FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 23, December, 2026, pp 150 – 161
DOI: https://doi.org/10.33003/fjs-2026-1023-6193

Modelling Loan Default Data of German Customers: Predictive Machine Learning Studies

Emmanuel Idowu Olamide

Department of Statistics, Federal University of Technology, P.M.B 704, Akure, Nigeria.

*Corresponding authors' email: eiolamide@futa.edu.ng

Received: 08 Oct. 2026
Accepted: 24 Oct. 2026
Published: 13 Nov. 2026

ABSTRACT
This study considered logistic regression and random forest models in modelling the loan default data of German Customers. The data were randomly partitioned into an 80% training set and a 20% test set. The train set was used to fit the data, while the test set served as the out-of-sample data used to evaluate the performance of the model. On the train set, the random forest performs better than the logistic regression model across all the evaluation metrics. On the test dataset, the random forest has superior specificity, precision, accuracy, and F1-score, but the logistic regression model produced a better sensitivity. This means that the model was better at identifying a bad loan; this is, in fact, the desire and priority of every financial institution – a model that is capable of detecting the risk of loan default. In terms of efficiency, the logistic regression model produced lower variance than the random forest, which had high variance.

Keywords: Loan Default, Logistic regression, Random Forest

INTRODUCTION

In recent times, the applications of statistics have been relied on increasingly to tackle several problems in the business sector. As the need to make critical decisions arises amid uncertainty, predictive analyses have been applied to predict and influence future occurrences, identify trends for more accurate planning, forecasting, and decision-making. More than ever, this branch of statistics has influenced several areas of the economy across the world. The banking sector is no exception, and the increasing requirement for financial institutions to have robust risk management has led to an interest in developing current methods of risk estimation (Daria and Johan, 2019). A key area where predictive analysis has been used in the modern world is in the detection of financial risk, such as fraudulent detection and loan default prediction.

The main importance of the financial institutions, particularly banks are to safeguard the money kept by their clients and make it accessible when need arises. They also advance loans to their customers (Kugiel & Jakobsen, 2009). The loan portfolio is typically the largest asset and the predominant source of income for banks (Aballey, 2009). In spite of the huge income generated from their loan portfolio, available literature shows that huge portions of bank loans are usually defaulted and therefore affect the financial performance of these institutions.

Loan defaults are on the rise, and this is a critical source of economic strain. For this reason, these defaults must be controlled and monitored (Daria and Johan, 2013). There has been a growing concern about the relative regression in loan performance in commercial banks.

Several studies have revealed that credit risk decisions are key determinants for the success of financial institutions because of huge losses that may result from wrong decisions. In fact, the issue of loan delinquency/default among banks and Microfinance Institutions has been discussed in many public lectures and forums as one of the reasons why commercial banks have not shown much interest in financing Micro, Small and Medium Enterprises (MSMEs) (Addae-Korankye, 2014). According to (KPMG, 2002), a high level of Non-Performing loans in a bank portfolio of investments affects the level of interest income.

The best method of solving this problem would be to create loan default models. In Kenya, several predictive models have been used to predict loan defaults (Evusa et al, 2015). These models include linear discriminant analysis, logistic regression models and generalized extreme value regression models. A technique that is widely used for estimating the probability of client default is Logistic Regression (Thomas et al, 2012). These models are parametric since they assume the response being investigated takes a particular functional form. However, in this study, we observe how tree models--which are non-parametric and non-linear in nature--can be used to solve a credit card default problem and compare its result to that of the traditional statistical technique.

A loan default occurs when the borrower does not make required payments or in some other way does not comply with the terms of a loan. (Murray, 2011), Failure to do something required under legislation or as ordered by a court, such as not making an appearance to answer charges. Default is the inability to repay the loan by either failing to complete the loan as per the loan agreement or neglect to service the loan. (Wakuloba, 2003). Consultative Group to Assist the Poor (CGAP) 2009 also defined loan default as when a borrower cannot or will not repay a loan, and the Microfinance Institution (MFI) no longer expects to be repaid (although it keeps trying to collect). Moreover, Pearson and Greeff (2006) defined default as a risk threshold that describes the point in the borrower's repayment history where he or she missed at least 3 installments within 24 months. This represents a point in time and indicator of behaviour, wherein there is a demonstrable increase in the risk that the borrower eventually will truly default, by increasing all repayments. The definition is consistent with international standards, and was necessary because consistent analysis required a common definition. This definition does not mean that the borrower had entirely stopped paying the loan and therefore been referred to collection or legal processes; or from an accounting perspective that the loan had been classified as bad or doubtful, or actually written off. Loan default can be defined as the inability of a borrower to fulfill his or her loan obligation as at when due (Balogun & Alimi, 1990).

MATERIALS AND METHODS

Logistic Regression

Consider a model with two predictors X_1 and X_2, and one binary (Bernoulli) response variable Y, which can be denoted by P = P(Y = 1). We assume a linear relationship between the predictor variables and the log-odds of the event that Y=1. This linear relationship can be written in the following mathematical form

$$\text{l}= log_b \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (1)$$

By simple algebraic manipulation, the probability that Y=1 is

$$P = \frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}} \quad (2)$$

Classification Tree

Decision tree learning is one of the predictive modeling approaches used in statistics, data mining, and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels (Wikipedia, n.d.).

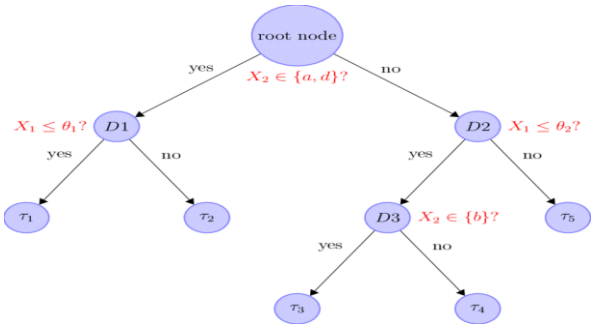


Figure 1: Simple Mathematical Illustration of a Decision Tree

Two main Types of Decision Trees

Classification Tree Analysis

This is when the predicted outcome is the class (discrete) to which the data belongs. (E.g., predict who wins an election, predict whether a patient is diabetic or not).

Regression Tree Analysis

It is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

Some techniques, often called ensemble methods, construct more than one decision tree:

Boosted trees incrementally build an ensemble by training each new instance to emphasize the training instances previously mis-modeled. Examples are: AdaBoost, XGBoost, LightGBM. These can be used for regression-type and classification-type problems.

A bootstrap aggregated (or bagged) decision tree, an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting on the trees for a consensus prediction.

A random forest classifier is a specific type of bootstrap aggregating

Random Forest

Random forest is an ensemble-based learning algorithm, which comprises of n collections of decorrelated decision trees (Hastie et al., 2009). It is built on the idea of bootstrap aggregation, which is a method for resampling with replacement in order to reduce variance. Random Forest uses multiple trees to average (regression) or compute majority votes (classification) in the terminal leaf nodes when making a prediction. Built on the idea of decision trees, random forest models have resulted in significant improvements in prediction accuracy as compared to a single tree by growing 'n' number of trees; each tree in the training set is sampled randomly without replacement (Breiman, 2001) Decision trees consist simply of a tree-like structure where the top node is considered the root of the tree that is recursively split at a series of decision nodes from the root until the terminal node or decision node is reached.

The decision tree algorithm is a top-down "greedy" approach that partitions the dataset into smaller subsets. Greedy algorithms are those that take the simplest solution rather than the most optimal solution, which is often more complex. The top of the decision tree is known as the root node and this corresponds to the best predictor variable. At each decision node, the features are split into two branches, and this process is repeated until the leaf nodes are reached, which are used to make the final prediction.

In a classification problem, there are three major methods of splitting features at each node: the Gini Impurity, Information Gain, and the Chi-square method.

Advantages of using a tree-like learning algorithm allow for training models on large datasets in addition to quantitative and qualitative input variables. Additionally, tree-based models can be immune to redundant variables or variables with high correlation, which may lead to overfitting in other learning algorithms. Trees also have very few parameters to tune when training the model and perform relatively well with outliers or missing values in a dataset. However, trees are prone to poor prediction performance; decision trees themselves are prone to overfitting noise in a training set, which ultimately leads to results with high variance. In other words, this means the model could accurately predict the same data it was trained on but may not possess the same performance on datasets without similar patterns and variations in the training set. Even fully-grown decision trees are notorious for overfitting and do not generalize well to unseen data; random forest solves the overfitting conundrum by using a combination or "ensemble" of decision trees where the values in the tree are a random, independent sample. Randomly sampling with replacement is known as bagging, and this results in a different tree being generated to train on; averaging the results from the 'n' number of trees will result in decreasing the variance and establishing a smoother decision boundary (Hastie, Tibshirani, & Friedman, 2009).

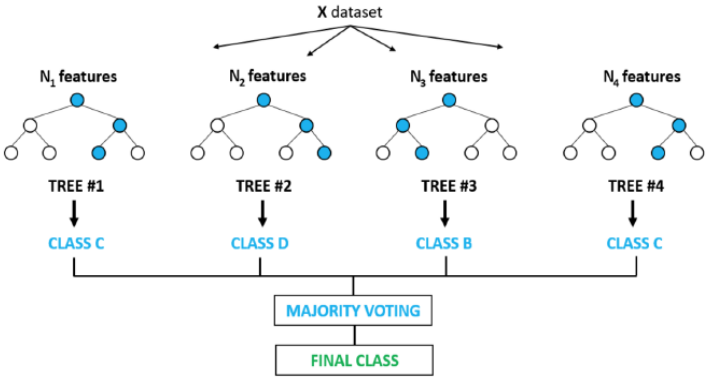


Figure 2: Random Forest Illustration

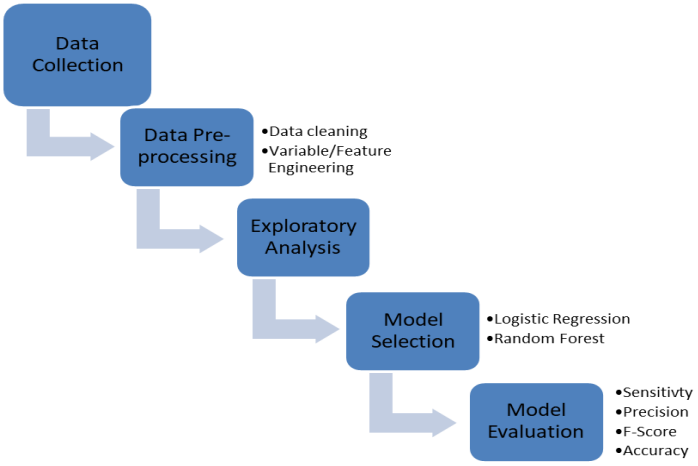


Figure 3: Workflow of the Modeling Process

Model Description

Logistic Regression

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent (explanatory) variables. The name logistic regression is used when the dependent variable has only two values, such as 0 and 1 or Yes and No. The name multinomial logistic regression is usually reserved for the case when the dependent variable has three or more unique values.

In logistic regression, a mathematical model of a set of explanatory variables is used to predict a logit transformation of the dependent variable. In logistic regression, a mathematical model of a set of explanatory variables is used to predict a logit transformation of the dependent variable.

$$l = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_p X_p \quad i = 1,2, \ldots p \qquad (3)$$

l, The log odds is the linear combination of the independent variables.

If p is the proportion of observations with an outcome of 1, then 1-p is the probability of an outcome of 0. The ratio $\frac{p}{1-p}$ is called the odds, and the logit (l) is the logarithm of the odds, or just log odds. Mathematically, the logit transformation is written (4)

$$l = logit(p) = \log\left(\frac{p}{1-p}\right) \qquad (4)$$

For a binary response, the logistic regression model can be expressed by summing over the linear combinations of input features and a corresponding weight (w) plus a bias term (b) for each instance, as shown in equations (5) and (6)

$$p\left(y^{(i)} = 1 \middle| x^{(i)}, w\right) = 1 - \frac{1}{1+e^{(w^T x^{(i)}+b)}} \qquad (5)$$

$$p\left(y^{(i)} = 0 \middle| x^{(i)}, w\right) = 1 - \frac{1}{1+e^{(w^T x^{(i)}+b)}} \qquad (6)$$

Therefore, the logistic regression equation is given as (7);

$$\pi(z) = \left[\frac{1}{1+e^{-z}}\right] \qquad (7)$$

Where,

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (8)$$

The function $\pi(z) = \left[\frac{1}{1+e^{-z}}\right]$ is called the logistic function (or sigmoid function); it looks like this:

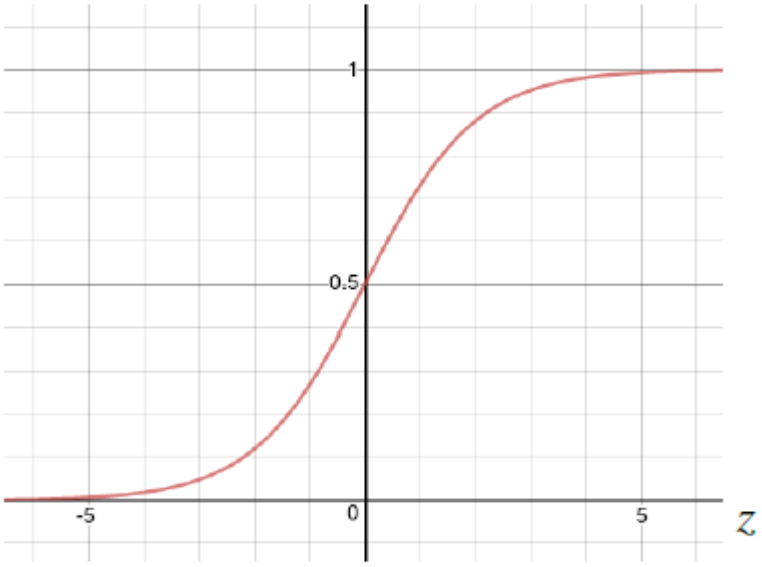


Figure 4: Logistic Regression (Sigmoid) Function

Random Forest

Random forest is an ensemble-based learning algorithm which comprises of n collections of de-correlated decision trees it is built on the idea of bootstrap aggregation, which is a method for resampling with replacement in order to reduce variance Random forest is an ensemble-based learning algorithm which comprises of n collections of de-correlated decision trees. It is built on the idea of bootstrap aggregation, which is a method for resampling with replacement in order to reduce variance.

A decision tree is the building block of a random forest and is an intuitive model. Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. The condition, or test, is represented as the "leaf" (node), and the possible outcomes as "branches" (edges). This splitting process continues until no further gain can be made or a preset rule is met, e.g., the maximum depth of the tree is reached.

For categorical variables, there are two major algorithms for selecting the best split for a classification problem: the Gini Impurity and Cross-entropy method. The Gini impurity splitting criterion is adopted in this research work. The Gini Impurity of a dataset is a number between 0 and 0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

Data Description and Source

Data Description

The data used for this research contains information on default payments, demographic factors, and credit data of German credit card client. This dataset classifies people described by a set of attributes as good or bad credit risks.

Data Source

The data was sourced from the University of California, Irvine (UCI) machine learning repository (hhtps://archive.ics.uci.edu), a top institute in the United States committed to research. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. The repository is supported by the National Science Foundations, United State. The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 to promote the progress of science; to advance the national health, prosperity, and welfare; to secure the national defense.

Table 1: Meta-Data

Data Set Characteristics:	**Multivariate**
Variable Characteristics:	Categorical, Integer
Sample Size:	1000
Number of Independent Variable:	12
Number of Dependent Variable:	1
Task	Classification

Model Performance Evaluation

i. ***Confusion matrix:*** (also known as an error matrix) refers to the table of result of the correctness of classification model. It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions. It contains the "true positive", "false positive", "true negative", and "false negative" "true positive", "false positive", "true negative", and "false negative" which form the basis of the most common performance measures of classification algorithms such as confusion matrix, accuracy, sensitivity, precision and specificity.

Table 2: Confusion Matrix Table

Matrix	Predicted Positive (PP)	Predicted Negative (PN)	Formula
Actual Positive (P)	True Positive (TP)	False Negative (FN)	$P = TP + FN$
Actual Negative (N)	False Positive (FP)	True Negative (TN)	$N = FP + TN$
	$PP = TP + FP$	$PN = FN + TN$	$Total = P + N$

ii. ***Accuracy:*** Accuracy measures how often the classifier makes the correct prediction. It is the ratio between the number of correct predictions and the total number of predictions. Accuracy will answer the question: what percentage of the models predictions were correct?

$$Accuracy = \frac{TP+TN}{P+N} \quad (9)$$

iii. ***Sensitivity (Recall or True Positive Rate)****:* measures the proportion of positives that are correctly identified. In this case, it is the proportion of bad creditors a model is able to predict correctly.

$$Sensitivity = \frac{TP}{TP+FN} = \frac{TP}{P} \quad (10)$$

iv. ***Specificity (True Negative Rate)****:* measures the proportion of negatives that are correctly identified. In this case, it is the proportion of good creditors that a model is able to predict correctly.

$$Specificity = \frac{TN}{TN+FP} = \frac{TN}{N} \quad (11)$$

v. ***Precision:*** it is the number of positive predictions divided by the total number of positive class values predicted. It is also called the Positive Predictive Value (PPV).

$$Precision = \frac{TP}{TP+FP} = \frac{TP}{PP} \quad (12)$$

*vi.****F1-score (F-score):*** is the harmonic mean of precision and recall score and is used as a metric in scenarios where choosing either of precision or recall score can result in compromise in terms of the model giving high false positives and false negatives respectively.

$$F1_{score} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (13a)$$

RESULTS AND DISCUSSIONS

Descriptive and Exploratory Analysis

Table 2: Summary Statistics of Quantitative Variables

	Age	Job	Credit amount	Duration	Present residence Since
Count	1000.000	1000.000	1000.000	1000.000	1000.000
Mean	35.564	1.904	3271.258	20.903	2.845
Std	11.375	0.654	2822.737	12.059	1.104
Min	19.000	0.000	250.000	4.000	1.000
25%	27.000	2.000	1365.500	12.000	2.000
50%	33.000	2.000	2319.500	18.000	3.000
75%	42.000	2.000	3972.250	24.000	4.0000
Max	75.000	3.000	18424.000	72.000	4.000

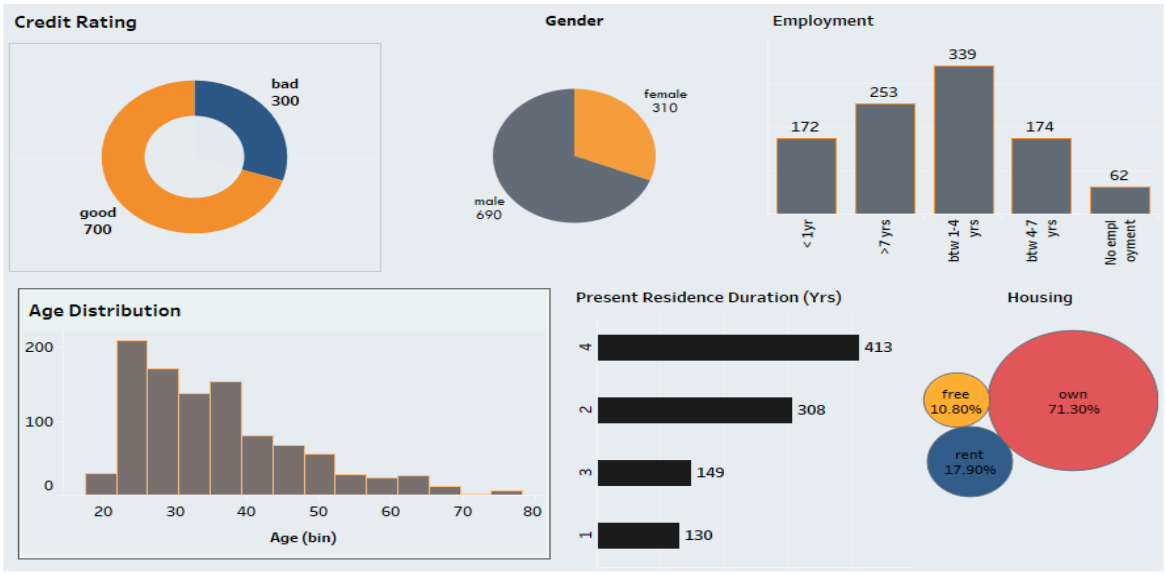


Figure 5: Demographic Characteristics of Credit Customers

The charts above highlight the imbalance in the response variable, with 70% of the customers having a good credit rating. The gender distribution was also highlighted, with 69% male and 31% female. It is revealed from the age distribution chart that the majority of the credit customers are of a younger age, with 75% of the population falling within the age of 42 or less.

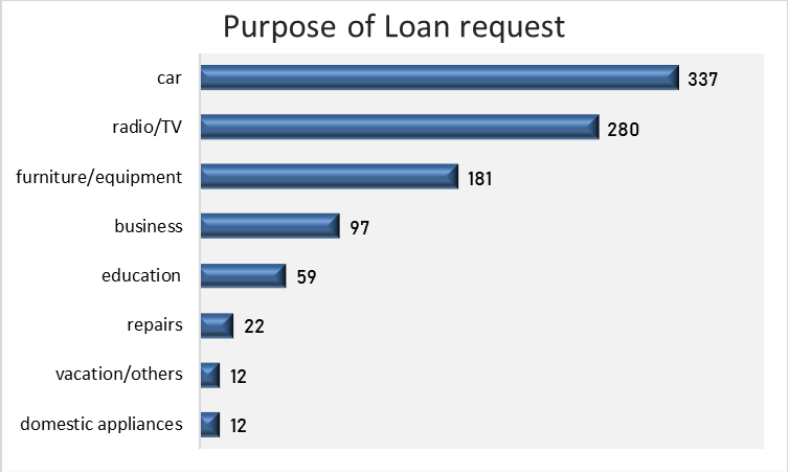


Figure 6: Purpose of Loan Request

More than one-third of the credit customers applied for a credit loan for car purchase. This could serve as a business opportunity for the financial institution as they could sell cars to the customers and give room for installment payment.

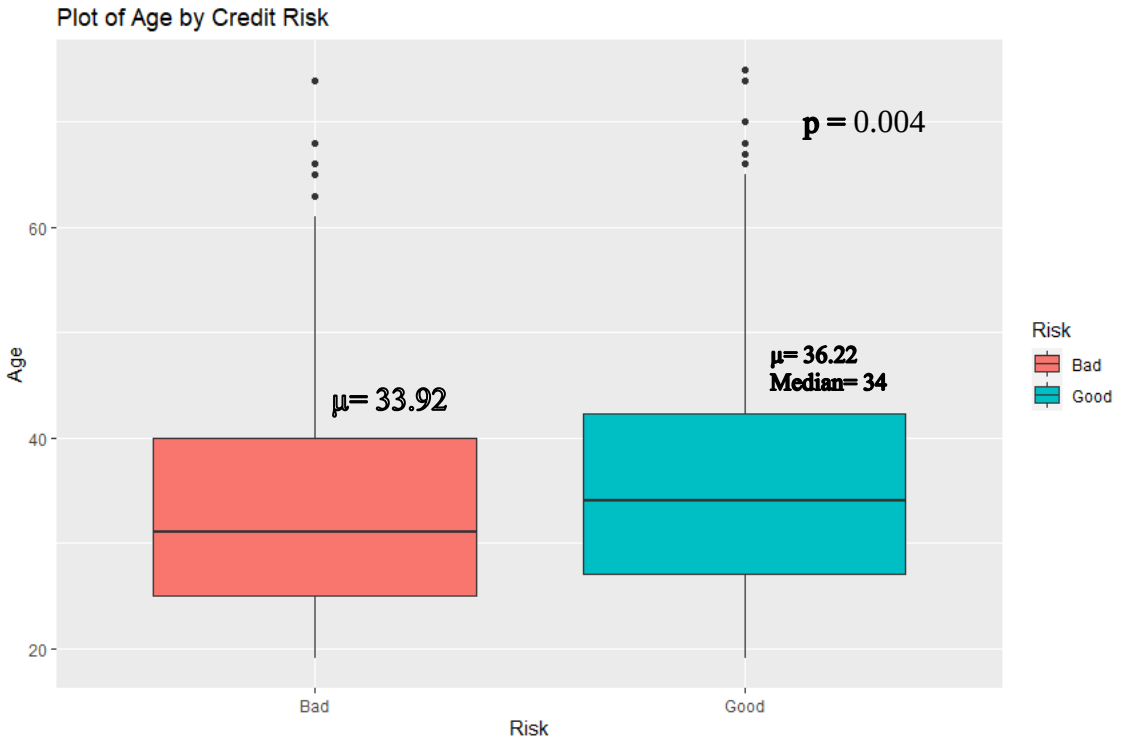


Figure 7: Effect of Independent Variables on The Response Variable (Good or Bad Loan)

The chart above shows that the median value for bad records is slightly lower than the median value for good records. It may be speculative to conclude that young individuals have terrible credit histories, but we may reliably presume it is riskier, as the independent sample Mann Whitney-U test confirms that there is significant difference between the mean ages of both groups.

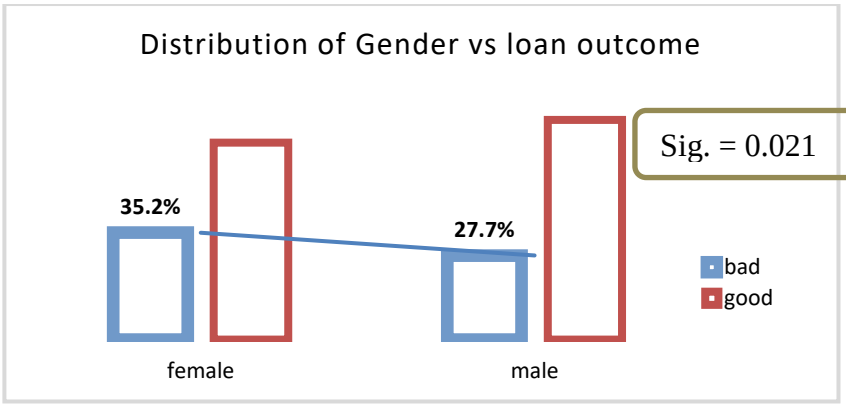


Figure 8: Gender Versus Loan Outcome

From Figure 8 above, the probability of loan default given female $P(bad/gender = female)$ is greater than $P(bad/gender = male)$. This is also revealed in the sex logistic regression coefficient. The chi-square p-value of 0.021< 0.05, the significance level, indicates that the loan default outcome significantly depends on gender. This means that females pose a higher credit risk than males.

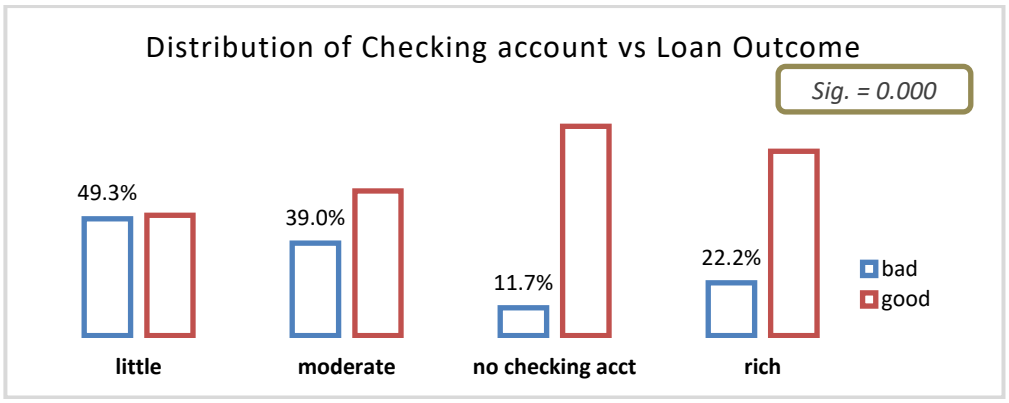


Figure 9: Checking Account Versus Loan Outcome

A checking account is a type of bank account that allows you to easily deposit and withdraw money for daily transactions. Customers with little and moderate credit amount poses more credit risk than the rich checking account. Those with no checking account are the safest. The p-value of 0.000 shows that there is strong dependence of checking account on credit risk.

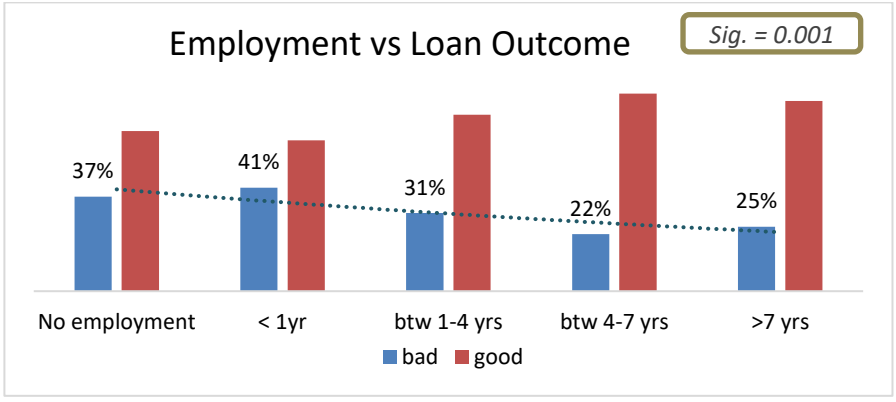


Figure 10: Employment Status Versus Loan Outcome

There is a downward slope of bad credit rating from unemployed customers to employees with over 7 years of service. The new employees have more tendency of defaulting than the old ones. The years of employment have a significant influence on the credit rating of German customers, with a significant value of 0.001< 0.05.

Modeling

The goal is to see how many loan defaulters are predictable by the model. This means that the Positive class is the 'bad' class, which is coded as 1, and the negative class is the 'good' class, which is coded as 0

Logistic Regression Model

Table 3: Variables and their Estimates

Variable	Coefficient	Standard Error	P-Value
Sex	-0.4418	0.186	0.018
Saving accounts(little)	0.4835	0.189	0.011
Employment ($<$ 1year)	0.6711	0.238	0.005
Employment (1-4 years)	0.3337	0.197	0.090
Checking account (rich)	-2.3792	0.440	0.000
Checking account (moderate)	-1.6131	0.301	0.000
Checking account (little)	-1.2801	0.309	0.000
Checking account (others)	-3.2267	0.320	0.000
Duration	0.0438	0.007	0.000

Logistic Regression Performance Measures

The fitted logistic regression model is presented in equation (13b)

$$\log(p/(1-p)) = -0.4418Sex + 0.4835(Savings\ account_little\ ) + 0.3337Employment_(btw1-4yrs) + 0.6711Employment_(< 1yr) - 2.3792Checking\ account_little - \quad (13b)$$

Table 4: Confusion Matrix of Train Set

		Predicted Class		
		Class = 0	**Class = 1**	**Column Total**
Actual Class	Class = 0	489	68	557
	Class = 1	119	124	243
	Row Total	608	192	800

Table 5: Confusion Matrix of Test Set

		Predicted Class		
		Class = 0	Class = 1	Column Total
Actual Class	Class = 0	122	21	143
	Class = 1	27	30	57
	Row Total	149	51	200

Table 6: Train Set Versus Test Set

Variable	Train set	Test set
Sensitivity	0.5103	0.5263
Specificity	0.8779	0.8531
Precision	0.6458	0.5882
Accuracy	0.7663	0.7600
F1 Score	0.5702	0.5556

Tables 4 and 5 show the confusion matrices of the train and test sets produced by the logistic regression model. On the train dataset, the model predicted 489 out of 557 (12.2%) good credit ratings correctly and 124 out of 243 (46.1%) bad credit ratings correctly. While on the test dataset, the model predicted 122 out of 143 (85.3%) good credit ratings correctly and 30 out of 57 (52.63%) bad credit ratings correctly.

Conversely, on the train dataset, 68 out of 557 (36.36% *False Positive Rate*) good credit ratings were wrongly detected as bad, and 119 out of 243 (48.97% *False Negative Rate*) bad credit ratings were wrongly detected as good. While on the test dataset, 21 out of 143 (43.75% *False Positive Rate*) good credit ratings were wrongly detected as bad, and 27 out of 57 (47.37% *False Negative Rate*) bad credit ratings were wrongly detected as good. The model predicted 76.6% and 76% of the training and test data sets accurately.

Random Forest Modeling

The random forest model works by combining several decision trees; each tree is fitted on the bootstrapped dataset (i.e., the dataset is sampled with replacement and each sample is trained by the decision tree model). These trees serve as the estimators of a random forest, the estimates generated is therefore averaged to determine the predicted value; this averaging process is called bagging. The output of the random forest model in this case is generated by bagging a total of 1788 estimators. The maximum depth of the trees is 110; this makes it difficult to visually display any of the trees.

The chart below displays the decision tree of the first estimator pruned at the third layer, i.e. $Maximum\ depth = 3$.

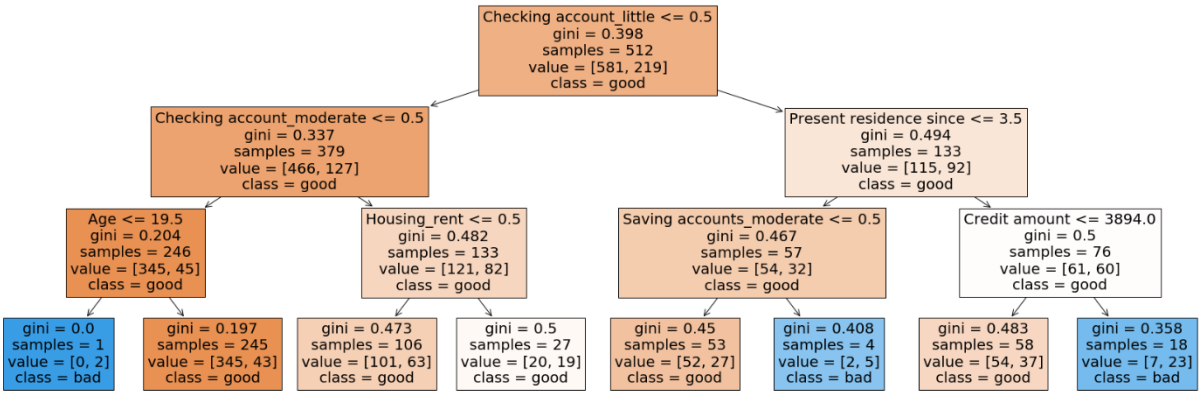


Figure11: Decision Tree of Loan Default With Max_Dept =3

Figure 11 gives an idea about how each decision tree is built. The blue nodes contain classes whose dominant class is "bad loan," while the red nodes are nodes dominated by the "good loan" class. The interpretation is intuitive and probabilistic. For instance, a new record with the following characteristics: Checking account_little <=0.5, Present_residence_since <= 3.5, and Credit amount <= 3894 has a 62% (i.e., (1-0.358)*100) risk of ending up as a bad loan.

The most important values in the tree above are the Gini impurities. Gini impurity is a metric for determining how the factors of a dataset should be split into nodes to form a tree in Decision Trees. It helps determine how pure a node is. Its value ranges between *0 and 0.5*, where 0 means the node is pure and 0.5 implies absolute impurity.

$$Gini\ impurity\ =\ 1 - Gini \quad (14)$$

$$Gini\ =(p^2+q^2) \quad (15)$$

Where,

$p\ is\ the\ probability\ of\ bad\ loan$

$q\ is\ the\ probability\ of\ good\ loan$

The leaf at the bottom left has a Gini impurity of 0 with dominant class = "bad"; this implies that the probability that a customer at that node will default is 1. The fourth leaf has a Gini impurity of 0.5; this implies that a customer who falls at that leaf has a 50% chance of defaulting.

Random Forest Performance Measure

Table 7: Confusion Matrix of Train Set

		Predicted Class		
		Class = 0	Class = 1	Column Total
Actual Class	Class = 0	555	2	557
	Class = 1	43	200	243
	Row Total	598	202	800

Table 8: Confusion Matrix of Test Set

		Predicted Class		
		Class = 0	Class = 1	Column Total
Actual Class	Class = 0	130	13	143
	Class = 1	29	28	57
	Row Total	159	41	200

Table 9: Train Set versus Test Set for Random Forest Modelling

Variable	**Train set**	**Test set**
Sensitivity	0.8230	0.4912
Specificity	0.9964	0.9091
Precision	0.9901	0.6829
Accuracy	0.9438	0.7900
F1 Score	0.8989	0.5713

Tables 7 and 8 show the confusion matrix of the train and test sets produced by the random forest model. On the train dataset, the model predicted 555 out of 557 (99.6% specificity) good credit ratings correctly and 200 out of 243 (82.3% sensitivity) bad credit ratings correctly. On the test dataset, the model predicted 130 out of 143 (90.91% specificity) good credit ratings correctly and 28 out of 57 (49.12% sensitivity) bad credit ratings correctly. Conversely, on the train dataset, 2 out of 557 (0.35% *False Positive Rate*) good credit ratings were wrongly detected as bad, and 43 out of 243 (50.9% *False Negative Rate*) bad credit ratings were wrongly detected as good. While on the test dataset, 13 out of 143 (9% *False Positive Rate*) good credit ratings were wrongly detected as bad, and 29 out of 57 (50.9% *False Negative Rate*) bad credit ratings were wrongly detected as good. The model predicted 94.3% and 79% of the training and test data sets accurately

Feature Importance

Unlike the logistic regression model, the random forest model is a complex non-linear model. It is an ensemble of decision trees; this makes it almost impossible to determine the

coefficient of each variable in reality. However, the importance of each feature to the model can be determined. The chart below displays the key variables that contribute to the training of the random forest model with their corresponding importance score.

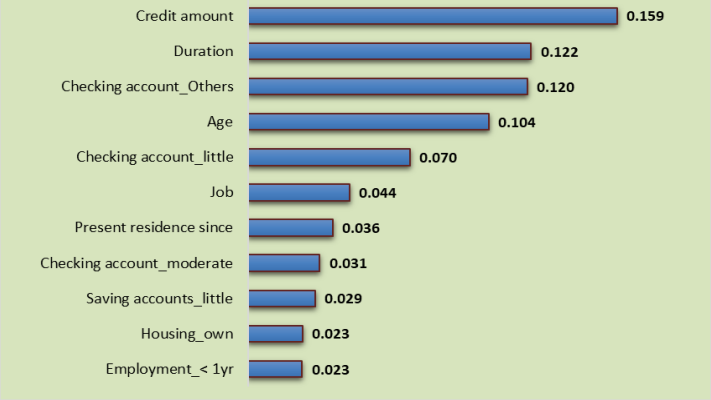


Figure 12: Random Forest Feature Importance

Comparison Between Both Models

Table 10: All Metrics for Train and Test Set

Metrics	**Train set**		**Test set**	
Score Metrics	Logistic Regression	Random Forest	Logistic Regression	Random Forest
Sensitivity	0.5103	0.8230	0.5263	0.4912
Specificity	0.8779	0.9964	0.8531	0.9091
Precision	0.6458	0.9901	0.5882	0.6829
Accuracy	0.7663	0.9438	0.7600	0.7900
F1 Score	0.5702	0.8989	0.5556	0.5713

Table 11: Sample Of 10 Test Data And Their Predicted Values

	Actual_ Y	**Logistic Reg**	**Random Forest**
562	0	0	0
356	0	0	0
915	1	0	1
164	0	0	0
579	0	0	0
890	0	1	0
409	1	0	0
208	0	1	1
938	1	1	1
629	0	0	0

CONCLUSION

The findings of this study reveal that the Random Forest model achieved better predictive performance on the test dataset than Logistic Regression. However, its tendency to overfit the training data suggests high variance and limited generalizability to unseen observations. Conversely, Logistic Regression demonstrated higher sensitivity, indicating a greater capacity to identify loan defaulters. This highlights its potential suitability for credit risk assessment, where accurately identifying potential defaulters is essential.

The results further suggest that Logistic Regression is a practical choice for small datasets, particularly when interpretability, simplicity, and computational efficiency are important. Nevertheless, the comparatively weaker performance of Random Forest may be partly attributable to the limited sample size. With larger datasets and appropriate hyperparameter tuning, tree-based models may better capture complex patterns and nonlinear relationships, potentially outperforming Logistic Regression.

In conclusion, the choice of model for loan default prediction should consider not only predictive performance but also sensitivity, generalizability, interpretability, and dataset size. While Logistic Regression may be preferable when detecting loan defaulters and ensuring model transparency are priorities, an appropriately optimized Random Forest model may provide stronger predictive performance when sufficient training data are available. These findings underscore the importance of selecting models based on the specific objectives and constraints of credit risk assessment.

COMPETING INTEREST

The author declares he has no conflicts of interest.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

Aballey, F. (2009). "Bad Loan Portfolio: A Case Study of Agricultural Development Bank", KNUST

Anioke, C. N. (2012). Loan Granting And Its Recovery Problems On Commercial Banks (A Case Study Of First Bank Plc, Ojo- Alaba Branch).

Baesens, B., Setiono, R., Mues, C., & Vanthienen, J. (2003). Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science, 49*(3), 312-329.

Breiman, L. (2001). Random Forests. *Machine Learning volume, 45*, 5-32.

Daria, G. and Johan, A. (2013). Interest Rates and Default in Unsecured Loan Markets. Quantitative Finance. (1925-1934), 13(12).

Daria, G. and Johan, A. (2019). Loan Default Prediction using Supervised Machine Learning Algorithms. pp.1.

Evusa, Z., Mudaki, J. S., and Ojala, D. O. (2015). Evaluation of the Factors Leading to Loan Default at Equity Bank, Kenya. Journal of Economics and Sustainability. *Vol.6*(ISSN 2224-607X), 8.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society, Series A – Statistics in Society, 160*(3), 523-541.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference and prediction. *Springer*.

Koh, H. C., & Chan, K. L. (2002). Data mining and customer relationship marketing in the banking industry. *Singapore Management Review, 24*(2), 1-27.

KPMG. (2002). An analysis of the impact of external debt on banks' performance. *Evidence from Nigeria. British Journal of Arts and Social Sciences., 12(1)*, 37-51.

Kugiel, L., & Jakobsen, M. (2009). Fund Transfer Pricing in a Commercial Bank. Master's 454 Thesis, MSC in Finance and International Business.

Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245-254.

Liang, D., Tsai, C.-F., Dai, A.-J., & Eberle, W. (2018). A novel classifier ensemble approach for financial distress prediction. *Knowledge and Information Systems, 54.2*, 437-462.

Lu, H., Wang, H., & Yoon, S. W. (2017). Real Time Credit Card Default Classification Using Adaptive Boosting-Based Online Learning Algorithm. *IIE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE)*.

Murray, J. (2011). Default on a loan. *United States Business Law and Taxes Guide*.

Paolo, G. (2001). Bayesian data mining, with application to benchmarking and credit scoring. *Applied Stochastic Models in Business and Society, 17*, 69-81.

Ronaghan, S. (2018). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark.

Sheikh, R. I., William, E., & Sheikh, K. G. (2018). Credit Default Mining Using Combined Machine Learning and Heuristic Approach.

Thomas, L. C. (2000). A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*, 149–172.

Thomas, L. C. Tong E. N., & Mues. C. (2012). "Mixture cure models in credit scoring: If and when borrowers default". *European Journal of Operational Research, 218*, 132-139.

White, A. (2020, October 12). *What are checking accounts*. Retrieved from Select: https://www.cnbc.com/select

Wikipedia. (n.d.). *Decision tree learning*. Retrieved July 10, 2021, from Wikipedia: https://en.wikipedia.org/wiki/

Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473-2480.

APPENDICES

Appendix 1

First 10 Rows of Dataset

SN	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Employment	Present residence since	Foreign worker	Risk
0	67	male	2	own	NA	little	1169	6	radio/TV	>7 yrs	4	YES	good
1	22	female	2	own	little	moderate	5951	48	radio/TV	btw 1-4 yrs	2	YES	bad
2	49	male	1	own	little	NA	2096	12	education	btw 4-7 yrs	3	YES	good
3	45	male	2	free	little	little	7882	42	Furniture/ equipment	btw 4-7 yrs	4	YES	good
4	53	male	2	free	little	little	4870	24	car	btw 1-4 yrs	4	YES	bad
5	35	male	1	free	NA	NA	9055	36	education	btw 1-4 yrs	4	YES	good
6	53	male	2	own	quite rich	NA	2835	24	furniture/equipment	>7 yrs	4	YES	good
7	35	male	3	rent	little	moderate	6948	36	car	btw 1-4 yrs	2	YES	good
8	61	male	1	own	rich	NA	3059	12	radio/TV	btw 4-7 yrs	4	YES	good
9	28	male	3	own	little	moderate	5234	30	car	No employment	2	YES	bad
10	25	female	2	rent	little	moderate	1295	12	car	< 1yr	1	YES	bad

Appendix 2

Variable Description Table

SN	Variable	Data Type	Class
1	Risk Rating (Response Variable)	Categorical	0 = Good 1 = Bad
2	Age	Integer	
3	Sex	Categorical	0 = Female 1 = Male
4	Job	Categorical	1 = unemployed/ unskilled - non-resident 2 = unskilled – resident 3 = skilled employee / official 4 = management/ self-employed/ highly qualified employee
5	Housing	Categorical	1= own 2 = free 3 = rent
6	Saving accounts	Categorical	1 = little 2 = moderate 3 = quite rich 4 = rich
7	Checking account		1 = little 2 = moderate 3 = no checking acct 4 = rich
8	Credit_amount	Integer	
9	Duration	Integer	
10	Purpose	Categorical	1 = business 2 = car 3 = domestic appliances 4 = education 5 = furniture/equipment 6 = radio/TV 7 = repairs 8 = vacation/others
11	Employment	Categorical	1 = '< 1yr' 2 = '>7 yrs' 3 = 'btw 1-4 yrs' 4 = 'btw 4-7 yrs' 5 = 'No employment'
12	Present residence	Integer	
13	Foreign worker	Categorical	0 = NO 1 = YES

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.