Key Determinants of Credit Default in Personal Loans: A Comparative Analysis of Logistic Regression and Random Forest

Yi Shen*

The Palatine Centre, Durham University, Durham, DH1 3LE, UK

*yishen66668@163.com

Abstract. Accurate credit default prediction is a core aspect of consumer credit risk management. This study based on a publicly available personal loan dataset containing 32,581 records, systematically compares the performance of logistic regression and random forest in default prediction through data preprocessing, exploratory data analysis, feature selection, and classification modeling. It focuses on examining the relative contributions of three types of variables—borrower basic information, loan application information, and credit history information, to default risk. The study finds that indicators reflecting repayment burden (loan amount as a percentage of income) are the strongest determinant of default prediction, followed by loan pricing (interest rate, credit rating). Borrower income and housing situation are the second most important dimensions, while the incremental predictive effect of credit history length is limited. Model comparisons show that random forest outperforms logistic regression in accuracy, AUC, and F1 score, indicating that default risk is significantly affected by nonlinearity and interaction effects; however, logistic regression still holds significant value in practice due to its interpretable coefficients. The study strengthens the internal validity of the conclusions through multiple robustness tests, including threshold adjustment, multicollinearity diagnosis, permutation importance, and SHAP interpretation. This research provides empirical evidence for financial institutions to optimize pre-loan risk assessment and variable selection.

Keywords: Credit default prediction; Logistic regression; Random forest; Feature importance; Personal loans

1 Introduction

1.1 Background

Accurate credit default forecasting is a fundamental component of consumer credit risk management. Loan decision-making depends critically on how well financial institutions distinguish between low-risk and high-risk borrowers. In consumer loans, inaccurate default predictions can directly increase financial losses, reduce portfolio quality, and lead to inefficient credit allocation. As consumer borrowing continues to expand,

© The Author(s) 2026
V. Erokhin et al. (eds.), *Proceedings of the 6th International Conference on Internet Finance and Digital Economy (ICIFDE 2026)*, Advances in Economics, Business and Management Research 412,
https://doi.org/10.2991/978-94-6239-830-6_2

the demand for reliable default forecasting models has become increasingly important in both academic research and practical risk management.

1.2 Brief Literature Review

Research on credit default prediction has evolved along two distinct trajectories. The first is the gradual expansion of the predictor set. The second is the comparison between traditional statistical models and newer machine learning methods. In a systematic review of 108 studies, Çallı and Coşkun (2021) show that early work mainly relied on demographic, socio-economic, and financial variables, while later studies increasingly incorporated behavioural, situational, and alternative forms of information. They also noted a broader shift toward non-traditional data sources and newer analytical techniques, indicating that credit risk is no longer considered a problem explained solely by financial variables.

The second strand of literature focuses more directly on model performance. Bhandary and Ghosh (2025) use a Taiwan credit card default dataset to compare linear discriminant analysis, logistic regression, support vector machines, XGBoost, random forest, and deep neural networks. Their results show that modern machine learning methods outperform traditional statistical methods in terms of F1 score, G-mean, and AUC, although logistic regression remains competitive and retains the advantage of explainability.

Aboagye et al. (2022) reached a relevant conclusion but emphasised the importance of predictors. Comparing ten data mining techniques, they found that both support vector machines and logistic regression performed strongly, and also showed that some variables were much more important than others. Their research demonstrates that using only the most important predictors can produce results close to those of a more complete model, and logistic regression maintained the same error rate and AUC before and after variable selection. This finding is particularly relevant to the present study, which aims not only to predict defaults but also to identify the key factors behind the prediction.

Another important issue in the literature is class imbalance. Alam et al. (2020) showed that classifier performance improved significantly after data rebalancing, and reported that oversampling methods generally outperformed undersampling methods. They found that K-means SMOTE combined with GBDT provided the best accuracy for the Taiwan dataset. Their findings emphasise the importance of attending not only to model selection but also to the structure of the data itself.

In addition to traditional statistical models and conventional machine learning methods, deep learning (DL) techniques have also been increasingly applied to credit default prediction. Wahab et al. (2024) conducted a comparative study using an artificial neural network (ANN), AdaBoost, and decision tree on a credit card default dataset from the UCI ML repository. Their results showed that AdaBoost and decision tree achieved the highest accuracy of 82%, outperforming the ANN model which reached 78% accuracy. More importantly, their study identified that hyperparameter tuning, including the number of hidden layers, choice of optimizer (Adam, SGD, RMSProp, Adamax), and activation functions (ReLU, Sigmoid, Tanh, ELU), which directly influenced predictive

performance. This finding highlights that while deep learning offers flexibility, its effectiveness in credit default prediction depends critically on appropriate model configuration.

1.3 Research Gap and Research Question

Despite the significant growth in the literature on credit default prediction, two issues remain important. First, many studies emphasise model comparison rather than the relative contribution of different categories of predictors. In other words, although it is often clear which model performs better, it is less clear whether the borrower's basic information, loan application information, or credit history information contributes most to default prediction. Second, some studies focus primarily on predictive accuracy, which can reduce attention to explainability. However, in real-world lending contexts, understanding why a borrower is classified as risky is equally important.

This study addresses these problems by focusing not only on predictive performance but also on identifying the key determinants of default. More specifically, it addresses the following research questions:

1) Which factors are most effective in predicting default risk in personal loan applications?

2) How do different dimensions of borrower information (basic information, loan application information, and credit history information) contribute to default prediction?

1.4 Study Objective and Contribution

The objective of this study is to examine the predictive role of borrower basic information, loan application information, and credit history information in a structured credit risk dataset. To achieve this goal, the study combines data preprocessing, exploratory data analysis, feature selection, and classification modelling. Logistic regression is used as the primary model due to its interpretability, while random forest is used as a benchmark model to test whether more flexible methods can improve predictive performance.

The contribution of this research is twofold. First, it provides evidence on which specific variables are most closely associated with default risk. Second, it compares the relative importance of three types of predictors rather than focusing only on model performance. In this way, the study aims to provide a better understanding of the most important factors in credit default prediction while retaining practical relevance for risk assessment.

In addition to comparing predictive performance and variable correlation, the empirical design also strengthens internal robustness through threshold adjustment, calibration analysis, multi-collinear diagnosis, robustness check that excludes relevant predictors, importance based on arrangement, SHAP-based interpretation and five-fold cross-validation. These extensions are designed to provide a more reliable basis for interpreting model performance and predictor importance in the absence of an independent external verification dataset.

2 Dataset

2.1 Data Source

This study uses the public credit risk data set obtained from the Kaggle platform. The data set contains structured information about individual loan applicants and their loan characteristics, making it suitable for checking default forecasts in the consumer loan environment. The binary target variable is loan_status, in which 0 means non-default and 1 means default. The original data set contains 32,581 loan records.

There are two reasons for choosing a data set. First of all, it provides a large enough sample for classification modelling and comparative analysis. Secondly, it includes variables that can be naturally divided into basic information of borrowers, loan application information and credit record information, which is closely related to the research objectives of comparing the predicted contributions of different information dimensions.

At the same time, the use of a single public data set imposes an important boundary on reasoning. Although the data set is suitable for structured empirical comparison, the publicly available metadata does not fully record the original sampling mechanism, geographical scope or collection period. Therefore, these findings should be interpreted as evidence of this structured benchmark setting, rather than the result that can be automatically promoted to all consumer loan markets. For this reason, the analysis complements the main results through extensive internal robustness checks, while recognising that independent external verification is still not possible in this study.

2.2 Sample and Variables

The dataset used in this study contains a total of 12 variables, including 11 explanatory variables and 1 binary target variable. The explanatory variables can be divided into three categories.

Category 1: Borrower Basic Information, including person_age, person_income, person_home_ownership, and person_emp_length. These variables describe the borrower's demographic and financial background.

Category 2: Loan Application Information, including loan_intent, loan_grade, loan_amnt, loan_int_rate, and loan_percent_income. These variables reflect the characteristics of the loan itself, including its purpose, price, assigned risk grade, and repayment burden relative to income.

Category 3: Credit History Information, including cb_person_default_on_file and cb_person_cred_hist_length, which capture prior default record and the length of the borrower's credit history.

The target variable loan_status identifies whether the borrower is in breach of contract. Since the main purpose of the study is to identify the key predictors of default, these three sets of explanatory variables constitute the basis for future feature selection and model comparison.

2.3 Relevance to the Research Question

The structure of the dataset is closely aligned with the research focus of this paper. Rather than using only one type of borrower information, the dataset allows comparison of the predictive value of three different dimensions: borrower basic information, loan application information, and credit history information. This is important because the study aims not only to build an accurate classifier but also to determine which type of information contributes most to default prediction.

Additionally, the dataset contains both numerical and categorical variables, enabling the analysis to combine descriptive statistics, feature screening, and classification modelling within a coherent framework. This makes the dataset suitable for both predictive analysis and interpretation, which is a key requirement of this study.

3 Methodology

This study follows the general logic of the CRISP-DM framework, which organises the empirical process into data understanding, data preparation, modelling, evaluation, and interpretation. Consistent with the research focus, the methodology aims not only to evaluate predictive performance but also to identify which variables and which broader information categories contribute most to credit default prediction.

3.1 Data Preprocessing

Before the model is estimated, in order to improve the data quality and reduce the influence of implausible observations, the original dataset was cleaned up. After the project design, two main data issues were addressed: implausible values and missing values.

Outlier treatment was applied to two variables. Borrower age (person_age) was restricted to the range of 18 to 70 years, while employment length (person_emp_length) was restricted to 0 to 47 years. These rules were used to remove unrealistic values that could distort both descriptive analysis and model estimation. After this step, the sample size declined slightly from 32,581 observations to 32,566, indicating that only a very small number of records were excluded.

Missing values were then imputed. Missing observations in person_emp_length were replaced using the rule age – 23, following the project guideline. If this calculation produced a negative value, the result was reset to zero. Missing values in loan_int_rate were filled with the mean of the cleaned sample. After these procedures, no missing values remained in the dataset.

The preparation of classification and numerical variables is carried out in the modelling stage. Categorical variables are encoded using one-hot encoding, while numerical variables are standardised where the model is required. In particular, Z-score standardisation is applied to the logistic regression pipeline so that the variables measured on different scales can be used consistently.

3.2 Exploratory Data Analysis

Exploratory data analysis (EDA) was carried out to understand the structure of the cleaned dataset and identify the preliminary patterns related to the default.

Step 1: Univariate analysis was performed on the main numerical variables, including borrower age, income, employment length, loan amount, loan interest rate, loan-to-income ratio, and credit history length. This allowed inspection of each variable's distribution and identification of general characteristics such as skewness and variability. Several financial variables, particularly income, loan amount, and loan burden relative to income, showed right-skewed distributions, which is common in consumer credit data.

Step 2: Group comparison involved comparing the means of key variables between the default and non-default groups. This comparison provided an initial understanding of differences between borrowers who default and those who do not. Descriptive results showed that defaulting borrowers tend to have lower income, shorter employment history, larger loans, higher interest rates, and heavier loan burdens relative to income.

Step 3: Correlation analysis generated a correlation heatmap for numerical variables to detect preliminary relationships with the target variable and identify possible multicollinearity among predictors. This stage served as a preliminary screening exercise before the more formal feature selection procedure.

3.3 Feature Selection

Feature selection is carried out using a combination of statistical filtering and actual interpretation. The purpose is to identify the variables that contribute most obviously to the default classification and check the redundant information at the same time.

The first step is single-variable filtering, based on the relationship between each predictor and the target variable (loan_status). Following the project design, the ANOVA F-value was used to evaluate the distinguishing power of each feature. This allows variables to be ranked according to the individual contributions of distinguishing between default and non-default observations. The strongest variables at this stage are mainly related to the loan burden, loan level and interest rate, followed by the financial variables of some borrowers and selected credit record indicators.

The second step is multicollinearity screening. Pairwise correlations among numerical variables were examined to determine whether variables with correlation coefficients above 0.80 might be redundant. The strongest overlap was found between person_age and cb_person_cred_hist_length, which is economically intuitive, because older borrowers usually have a longer credit history. This relationship was noted in the explanation, but the two variables were retained in the main analysis because they captured the relevant but not the same aspects of the borrower's personal data.

In this study, the feature screening stage is used as a preliminary rather than final evaluation of the correlation of predictors. The F value of variance analysis provides a single-variable ranking of distinguishing ability, and the pairing correlation is used to identify potential overlap between numerical variables. Since these two methods are still limited in multivariate and nonlinear settings, the analysis does not rely only on

them. On the contrary, these initial results are explained together with multivariate coefficients, model-specific importance, arrangement-based importance, SHAP analysis, polycolinear diagnosis and robustness checks reported later in this article.

3.4 Model Development

In order to answer the research questions, the study developed a main model and a benchmark model.

Primary model: Logistic regression was chosen because of its interpretability and long-standing use in credit scoring research. Since one of the main goals of this paper is to understand the direction and relative importance of predictors, a model with explainable coefficients is particularly useful.

Benchmark model: Random forest was included because it can capture non-linear relationships and interactions among variables more effectively than linear classifiers. It also provides an alternative measure of feature importance, making it suitable for comparison with logistic regression.

The cleaned dataset was divided into training and test subsets using stratified 80:20 split to preserve the class distribution of the target variable in both samples. Based on the final cleaned dataset of 32,566 observations, this resulted in 26,052 observations in the training set and 6,514 in the test set. Logistic regression was estimated using one-hot encoded categorical variables and standardised numerical variables. Random forest used the same train-test split and the same encoded predictors, but no numerical scaling was required.

3.5 Model Evaluation

A broader set of classification and interpretation criteria is used to evaluate the performance of the model. Since default forecasting is an unbalanced classification problem, in which the cost of missing the actual defaulter may be much higher than the cost of labelling safe borrowers, the overall accuracy is reported, but it is not considered sufficient in itself.

First, use ROC AUC and Precision-Recall AUC to evaluate the discrimination performance. ROC AUC captures the ranking ability independent of the threshold, while Precision-Recall AUC is particularly informative in a minority environment, because it focusses more directly on the trade-off between identifying actual defaulters and avoiding false alerts.

Secondly, both models report threshold-dependent indicators, including accuracy, recall, F1 score, specificity, equilibrium accuracy and confusion matrix. These indicators are used to evaluate the effectiveness of each model in identifying default cases, not just to achieve high overall accuracy through multiple class predictions.

Third, the threshold adjustment is carried out by scanning the replacement classification cut-off value and determining the threshold to maximise the F1 score in the retention test set. This allows the analysis to check whether the traditional 0.5 decision threshold is suitable for credit risk screening.

Fourth, evaluate the probability quality through Brier score and calibration curve. These measures help to evaluate whether the default probability of the forecast is consistent with the default frequency observed, not just ranking the observations.

Fifth, the predictor interpretation is based on a variety of complementary methods. In logical regression, coefficient symbols and sizes are used for conditional interpretation. In the random forest model, the characteristic importance based on impurities is reported for baseline comparison, but it is not considered sufficient in itself. In order to enhance interpretability, the analysis also reports the importance of arrangement and global interpretation based on SHAP.

Finally, the coefficient of variance expansion (VIF) is used to check the numerical polycolinearity, and the robustness is checked by excluding high correlation predictors. In addition, the five-fold cross-validation is used to assess whether the main performance mode remains stable in the repeated train test partition. These evaluation procedures are designed to enhance internal robustness, although there is no independent external verification dataset in this study.

3.6 Interpretation Strategy

The final stage of the methodology focuses on interpretation. Since the purpose of this paper is not only predictive but also explanatory, the model outputs are interpreted in two ways.

First, attention is paid to whether the same variables emerge as important across different stages of analysis, including EDA, feature selection, logistic regression, and random forest. This helps identify robust predictors rather than relying on a single model result.

Second, predictors are interpreted according to their broader categories. All explanatory variables are divided into three dimensions: borrower basic information, loan application information, and credit history information. This allows the study to examine not only which individual variables are the strongest predictors but also which types of information contribute most to overall default classification.

In this way, the methodological framework is directly aligned with the research questions, which concern both the key determinants of credit default and the relative contribution of different types of borrower information.

4 Results and Discussion

4.1 Descriptive Results

After preprocessing, the final sample contains 32,566 observations, while the original dataset contains 32,581. This means that only a few records were deleted during outlier processing, so the overall structure of the dataset was preserved. In addition, no missing values remained after imputation.

Class distribution shows that the dataset is imbalanced, but not severely so. The non-default group (loan_status = 0) contains 25,459 observations, accounting for 78.18% of

the sample, while the default group (loan_status = 1) contains 7,107 observations, accounting for 21.82%. This shows that default is a minority, which is typical in credit risk analysis.

Descriptive comparison between the two groups revealed several meaningful differences, as shown in Table 1, Figure 1 and Figure 2.

Table 1. Descriptive Statistics by Default Status.

	loan_status	person_age	person_income	person_emp_length	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length
1	0	27.77	70558.05	4.95	9237.36	10.49	0.15	5.83
2	1	27.48	49124.26	4.15	10847.10	12.87	0.25	5.69

The age of non-default borrowers was slightly older (27.77 years on average) compared to defaulting borrowers (27.48 years). The income gap was the largest: the average income of the non-default group was 70,558.05, while that of the default group was 49,124.26. This indicates that income is closely related to repayment capacity.

Employment history also differed between the two groups. The average employment length for non-default borrowers was 4.95 years, compared to 4.15 years for defaulting borrowers. At the same time, defaulting borrowers took out larger loans. The average loan amount for the non-default group was 9,237.36, while for the default group it was 10,847.10. Loan pricing also showed a large difference. The average interest rate for non-defaults was 10.49, and for defaulters it was 12.87.

The most striking comparison appeared in loan_percent_income: the non-default group averaged 0.15, while the default group averaged 0.25. This shows that the repayment burden relative to income is one of the most important descriptive differences in the sample.

Credit history length was also slightly lower among defaulting borrowers. The average credit history length for the non-default group was 5.83 years, compared to 5.69 years for the default group. Although this gap is smaller than those for income, interest rates, and loan burden, it is still consistent with the view that more mature borrowers are slightly less likely to default.

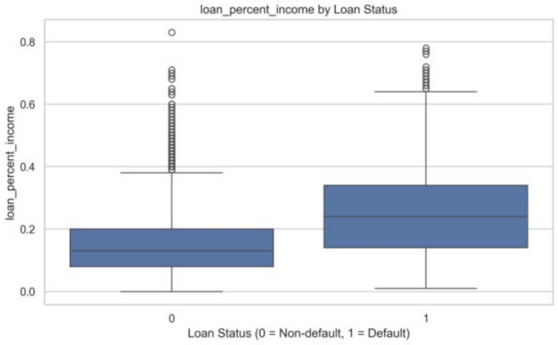


Fig. 1. Boxplot of Loan Percent Income by Default Status.

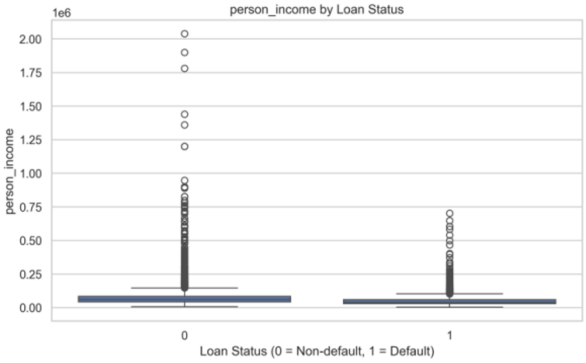


Fig. 2. Borrower Income by Default Status.

In summary, the descriptive results indicate that defaulting borrowers tend to have lower income, shorter employment history, larger loans, higher interest rates, heavier loan burdens relative to income, and slightly shorter credit histories. Because the default class accounts for 21.82% of the final sample, later model evaluation places particular emphasis on minority-class performance rather than relying only on overall accuracy.

4.2 Correlation and Feature Screening Results

Correlation analysis reinforced the descriptive findings. Table 2 presents the correlation matrix of numerical variables.

Table 2. Correlation Matrix of Numerical Variables.

	per-son_a ge	per-son_in-come	per-son_emp_l ength	loan_ amnt	loan_in t_rate	loan_per-cent_income	cb_per-son_cred_hist_l ength	loan_s tatus
person_age	1.00	0.14	0.20	0.05	0.01	-0.04	0.88	-0.02
person_income	0.14	1.00	0.15	0.32	0.00	-0.29	0.12	-0.17
per-son_emp_length	0.20	0.15	1.00	0.11	-0.05	-0.06	0.17	-0.08
loan_amnt	0.05	0.32	0.11	1.00	0.14	0.57	0.04	0.11
loan_int_rate	0.01	0.00	-0.05	0.14	1.00	0.11	0.02	0.32
loan_per-cent_income	-0.04	-0.29	-0.06	0.57	0.11	1.00	-0.03	0.38
cb_per-son_cred_hist_l ength	0.88	0.12	0.17	0.04	0.02	-0.03	1.00	-0.01
loan_status	-0.02	-0.17	-0.08	0.11	0.32	0.38	-0.01	1.00

Among the numerical predictors in Figure 3, the strongest positive correlation with loan_status was loan_percent_income, with a coefficient of 0.38. The second strongest positive relationship was loan_int_rate, with a coefficient of 0.32. These results indicate that repayment burden and loan pricing are closely associated with default risk.

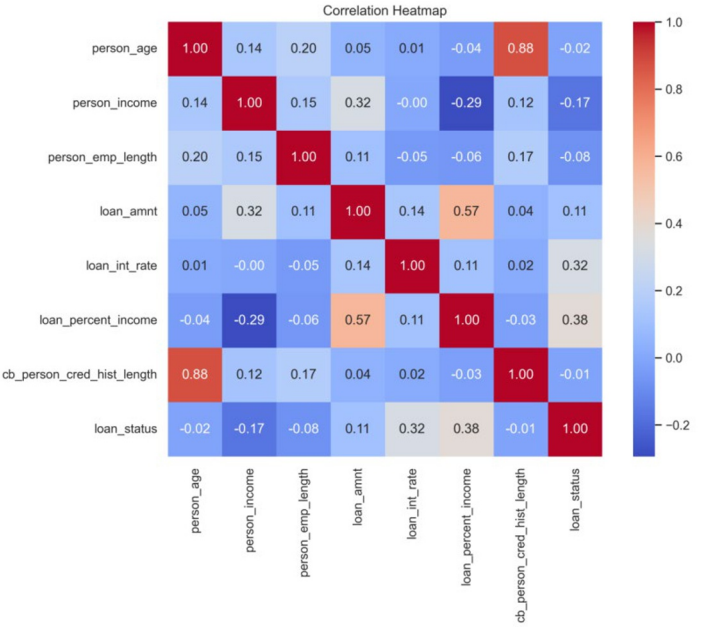


Fig. 3. Correlation Heatmap of Numerical Variables.

Several variables were negatively correlated with default. Among these, the strongest was person_income, with a correlation of -0.17. Employment length was also negatively correlated with default, with a coefficient of -0.08. Age and credit history length were only slightly negative, at -0.02 and -0.01 respectively. These results suggest that financial strength and stability reduce default risk, although their raw linear relationships are weaker than those for repayment burden and interest rates.

The feature screening results based on ANOVA F-values were highly consistent with the correlation analysis, as shown in Table 3.

Table 3. ANOVA F-value Ranking of Predictors.

feature	f_value	p_value
loan_percent_income	5470.113348	0
loan_grade	5263.104784	0
loan_int_rate	3696.382265	0
person_home_ownership	1527.296947	0
cb_person_default_on_file	1077.328406	1.74E-232
person_income	951.7582747	5.12E-206
loan_amnt	364.4038067	8.56E-81
person_emp_length	211.8169308	7.80E-48
loan_intent	140.4596782	2.46E-32
person_age	12.75355378	0.000355845
cb_person_cred_hist_length	7.233837344	0.007157857

The strongest variable is the loan_percent_income, and the F value is 5470.11. Followed by loan_grade (5263.10) and loan_int_rate (3696.38). Other high-ranking predictors include person_home_ownership (1527.30), cb_person_default_on_file (1077.33) and person_income (951.76). Loan_amnt and person_emp_length also show meaningful prediction capabilities, with F values of 364.40 and 211.82 respectively. In contrast, person_age (12.75) and cb_person_cred_hist_length (7.23) are much weaker in single-variate screening, although the two are still statistically correlated.

Multicollinearity inspection identified one notably strong relationship, as shown in Table 4.

Table 4. Highly Correlated Variable Pairs.

feature_1	feature_2	correlation
person_age	cb_person_cred_hist_length	0.87841188

The correlation between person_age and cb_person_cred_hist_length was 0.8784, exceeding the threshold of 0.80 set in the project design and indicating significant overlap between age and cumulative credit history. However, this relationship is economically intuitive, and no other pair exceeded the threshold, so serious multicollinearity does not appear to be widespread in the dataset.

Overall, the correlation analysis and ANOVA screening reached the same conclusion: the most informative predictors are concentrated in loan application information, particularly variables related to affordability, loan pricing, and risk grade. However, pairing correlation alone is not enough to determine whether the overlap between predictors substantially distorts the multivariate interpretation. For this reason, the analysis uses VIF statistics and robustness check to exclude the relevant variables in Section 4.5 to further examine the numerical polycolinearity.

4.3 Logistic Regression Results

Logical regression is the main explainable model, as shown in Table 5 and Figure 4. In the evaluation, its performance is evaluated using a wider range of classification indicators.

As described in Table 5, on the retention test set, the accuracy of the logical regression model is 0.8643, and the ROC AUC is 0.8623. However, the additional minority indicators show a more subtle picture. The accuracy of the default class is 0.7712, the recall is 0.5380, the F1 score is 0.6338, the balance accuracy is 0.7467, the PR AUC is 0.7155, and the Brier score is 0.1039. These results show that the overall performance of the model is quite good, but when the traditional 0.5 classification threshold is used, it still lacks a large number of actual defaulters.

The confusion matrix below the threshold of 0.5 confirms this pattern. The model correctly classified 4,865 non-default borrowers and 765 defaulted borrowers, and produced 227 false positives and 657 false negatives. In fact, this means that the model is relatively conservative: it protects specificity, but at the expense of reducing sensitivity to risky borrowers.

The threshold adjustment further shows that under the F1 standard, the default cut-off value of 0.5 is not the best value.

Table 5. Extended Classification Performance of Logistic Regression on the Hold-Out Test Set.

Model	Accu-racy	AUC	Preci-sion	Recall	F1_scor e	Speci-ficity	Balanced_Ac-curacy	PR_AU C	Brier_S core	TN	F P	F N	T P	NPV
logistic_re-gression	0.86429 229	0.86227 134	0.77116 935	0.53797 468	0.63380 282	0.95542 027	0.74669748	0.71553 262	0.10393 939	48 65	22 7	65 7	76 5	0.88102 137

As shown in Table 6, the optimal threshold for logical regression is 0.35, not 0.50. At this threshold, the accuracy decreased to 0.6735, but the recall increased to 0.6660, and the F1 score increased to 0.6697. The balance accuracy is also increased from 0.7467 to 0.7879. In the context of credit risk, this trade-off is important because reducing the number of missed defaulters can prove that a moderate increase in error alerts is reasonable.

Table 6. Threshold Tuning Results under the F1 Criterion.

Model	Best_Thresh-old_by_F1	Preci-sion_at_Best	Recall_at_Best	F1_at_Best	Balanced_Accu-racy_at_Best
logistic_regression	0.35	0.673541963	0.665963432	0.669731259	0.787911017

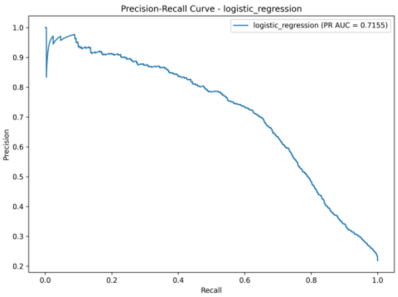


Fig. 4. Precision–Recall Curve for Logistic Regression.

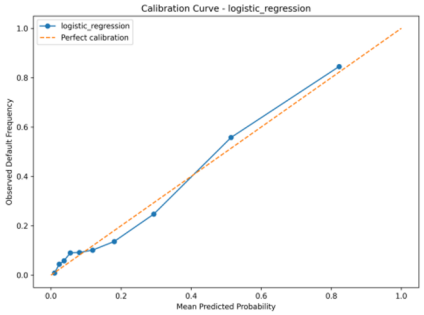


Fig. 5. Calibration Curve for Logistic Regression.

The calibration curve in Figure 5 shows that the predicted probability of the model is still roughly consistent with the observed default frequency, although the calibration is not perfect in all probability warehouses.

The coefficient estimates in Table 7 continue to provide useful conditional explanations.

Table 7. Logistic Regression Coefficients.

	coefficient	abs_coefficient
loan_grade_G	4.48197993	4.48197993
loan_grade_F	2.96452869	2.96452869
loan_grade_E	2.60646515	2.60646515
loan_grade_D	2.4083158	2.4083158
person_home_ownership_OWN	-1.6940391	1.6940391
loan_percent_income	1.46443705	1.46443705
loan_intent_VENTURE	-1.1021626	1.1021626
person_home_ownership_RENT	0.84995597	0.84995597
loan_intent_EDUCATION	-0.8419616	0.84196157
loan_amnt	-0.7206047	0.72060471
loan_intent_PERSONAL	-0.6156569	0.61565694
loan_grade_C	0.34541407	0.34541407
person_home_ownership_OTHER	0.28520066	0.28520066
loan_int_rate	0.23142503	0.23142503
loan_intent_MEDICAL	-0.1998586	0.19985857
loan_grade_B	0.15211907	0.15211907
person_income	0.11187617	0.11187617
loan_intent_HOMEIMPROVEMENT	0.05708163	0.05708163
person_age	-0.0513988	0.05139878
person_emp_length	-0.0421052	0.04210518
cb_person_cred_hist_length	0.03866159	0.03866159
cb_person_default_on_file_Y	0.00848818	0.00848818

The strongest positive coefficient is related to low-cost loan grades, especially loan_grade_G, loan_grade_F, loan_grade_E and loan_grade_D, indicating that the lower internally allocated loan quality is closely related to the higher probability of default. Loan_percent_income is still one of the most important predictors, with a positive coefficient of 1.4644, and loan_int_rate is positively related to the risk of default. The housing situation is also informational: person_home_ownership_RENT is positive and person_home_ownership_OWN is negative, indicating that housing instability is related to greater repayment difficulties.

A coefficient still needs to be carefully interpreted, that is, person_income. Although the descriptive correlation results show that the original relationship with the default is negative, the multivariate coefficient is positive, but it is very small. This will not invalidate the correlation of variables. On the contrary, it shows that in the condition setting of logical regression, income is evaluated together with the loan burden, loan amount, loan pricing and other borrower characteristics. Later model agnostic analysis

shows that the predictive role of income is indeed important, but it is more complicated than simple one-way linear effects.

4.4 Random Forest Benchmark Results

Random forests are used as reference models to test whether more flexible nonlinear classifiers can improve prediction efficiency. The assessment shows that the stochastic forest model is superior to logical regression in terms of threshold independence and threshold dependence.

Table 8. Extended Classification Performance of Random Forest on the Hold-Out Test Set.

Model	Accuracy	AUC	Precision	Recall	F1_score	Specificity	Balanced_Accuracy	PR_AUC	Brier_Score	TN	FP	FN	TP	NPV
random_forest	0.92815474	0.92498727	0.97134387	0.69127989	0.80772391	0.99430479	0.84279234	0.87209514	0.06423 53	5063	29	439	983	0.92021083

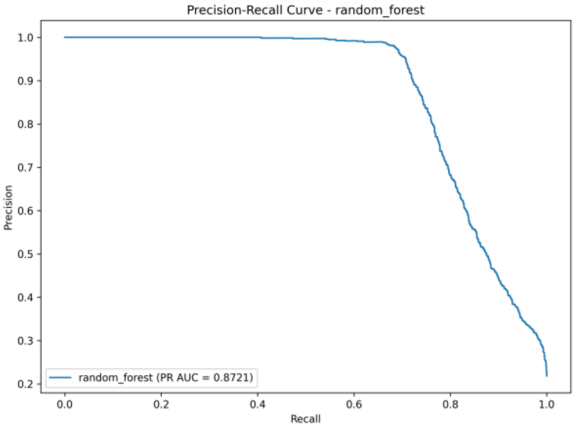


Fig. 6. Precision–Recall Curve for Random Forest.

As shown in Figure 6, the precision-recall curve further demonstrates the strong minority-class discrimination of the random forest model, with a PR_AUC of 0.8721. As shown in Table 8, in the retention test set, the accuracy of the random forest is 0.9282, and the ROC AUC is 0.9250. More importantly, it also achieved an accuracy of 0.9713, a recall of 0.6913, an F1 score of 0.8077, a balance accuracy of 0.8428, a PR AUC of 0.8721 and a Brier score of 0.0642. These values show that the benchmark model is not only better in ranking observations, but also more effective in identifying actual defaulters.

Under the traditional 0.5 threshold, the random forest correctly classified 5,063 non-default borrowers and 983 defaulted borrowers, with only 29 false positives and 439 false negatives. Compared with logical regression, the two main types of error are greatly reduced. This means that the improvement in random forest performance is not

only driven by the accuracy of the majority class; it also reflects the stronger detection of the minority class.

As shown in Table 9, the threshold adjustment produces the optimal F1 threshold of 0.45.

Table 9. Threshold Tuning Results under the F1 Criterion.

Model	Best_Threshold_by_F1	Precision_at_Best	Recall_at_Best	F1_at_Best	Balanced_Accuracy_at_Best
random_forest	0.45	0.944652908	0.708157525	0.809485531	0.848285361

Under this threshold, the accuracy is 0.9447, the recall is 0.7082, the F1 score is 0.8095, and the balance accuracy is 0.8483. Relative to the 0.5 threshold, there is little change, which indicates that the model remains relatively stable in the nearby decision-making cut-off value.

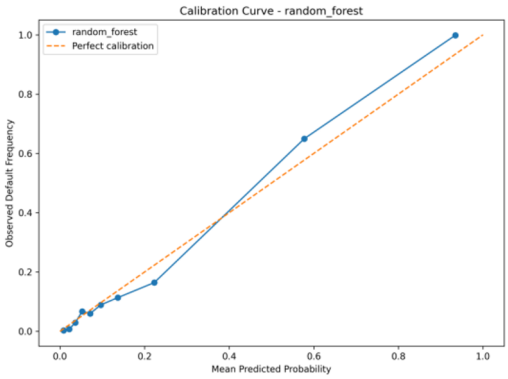


Fig. 7. Calibration Curve for Random Forest.

The calibration curve in Figure 7 shows a general reasonable consistency between the predicted probability and the default frequency of realisation, and compared with logical regression, the lower Brier score indicates a stronger overall probability quality.

The stability of the two models was further evaluated by using five-fold cross-validation.

Table 10. Five-Fold Cross-Validation Performance Summary.

Model	Metric	Mean	Std	Min	Max
logistic_regression	accuracy	0.86596459	0.00252179	0.86288961	0.86903117
logistic_regression	roc_auc	0.86916482	0.00684325	0.85823906	0.87937816
logistic_regression	precision	0.76640878	0.01167229	0.75239006	0.78419453
logistic_regression	recall	0.55522933	0.00748367	0.5443038	0.56579873
logistic_regression	f1	0.64387419	0.00608597	0.63739837	0.65339293
logistic_regression	balanced_accuracy	0.75396891	0.00359501	0.75072699	0.75972576
random_forest	accuracy	0.93069468	0.00228616	0.92784771	0.93459235

random_forest	roc_auc	0.93105515	0.00342324	0.9253596	0.93570713
random_forest	precision	0.96372997	0.00403243	0.95937211	0.97058824
random_forest	recall	0.70916333	0.0124566	0.69620253	0.73117523
random_forest	f1	0.81699545	0.0074274	0.80863192	0.8298722
random_forest	balanced_accuracy	0.85085021	0.00591244	0.8451302	0.86126711

As shown in Table 10, the average cross-verification value of the random forest is 0.9307, the ROC AUC is 0.9311, the accuracy is 0.9637, the recall is 0.7092, the F1 score is 0.8170, and the balance accuracy is 0.8509. The corresponding standard deviation is very small, indicating that the performance of the model does not depend on a single train test split.

In summary, these results support the interpretation that the default risk in this dataset is not purely driven by simple linear relationships. On the contrary, stronger benchmark performance suggests that nonlinear effects and interactions between affordability, pricing, borrower background and selected credit signals are also important.

4.5 Discussion of Key Predictors

When descriptive comparison, correlation analysis, variance analysis screening, logical regression coefficient, random forest benchmark results, arrangement importance and SHAP-based interpretation are considered together, a clear and robust pattern will emerge.

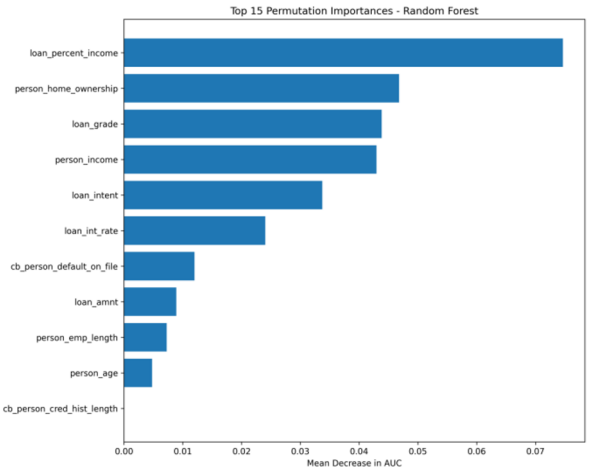


Fig. 8. Top Permutation Importances from the Random Forest Model.

As shown in Figure 8, loan_percent_income has the highest permutation importance, followed by person_home_ownership and loan_grade. The most stable and influential predictor in the dataset is the loan_percent_income. It shows the strongest positive correlation with the default value, the highest variance analysis F value, the positive coef-

ficient in logical regression, the highest arrangement importance and the highest average absolute SHAP value. In all these methods, the same conclusion is drawn: the repayment burden relative to income is the most powerful determinant of the risk of default in this case.

The second group of important predictors involves loan pricing and underwriting conditions. Variables such as loan_grade and loan_int_rate are still important in model-specific and model-different interpretation methods. Logical regression shows that poorer loan levels are associated with higher probability of default, and random forest, permutation analysis and SHAP ranking confirm that loan pricing and internally allocated credit quality contain a lot of predictive information. This supports the view that contract loan conditions account for a large part of the realised risk of breach of contract.

The basic information of the borrower is still the second most important information dimension, not a negligible supplement. In particular, personal income and housing ownership status have been repeatedly identified as important. The results that are not related to the model are particularly informative here. In permutation analysis, person_income and person_home_ownership are one of the most important original predictors. In SHAP analysis, person_home_ownership_RENT is one of the strongest specific indicators. The swarm model shows that leasing tends to push the model to achieve higher default forecasts. Person_income also ranks first in SHAP and permutation analysis, indicating that its predictive role is true, although its logical regression coefficient is small and has conditional symbols. This shows that income is important in a non-linear and dependent interaction way, rather than just through simple linear channels.

Credit history information seems to be relevant, but overall it is obviously weak. Cb_person_default_on_file retains some predictive value, but cb_person_cred_hist_length is still weak in the whole analysis. Among the main predictors, its variance analysis screening value is the lowest, the correlation with the target variables is very small, the sorting importance is very low, and the SHAP importance ranking is very low. This pattern shows that once affordability, pricing and borrower background variables are included, the length of the credit record only contributes to the limited incremental forecasting capacity.

The analysis also addresses concerns that this conclusion may be distorted by polycolinearity.

Table 11. Variance Inflation Factors for Numerical Predictors.

feature	VIF
person_age	4.44099982
person_income	1.800909758
person_emp_length	1.076761492
loan_amnt	2.471054063
loan_int_rate	1.026849731
loan_percent_income	2.402360415
cb_person_cred_hist_length	4.381800496

The pairing correlation shows a strong overlap between person_age and cb_person_cred_hist_length, but Table 11 shows that the VIF values of these variables are still lower than the regular critical thresholds, which are 4.44 and 4.38 respectively. There is no VIF with a numerical predictor greater than 5.

Table 12. Robustness Checks Excluding Correlated Predictors.

Specification	Num_Features	Accuracy	AUC	Precision	Recall	F1_score
full_model	7	0.86429229	0.86227134	0.77116935	0.53797468	0.63380282
drop_age	6	0.86459932	0.86239812	0.77108434	0.54008439	0.63523573
drop_credit_history_length	6	0.86367823	0.86242864	0.76915323	0.53656821	0.63214582
drop_both_age_and credit_history	5	0.86521339	0.86239909	0.77419355	0.54008439	0.63628832

In addition, the robustness check in Table 12 shows that the exclusion of age, credit record length or exclusion of these two variables can only be ignored in terms of accuracy, AUC, precision, recall and F1 score. This means that the main finding is not driven by severe redundancy between these two variables.

The explanation of model agnosticism further strengthens the ranking of the prediction group.

Table 13. Permutation Importance of Raw Predictors in the Random Forest Model.

feature	permutation_importance_mean	permutation_importance_std
loan_percent_income	0.074671709	0.002108027
person_home_ownership	0.046803983	0.002237345
loan_grade	0.043838326	0.001729028
person_income	0.042969557	0.002284057
loan_intent	0.033739157	0.002146282
loan_int_rate	0.024033908	0.002421484
cb_person_default_on_file	0.01199253	0.001578666
loan_amnt	0.008894789	0.001262567
person_emp_length	0.007267778	0.000733752
person_age	0.004777012	0.001337978
cb_person_cred_hist_length	6.08E-05	0.000461687

The permutation analysis in Table 13 shows that the most important original predictors are loan_percent_income, person_home_ownership, loan_grade, person_income, loan_intent and loan_int_rate, while cb_person_cred_hist_length contributes almost no incremental forecasts.

Table 14. SHAP-Based Global Importance Ranking for the Random Forest Model.

feature	mean_abs_shap
loan_percent_income	0.086036602
person_home_ownership_RENT	0.052912438
person_income	0.047308612
loan_grade_D	0.042920341
loan_int_rate	0.042472427
cb_person_default_on_file_Y	0.016349732
loan_intent_VENTURE	0.015515901

person_emp_length	0.013904192
person_home_ownership_OWN	0.013125676
loan_amnt	0.012259929
loan_intent_HOMEIMPROVEMENT	0.011356181
loan_grade_E	0.009915
loan_intent_EDUCATION	0.009152352
loan_grade_C	0.008644348
loan_intent_MEDICAL	0.007944698
loan_intent_PERSONAL	0.005915886
person_age	0.005254871
loan_grade_B	0.002969451
cb_person_cred_hist_length	0.002740574
loan_grade_F	0.001560142
loan_grade_G	0.000468444
person_home_ownership_OTHER	4.58E-05

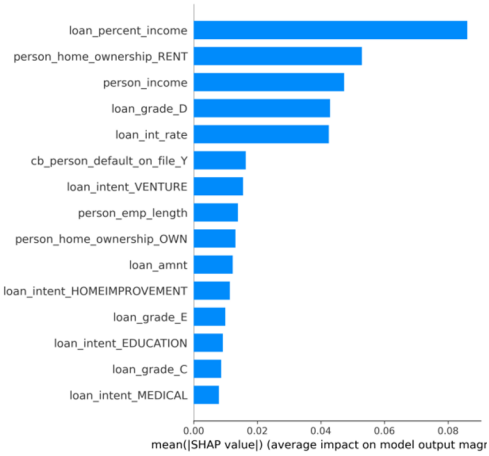


Fig. 9. SHAP Summary Bar Plot for the Random Forest Model.

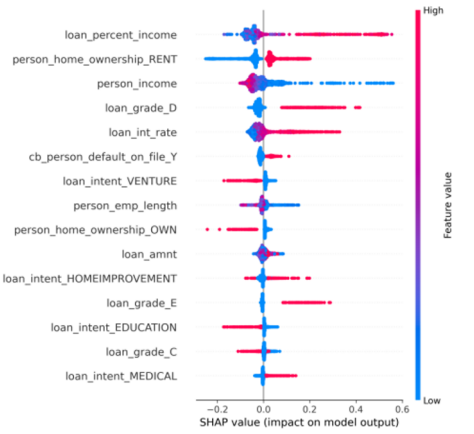


Fig. 10. SHAP Summary Beeswarm Plot for the Random Forest Model.

The SHAP ranking in Table 14, Figure 9 and Figure 10 draws the same broad conclusion, among which loan_percent_income, person_home_ownership_RENT, person_income, loan_grade_D and loan_int_rate become the most influential specific characteristics.

Overall, the evidence provides a more powerful answer to the second research question. The most powerful forecast signals come from loan application information, especially affordability, loan pricing and risk level. The borrower's basic information, especially the income and house ownership status, constitutes the second most important layer. Credit history information is not irrelevant, but its incremental contribution is weaker than the other two dimensions. This ranking is now supported not only by descriptive and model-specific evidence, but also by multi-collinear diagnosis, robustness check, sorting importance, SHAP analysis and cross-verification.

5 Conclusion

5.1 Summary of the Key Findings

This study used a structured personal loan dataset to examine the key predictors of credit default. The analysis combined data preprocessing, exploratory data analysis, feature screening, logistic regression, and random forest to evaluate predictive performance and the relative contribution of different types of borrower information.

Three main conclusions are drawn.

First, several variables consistently emerged as powerful predictors of default across different stages of analysis. In particular, loan_percent_income, loan_int_rate, loan_grade, person_income, and person_home_ownership were repeatedly identified as important in descriptive analysis, feature selection results, and model outputs. Among these, loan_percent_income was the most reliable predictor, indicating that repayment burden relative to income is the single strongest indicator of default risk in this dataset.

Second, the contribution of different information dimensions is unequal. Loan application information provides the strongest predictive signal, especially variables related to loan burden, interest rate, and loan grade. Borrower basic information, particularly income and housing ownership, constitutes the second most important category. Credit history information remains relevant, but its overall contribution is weaker than that of the other two dimensions. This directly answers the second research question and shows that in this setting, loan characteristics and borrower repayment capacity are more important than credit history length.

Third, the model comparison shows that random forest outperforms logistic regression. Logistic regression achieves robust performance and remains useful due to its interpretability, while random forest provides higher accuracy and AUC, indicating that default risk is influenced not only by linear effects but also by more complex non-linear patterns and interactions. In the analysis, these conclusions are strengthened through threshold adjustment, calibration check, multi-collinear diagnosis, robustness testing that excludes relevant predictors, sorting importance, SHAP interpretation and five-fold cross-validation, all of which support the broad ranking of predictor correlation.

5.2 Limitations

This study has several limitations. First of all, it is based on a single public dataset obtained from Kaggle. Although the dataset is suitable for structured empirical comparison, the publicly available metadata does not fully record the original sampling mechanism, geographical scope or collection period. Therefore, the survey results should not be interpreted as automatically promoted to all consumer loan markets. In particular, this study does not provide independent external verification for the second dataset from other sources or regions.

Secondly, although the analysis greatly enhances internal robustness through threshold adjustment, calibration analysis, VIF diagnosis, robustness check that excludes relevant predictors, sorting importance, SHAP-based interpretation and five-fold cross-verification, these procedures cannot completely replace the real external verification. They increase confidence in finding internal stability, but they do not eliminate the need to test generalisability in another empirical environment.

Third, the dataset is still limited to structured borrower, loan and credit record variables. It does not include behavioural information such as expenditure patterns or repayment trajectory, nor does it include broader macro-financial conditions, such as unemployment, interest rate system or economic shock. Fourth, for this analysis, the dataset is actually cross-sectional, so the time change of default behaviour cannot be directly evaluated. Finally, although the importance of SHAP and arrangement is more explainable, the benchmark random forest model is still not as transparent as logical regression at the level of simple coefficient table.

5.3 Future Work

Future research can expand this research in several directions. The first task is to use comparable datasets from other lenders, platforms, countries or time periods for independent external verification. This is the most important step to assess whether the ranking of current affordability, pricing, borrower background and credit history signals remains stable outside the current benchmark dataset.

Secondly, a richer data environment can be used to incorporate behavioural and transaction variables, such as repayment trajectory, expenditure pattern, utilisation rate or dynamic delinquency signals. Third, time index data will allow you to analyse whether the importance of predictors changes in economic conditions or credit cycles. Fourth, additional model classes, including gradient-up models and other advanced machine learning architectures, can be compared under the same extended evaluation framework used in revision studies.

Fifth, future work can deepen the interpretability by combining SHAP with local case-level interpretation and fair-oriented diagnosis, especially if the model is designed to be used for operating loans. From a practical point of view, the current survey results have shown that lenders should pay more attention to affordability measures, loan pricing and the stability of borrowers related to housing. However, confirming the external validity of this conclusion remains the core task of subsequent research.

Reference

1. Aboagye, E.F., Borketey, B., Danquah, K. and Borketey, D. (2022). A PREDICTIVE MODELING APPROACH FOR OPTIMAL PREDICTION OF THE PROBABILITY OF CREDIT CARD DEFAULT. *International Research Journal of Modernization in Engineering Technology and Science*, 04(08). doi:https://doi.org/10.56726/irjmets29286.
2. Alam, T.M., Shaukat, K., Hameed, I.A., Luo, S., Sarwar, M.U., Shabbir, S., Li, J. and Khushi, M. (2020). An Investigation of Credit Card Default Prediction in the Imbalanced Datasets. *IEEE Access*, 8, pp.201173–201198. doi:https://doi.org/10.1109/access.2020.3033784.
3. Bhandary, R. and Ghosh, B.K. (2025). Credit Card Default Prediction: an Empirical Analysis on Predictive Performance Using Statistical and Machine Learning Methods. *Journal of Risk and Financial Management*, 18(1), p.23. doi:https://doi.org/10.3390/jrfm18010023.
4. Wahab, F., Khan, I., & Sneha Sabada. (2024). Credit Card Default Prediction Using ML and DL Techniques. Internet of Things and Cyber-Physical Systems. https://doi.org/10.1016/j.iotcps.2024.09.001
5. Çallı, B.A. and Coşkun, E. (2021). A Longitudinal Systematic Review of Credit Risk Assessment and Credit Default Predictors. *SAGE Open*, [online] 11(4), pp.1–19. doi:https://doi.org/10.1177/21582440211061333.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

