6006CEM ML

Personal loan eligibility with AI

Contents

[Abstract 2](#_Toc152813996)

[Literature Review 2](#_Toc152813997)

[Methodology 3](#_Toc152813998)

[Dataset 3](#_Toc152813999)

[Pre-Processing 4](#_Toc152814000)

[Logistic Regression 4](#_Toc152814001)

[Neural Networking 5](#_Toc152814002)

[K Nearest Neighbour 7](#_Toc152814003)

[Linear Regression 8](#_Toc152814004)

[Conclusion 9](#_Toc152814005)

[References 10](#_Toc152814006)

# Abstract

With new developments in AI The aim of this work is to investigate the optimal Machine Learning (ML) method to predict loan approval for individuals. The ML methods investigated are: Linear regression, logistic regression, neural networking, and K nearest neighbour (KNN). Bad bank loans have previously crippled the global economy leading to the economic crash of 2008, a banks ability to assess risk in loaning money is integral to its continuing functioning (Lal, 2010). Every loan has the risk of defaulting, the failure to make payment, and in these situations a bank must ensure the loan is backed by a substantial collateral, i.e., a house. AI integration in this sector could ensure the reliability of the recipient of the loan and ensure that the recipient gets a loan that fits their needs, benefiting both the recipient and the bank. Estimating whether an individual should receive a loan is uncertain as one factor may outweigh others, for instance an applicant with a bad credit rating but high income may get rejected because the credit rating shows they don’t pay their loans on time. Ensuring accurate machine learning models is imperative for financial integration. Inaccuracies can lead to biased decisions, increasing social inequalities and potential negative effects on the economy. Accurate models strengthen risk management for banks and people looking for loans; precise assessments increase consumer trust in banks and the lending process. As machine learning is shaping the financial landscape and an increasing number of tasks being automated refined machine learning models are integral to upholding ethical lending practices.

# Literature Review

A literature review was conducted in order to compare and determine the best Machine Learning (ML) models for loan eligibility. The Aim to present the ML methods used by researchers to predict an individual’s chance of loan approval. According to Statistica (2020) machine learning was the most commonly used AI application by investment banks, with 63% reporting use. Bensic (2005), investigated small business lending, however they used a relativly small dataset of only 160 applicants. They used supervised learning with an unsupervised kernel in the hidden layer and back proporgation. Their neural network had a score of 77.37, correctly classifying 84.62% of the data with an error rate of 15.38%. In B. Aditya’s (2022) paper a 77.27% accuracy was found for logistic regression in predicting loan eligability, whereas the KNN methodology only resulted in a 72.27% accuracy on their data set. A difference of 5% may not seem like much, but in the finance sector can be a difference of millions of dollars. Baesens (2005) investigated the use of neural networking for personal loan eligibility using data from a UK financial institution in order to predict which customers are more likely to default. However in their subset data of 15,000 observations there were gaps in data which they filled in with the mean value for continuous variables, they do not specify how many values were inputted this way. This will have smoothed their data even just slightly leading to potentially skewed results. Their results showed their logistic regression classifier had an accuracy of 69.08% and the neural network had an accuracy of 70.36%. With only a 1.28% difference the potentially skewed data could have effected this. As B.Aditya (2022) found logistic regression to be the most effective when compared to KNN and Baesens (2005) paper concluded neural networks to be most accurate, all three will be investigated and comapred. How do all of these comapare to a more classic type of machine learning such as linear regression? The Linear regression model serves as a robust tool for organising data analysis. It enables the examination of two variables exhibit a linear relationship and allows for the calculation of degree of strength of that linear relationship.

# Methodology

## Dataset

This investigation was done using Jupyter Notebook with data from  
Bank\_Personal\_Loan on Kaggle (Tofighizavareh). The dataset consisted of 14 columns and  
5000 rows and is summarised as such:

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Data Type |
| ID | Numerical identification of individual | Numeric |
| Age | Age of individual | Numeric |
| Experience | Years of professional experience | Categorical |
| Income | Annual income of individual | Numeric |
| ZIP Code | Home address ZIP code | Categorical |
| Family | Size of individuals family | Numeric |
| CCAvg | Average credit card spending per month | Numeric |
| Education | Education level where: Undergraduate=1, Graduate=2, Advanced=3 | Categorical |
| Mortgage | Value of House Mortgage | Numeric |
| Personal Loan | Did individual receive a personal loan in last campaign? | Categorical |
| Securities Account | Does the individual have a preexisting securities account with the bank? | Categorical |
| CD Account | Does the customer have a certificate of deposit account with the bank? | Categorical |
| Online | Does individual use online banking? | Categorical |
| Credit Card | Does individual use credit card issued by Universal Bank? | Categorical |

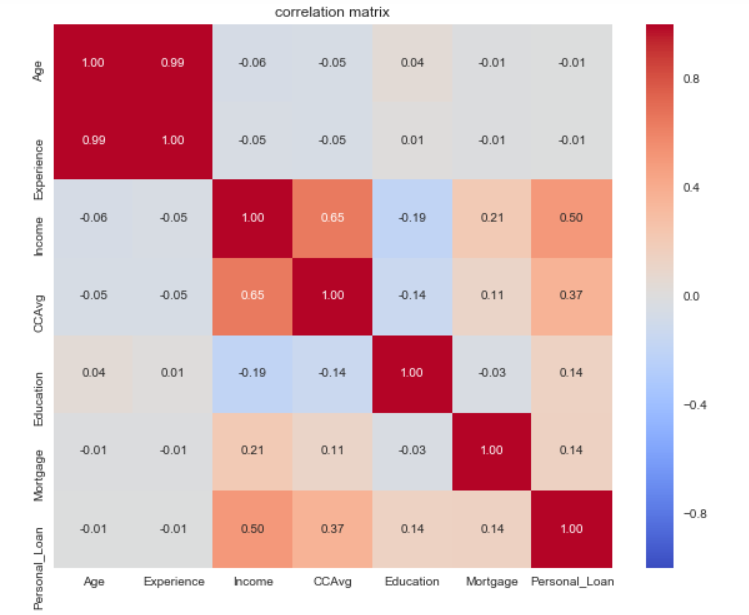


Figure 1: Correlation Matrix

The above figure shows the correlation matrix between: Age, Experience, Income, CCAvg, Education, Mortgage and Personal Loan. The matrix shows that there is a positive relationship between Income and Personal Loan of +0.5, this is the strongest relationship that affects Personal Loan. The influence of variables Education, Age, Income, CCAvg and Mortgage have particular effect on Personal\_Loan in the correlation matrix.

# Pre-Processing

Exploratory data analysis (EDA) was used to assess the dataset and identify patterns. During this phase the dataset is processed to remove unwanted data and assess what to do with incomplete data. The data straight from the CSV file were all object types but contained only numerical values so had to be converted with pandas.to\_numeric before any review of the data could be done. All Null values were checked for and removed with df.isnull().values.any().

# Logistic Regression

Logistic regression is a classification algorithm based on a binary variable of the dataset. In this case it is acceptance/ rejection of a loan application denoted as 1/0 in the ‘Personal\_Loan’ column, making this a supervised learning model. To calculate logistic regression the binary variable must be categotical where as the dependent variables, which are the other columns, can be numeric or categorical.

The Logistic function:



Where:

- intercept



- Regression coefficent



1 + e - exponential function

The dataset, having multiple columns to compare is classed as a multiclass case, because of this the algorithm uses a one-vs-rest (OvR) scheme with the one being Personal\_Loan.

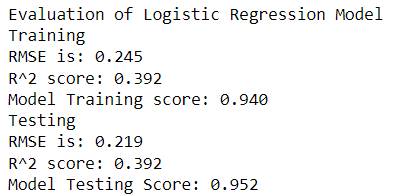


Figure 2: Output from Logistic regression training and testing

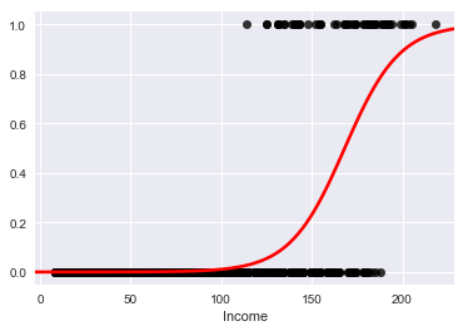
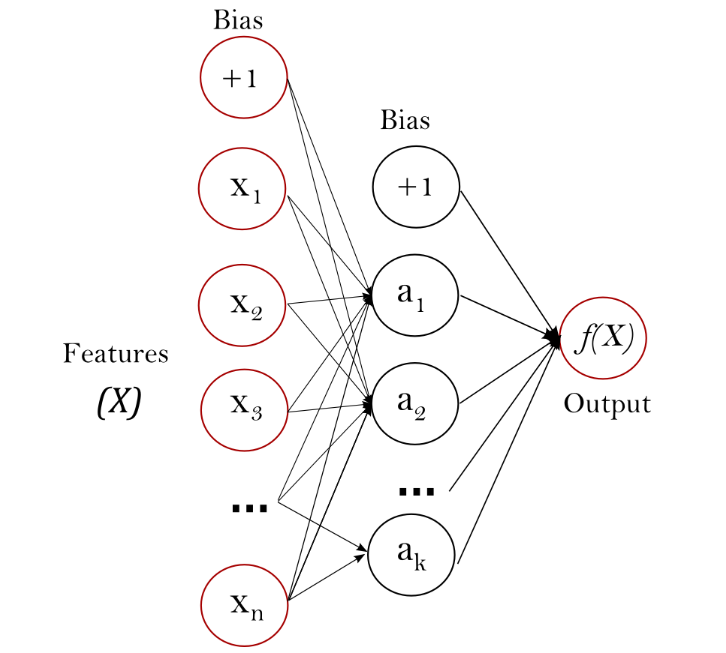


Figure 3: Logistic Regression

# Neural Networking

The neural network function used was a multi-layer perceptron (MLP) learning algorithm. It aims to act as the neurons of the brain, creating links between data points to solve patter recognition and interpolation. Figure 2 below shows an input x which is passed to the hidden layers under ‘Bias’. Each layer transforms the data with weighted linear summation, the output is then put into f(X) a non-linear activation function before becoming output values.



**Figure 4: Example of a one hidden layer MLP (scikit-learn, 2023)**

The MLP regressor implemented trains the MLP with back propagation, circumventing the output layer, which fine tunes the weights of the neural network. It uses the square error for the loss function to tune the weights (Murtagh, 1991).

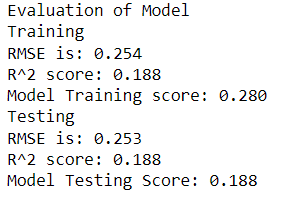


Figure 5: output from Neural Network training and testing

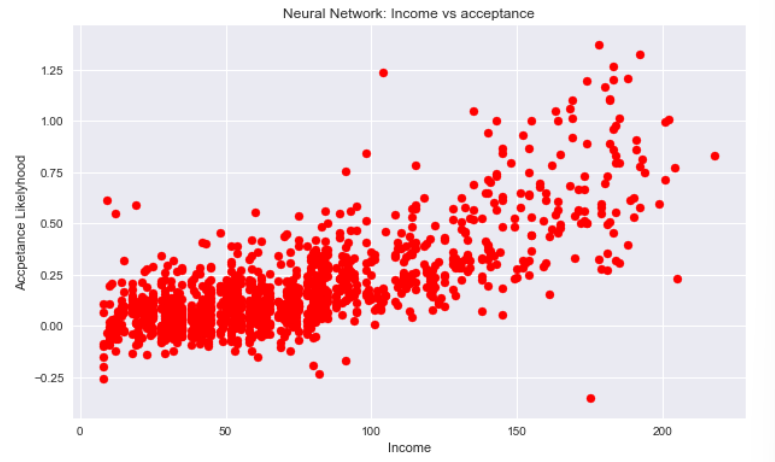
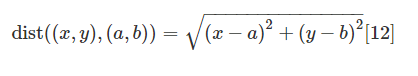


Figure 6:neural network

# K Nearest Neighbour

K nearest neighbour (KNN) is a machine learning algorithm that calculates the distance between a test sample and training samples using Euclidian distance (U.E. Orji, 2022). When a point is to be predicted it’s based on the average values of its k nearest neighbours.



(x,y) and (a,b) are the coordinates of the two points. It does not make any assumptions about the data it is processing making it a non parametric algorithm. As it searches for patterns in data without any assumptions it makes for great analysis on practical data such as clarifying what factors effect loans without any outside prejudice (M.A. Mukid, 2018).

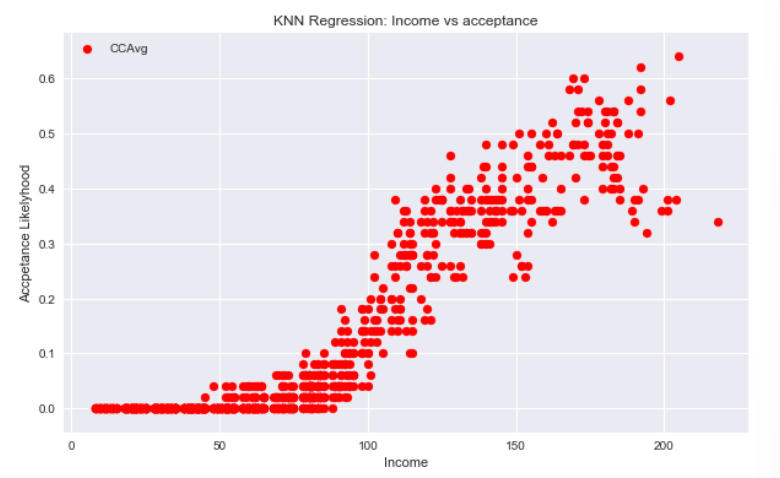


Figure 7: K nearest neighbours

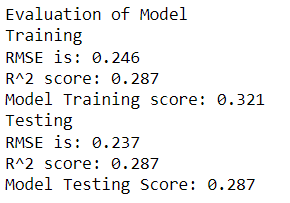
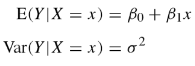


Figure 8: k nearest neighbours testing and training scores

# Linear Regression

Linear regression is a simple analysis technique to predict data that creates a line of best fit. The model uses mean fucntion and variance function thus:



(Weisberg, 2005)

Where:

is equal to E(Y|X = x) when x is zero



is the rate of change in E(Y|X = x)

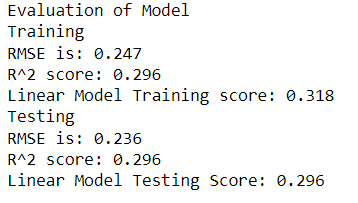


Figure 9: output from training and testing linear model

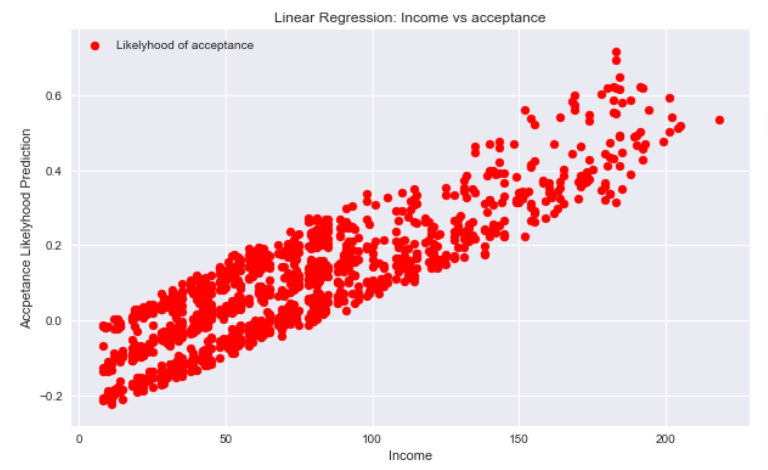


Figure10: linear model

The graph shows a correlation between an individual’s income and their likelihood of acceptance, despite having a very low training and test score of 0.318 and 0.296 respectively.

# Conclusion

|  |  |  |
| --- | --- | --- |
| Model | Training score | Testing score |
| Logistic regression | 0.940 | 0.952 |
| Neural network | 0.280 | 0.188 |
| KNN | 0.32 | 0.287 |
| Linear regression | 0.318 | 0.296 |

The Graphs in both neural network and KNN models shows a casual association between classifiers and whether an individual gets offered a loan within this dataset. Logistic Regression is the obvious best choice for this modelling with the highest training and test score showing the performance of the model.

More data about the individuals such as race and background, would be needed to make a more accurate model for predicting a loan offer. This sort of data would show any prejudice within the financial sector. Machine learning methods are only as good as the data that they are enacted on. Algorithms may learn to reproduce these biases and in sectors such as finance this could bankrupt individuals. Scrutinizing model outputs and diversifying training data is central to mitigating these biases and promoting equitable lending practices.

# References

B, A. (2023). Improved Decision Support System for Personal Loan Elibibility Using Artificial Neural Networks.

B. Aditya, V. N. (2022). Prophecy of Loan Approval by Comparing Decision Tree with Logistic Regression, Random Forest, KNN for better Accuracy. *Jornal of Pharmaceutical Negative Results*, 759-768.

Baesens, B. (2005). Neural Network Survivial Analysis for Personal Loan Data. *Jornal of the Operational Research Society Vol. 56 (9)*, 1089-1098.

Bensic, M. (2005). Modeling small-business credit scoring by using logitic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal. Vol. 13 (3)*, 133-150.

D. Thrinath Reddy, L. R. (2023). Predicting Vehicle Loan Eligibility using Random Forest Comparing with Linear Regression Based on Accuracy. *AIP Conference Proceedings (Vol. 2822, No.1).* AIP Publishing.

Lal, D. (2010). The Great Crash of 2008: Causes and Consequences. *Cato J. Vol.30*, 265.

LaValley, M. (2008). Logistic Regression. *Circulation. Vol.117 (18)*, 2395-2399.

M.A. Mukid, e. a. (2018). Credit scoring anlysis using weighted k nearest neighbour. *Jornal of Pysics: Conference Series. Vol.1025 (1)*, p. 012114.

Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, 183-197.

Reddy, C. (2022). Machine Learning based Loan Eligability Prediction using Random Forest Model. *2022 7th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1073-1079). IEEE.

Sarungu, C. (n.d.). *Loan Eligibility Prediction Using Logistics Regression Algorithm.* Indonesia: WhiteGate.

scikit-learn. (2023). *1.17. Neural network models (supervised)*. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/neural\_networks\_supervised.html

Statistica. (2020, October). *Most Commonly used A.I application in Investment Banking Worldwide 2020 by Types*. Retrieved from Statistica: https://www.statista.com/statistics/1246874/ai-used-in-investment-banking-worldwide-2020/

Tofighizavareh, Z. (n.d.). *Bank\_Personal\_Loan.* Retrieved from Kaggle: https://www.kaggle.com/datasets/zohrehtofighizavareh/bank-personal-loan/data

U.E. Orji, C. U. (2022). Machine Learning Models for Predicting Bank Loan Eligibility. *IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development*, 1-5.

Weisberg, S. (2005). *Applied Linear Regression. Vol 538.* John Wiley & Sons.