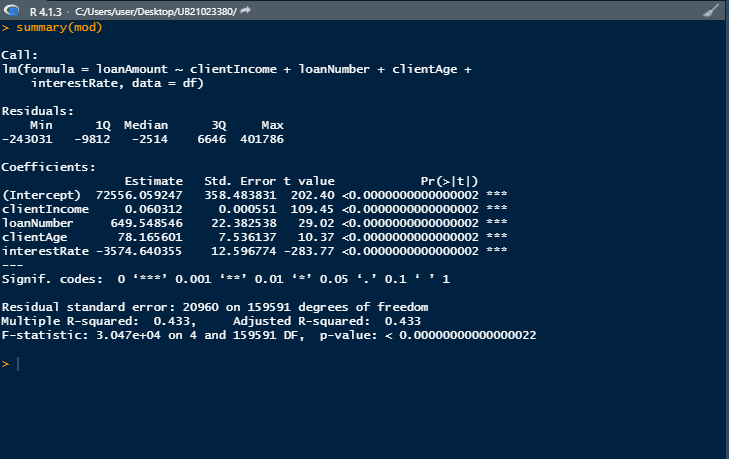
****

**Figure 8: Distribution of Loan Amount**

## Predictive Analysis

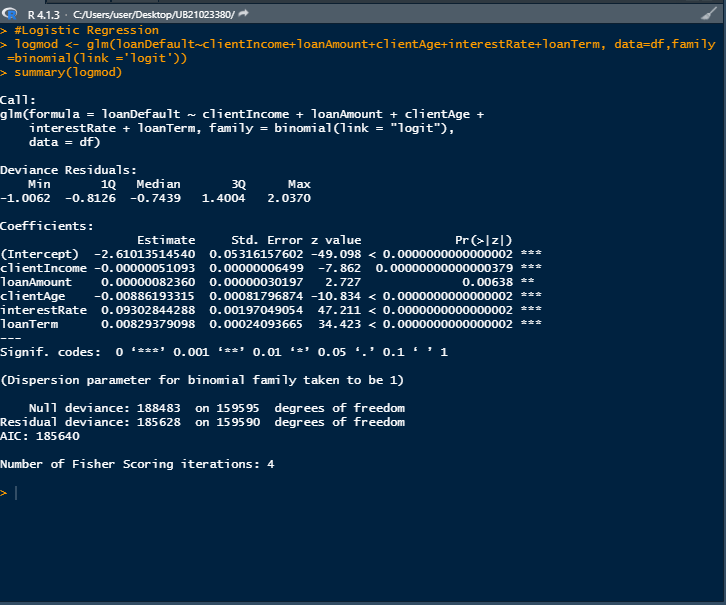
Predictive analysis such as conducted using Paylater dataset

### Multiple Linear Regression



**Figure 9: Multiple Linear Regression Result**

### Logistic Regression

****

**Figure 10: Logistic Regression Result**

## Interpretation of Results

The result of the predictive analysis conducted using the loan data was presented in this section.

### Interpretation of Multiple Linear Regression

A multiple linear regression was employed to examine the impact income of the clients, and age of the clients, number of loan taken, and interest rate have on loan amount. The estimated regression model was given as:

loanAmount = 72556.0592 + 0.06031clientIncome + 649.5485loanNumber + 78.1656clientAge - 3574.6404interestRate

According to the estimated regression model, a unit increase in client will increase loan amount by 0.06031, holding all other variables constant. A unit increase in loan number will increase loan amount by 649.55, holding all other variables constant. A unit increase in client age will increase loan amount by 78.17, holding all other variables constant. A unit increase in interest rate will lead to a 3574.64 decrease in loan amount. The constant in the model was 72556.06. This is the value of loan amount when all the variables are held constant. The p-value of client income was 0.000, a value less than 0.05; this showed that it was statistically significant to the model. The p-value of loan number was 0.000, a value less than 0.05; this showed that it was statistically significant to the model. The p-value of client age was 0.000, a value less than 0.05; this showed that it was statistically significant to the model. The p-value of interest rate was 0.000, a value less than 0.05; this showed that it was statistically significant to the model. The coefficient of determination was 0.43. This showed that the proportion of variation in loan amount explained by all the independent variables combined () was 43%.

### Interpretation of Logistic Regression

To examine how income of the client, number of loan taken, loan amount, age of the clients, interest rate, and loan term affect the chances of loan default, a logistic regression was conducted. The dependent variable- loan default was one that can take only two possible values, which are 0 and 1. This made logistic regression suitable for this purpose. The estimated model was given as:

p = exp(-2.6101 – 0.0001clientIncome – 0.000loanAmount – 0.0089clientAge + 0.0930interestRate + 0.0083loanTerm )/[1+exp(-2.6101 – 0.0001clientIncome – 0.0001loanAmount – 0.0089clientAge + 0.0930interestRate + 0.0083loanTerm)]

From the output of the logistic regression model, the coefficient estimates of each independent variable and the intercept were obtained.

* The coefficient of client income was -0.0001, which is a negative value. This shows that an increase in client income can decrease the probability of loan default.
* The coefficient of loan amount was -0.0001, which is a negative value. This shows that an increase in loan amount can decrease the probability of loan default.
* The coefficient of client age was -0.0089, a negative value. This shows that an increase in client age can decrease the probability of loan default.
* The coefficient of interest rate was 0.0930, a positive value. This shows that an increase in interest rate can increase the probability of loan default.
* The coefficient of loan term was 0.0083, a positive value. This shows that an increase in loan term can increase the probability of loan default.
* The intercept in the estimated model is -2.6101

## Recommendations for Performance Improvements

Based on the findings made, it is recommended that Paylater carry out regular media campaign to create awareness about its products and services. This way the company will be able to increase its client base. Also, the company should give out loans to applicants at a friendly interest rates. This will help reduce the rate at which loan defaulting occurs, and consequently help the company boost its revenue.

# References

ALTEN Calsoft Labs. (2018). *Loan Default Prediciton.*

Chang, V., & Jing, L. (2018). A Discussion Paper on the Grey Area – The Ethical Problems Related to Big Data Credit Reporting. *In Proceedings of the 3rd International Conference on Internet of Things, Big Data and Security (IoTBDS 2018)* (pp. 348-354). SCITEPRESS - Science and Technology Publication.

Cheema, A., & McGaugh, M. (2019). *Loan Default Prediction.* Oklahoma State University. Oklahoma: Oklahoma State University.

Eletter, S. F., Yaseen, S. G., & Elrefae, G. A. (2010). Neuro-Based Artificial Intelligence Model for Loan Decisions. *American Journal of Economics and Business Administration*, 27-34.

Ereiz, Z. (2019). Predicting Default Loans Using Machine Learning (OptiML). *27th Telecommunications forum TELFOR 2019.* *27*, pp. 19-22. Serbia, Belgrade: IEEE.

Guzelian, C. P., Stein, M. A., & Akiskal, H. S. (2015). Credit Scores, Lending, and Psychosocial Disability. *BUL Rev*.

Hurley, M., & Adebayo, J. (2016). ‘Credit scoring in the era on big data. *Yale JL & Tech*.

Insider Intelligence. (2022). *Chatbot market in 2022: Stats, trends, and companies in the growing AI chatbot industry*. Retrieved April 10, 2022, from Business Insider: https://www.businessinsider.com/chatbot-market-stats-trends?IR=T

Irena, A., Kwang-Cheng, C., Ahmed, E. K., & Malamati, L. (2020). Data science and artificial intelligence. *IEEE Communication Magazine*.

Kene-Okafor, T. (2019). *Nigerian digital bank Carbon hit $240M in payments processed last year, up 89% from 2019*. Retrieved April 01, 2022, from TechCrunch: https://techcrunch.com/2021/02/01/nigerian-digital-bank-carbon-hit-240m-in-payments-processed-last-year-up-89-from-2019/

Markets & Markets. (2019) *Artificial Intelligence Market worth $190.61 billion by 2025 with a Growing CAGR of 36.6%*.  <https://www.marketsandmarkets.com/PressReleases/artificial-intelligence.asp%20.asp> Accessed 11 April 2022

Mhlanga, D. (2021). Financial Inclusion in Emerging Economies: The Application of Machine Learning and Artificial Intelligence in Credit Risk Assessment. *International Journal of Financial Studies*.

Mhlanga, D., & Varaidzo, D. (2021). Determinants of Financial Inclusion in Southern Africa. *Studia Universitatis Babes-Bolyai*, 39-52.

Mogaji, E., Soetan, T. O., & Kieu, T. A. (2020). The implications of artificial intelligence on the digital marketing of financial services to vulnerable customers. *Australasian Marketing Journal*, 235-242.

Ramachandran, K. K., Apsara, M. S., Shibani, H., Asokk, D., Bhaskar, B., & Pitroda, J. R. (2022). Machine learning and role of artificial intelligence in optimizing work performance and employee behavior. *materialstoday*, *51*, pp. 2327-2331. doi:https://doi.org/10.1016/j.matpr.2021.11.544

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: A modern approach* (3rd ed.). Pearson Education Limited,.

Saito, K., & Daisuke, T. (2018). Information asymmetry in small and medium enterprise credit guarantee schemes: Evidence from Japan. *Applied Economics*, 69-85.

Shourya, S., & Abhijit, b. (2020). Impact of Data Science, Machine Learning & Artificial Intelligence on Humans. *IOSR Journal of Computer Engineering (IOSR-JCE)*, 26-35.

Startcredits. (2022, March 23). *Paylater*. Retrieved from Start Credits: https://startcredits.com/loans/paylater/

Sweezey, M. (2019). *Key Chatbot Statistics to Know in 2019*. Retrieved April 10, 2022, from Sakes Force: https://www.salesforce.com/blog/chatbot-statistics/

Tfaily, A. (2017). Century Managing Information Asymmetry and Credit Risk—A Theoretical Perspective. *. Proceedings of the 11 the International Management Conference the Role of Management in the Economic Paradigm of the XXI.*

Turiel, J. D., & Aste, T. (2020). Peer-to-peer loan acceptance and default prediction with artificial intelligence. *Royal Society Open Science*.

Uberall. (2019). *STUDY: 80% of Consumers Report Chatbot Experiences as Positive Yet nearly 60% still lack strong enthusiasm for the technology, according to new Uberall study*. Retrieved April 10, 2022, from uberall: https://uberall.com/en-us/company/press-releases/study-80-of-consumers-report-chatbot-experiences-as-positive-yet-nearly-60-still-lack-strong-enthusiasm-for-the-technology-according-to-new-uberall-study

Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights*.

Wu , C.-F., Mahdi, F., David, M. C., & Panagote, M. P. (2020). Credit guarantee mechanism with information asymmetry: A single-sourcing model. *International Journal of Production Research* , 77-93.

Wu, W. (2018). *Final Project: Prediction of Loan Default with Machine Learning Method.* University of California, Berkeley, Industrial Engineering & Operations Research. California, Berkeley: University of California, Berkeley.

Appendix

