Contents

1 Abstract

Credit risk assessment is a central task in retail banking, ensuring that financial institutions can effectively discriminate between creditworthy and non-creditworthy applicants. Modern credit scoring increasingly relies on data-driven methods that combine statistical modelling, machine learning techniques and domain-specific transformations to produce interpretable and stable credit-risk predictions. This project applies a **Weight-of-Evidence (WoE)** and **Information Value (IV)**–driven scorecard modelling framework using the *German Credit* dataset. The workflow includes data preparation, IV-based feature assessment, supervised binning, WoE transformation, logistic regression modelling, scorecard generation and out-of-sample validation. Performance is evaluated through stability measures such as the **Population Stability Index (PSI)** and accuracy metrics including AUC, Gini and RMSE. The resulting scorecard provides a transparent, regulator-compliant and empirically robust tool for credit-risk prediction.

2 Introduction

A central function of banks is to provide loans to individuals and companies. Credit scoring models are an essential tool for banks to assess the creditworthiness of lenders and predict how likely they are to meet their financial obligations. Popular models are based on the 3C's, 4C's, or 5C's, which stand for character, capital, collateral, capacity, and condition. However, with advancing technology, more novel approaches have emerged. In the subsequent paper, a credit scoring model is built and evaluated based on German data. The objective of the paper is to investigate the predictive value of demographic attributes. The objective of this study is not to enhance accuracy, but rather to offer insight into the structure of creditworthiness within the German context.

Credit risk modelling aims to quantify the likelihood that a borrower will fail to meet contractual repayment obligations. As banks, insurers and financial intermediaries increasingly rely on data-driven decision frameworks, scorecards have become the industry standard for credit underwriting due to their transparency, interpretability and regulatory acceptance (Hand and Henley, 1997; Thomas et al., 2002). In contrast to purely algorithmic black-box models, scorecards preserve a clear link between economic reasoning, data transformations and model parameters—an aspect that is essential for auditability and explainability in regulated environments.

The aim of this project is to build and validate a **predictive scorecard model** based on the *German Credit* dataset. To achieve this, the project adopts a structured analytical workflow aligned with established credit-risk modelling guidelines (Anderson, 2007; Siddiqi, 2017). The key methodological components are:

1. **Data Filtering and Variable Selection** – Choosing a subset of variables that capture behavioural and financial characteristics of borrowers.
2. **Information Value Analysis (IV)** – Quantifying the predictive strength of candidate variables.
3. **Supervised Binning and Weight-of-Evidence Transformation (WoE)** – Creating monotonic predictor–default relationships and ensuring stable logistic-regression estimation.
4. **Logistic Regression Modelling** – Fitting a parsimonious and interpretable model linking WoE-transformed predictors to the probability of default.
5. **Scorecard Generation** – Translating the regression coefficients and bin structures into a practical scorecard with additive scorepoints.
6. **Model Validation** – Assessing predictive accuracy (AUC, Gini, RMSE), calibration, and population stability (PSI) for both in-sample (IS) and out-of-sample (OoS) samples.

Using WoE transformations is beneficial because it enforces monotonicity, reduces the influence of outliers and yields logistic models with minimal multicollinearity (Siddiqi, 2017). This ensures that the final scorecard is both empirically robust.

Overall, this project demonstrates how predictive analytics can be applied to credit-risk modelling to produce a validated scorecard. The methodological steps should lead to a replicable blueprint for building credit-risk models.

3 Methodology

In order to build and assess these models, the Weight-of-Evidence (WoE) approach was taken. In it, the raw features of individuals are transformed in that they are sorted in bins based on their predictive strength to distinguish between defaulters and non-defaulters. To calculate it, the logarithm of the ratio of non-defaulters to defaulters within a group is taken, resulting in a score that represents credit risk. In doing so, the WoE approach allows to convert categorical or numerical values into values that are then processed in a logistic regression or scorecard model. In the underlying general formula, the event is defined in the underlying case as being a *good customer*, meaning that this individual repays their debt towards the bank, respectively in each group i.

$$WoE_i = \ln\left(\frac{Event_i\%}{NonEvent_i\%}\right) \tag{1}$$

The resulting predictor variables were then used in two models, first a logistic regression and secondly a scorecard model, to compare the accuracy of each. Both models aim to predict the probability of default $P(Y = 1)$ for the given set of predictors with the logistic approach using a sigmoid function. Sigmoid functions are mathematical functions, usually shaped like an S, used for classification problems by taking real numbers as an input and transforming them to values between 0 and 1 (probabilities). The model takes the form

$$P(\text{default} = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}} \tag{2}$$

where the coefficients β represent the individual contribution of each variable to the likelihood of default.

On the other hand, a scorecard model was implemented to serve as a more transparent comparison to the logistic regression. After the variables are initially binned and converted into their respective WoE values, each bin of each variable was assigned a fixed number of score points that reflect its contribution to the credit risk. The scorecard is derived from the logistic regression, which models the log-odds of default as:

$$\log\left(\frac{P(Y = 1 \mid X)}{1 - P(Y = 1 \mid X)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j \, WoE_j \tag{3}$$

Each bin-specific score is then calculated as the scaled product of the corresponding regression coefficient and WoE value,

$$\text{Score}_{j,b} = \text{Factor} \cdot \beta_j \cdot WoE_{j,b} \tag{4}$$

and combined with a base score derived from the regressional intercept. The final score of each individual applicant is therefore calculated as

$$\text{Total Score} = \text{BaseScore} + \sum_{j=1}^{k} \text{Score}_{j,b(j)} \tag{5}$$

where $b(j)$ denotes the bin assigned to variable X_j. The full scorecard was implemented using the *scorecard*() function from the eponymous package in R.

To evaluate and the performance of these models, a selected number of accuracy metrics, namely Area Under the ROC Curve (AUC) and the Gini coefficient, was used. These metrics measure how well the model is able to differentiate between defaulters and non-defaulters. Furthermore, the Root Mean Squared Error (RSME) quantifies the models' accuracy in their predicted probabilities by quantifying the average deviation between predicted and actual outcomes (in the train set). Additionally, matrices summarizing the performance by reporting true positives, false positives, true negatives and false negatives were generated

(confusion matrices). The stability or robustness of the models' scores and probability distributions were assessed with the Population Stability Index (PSI). This metric compares the distributions in the training set with those of the validation set to detect shifts in population characteristics across different samples. In this context, a low score is indicative of a stable model whereas high values suggest a potential drift or deterioration.

4 Data Selection and Preparation

To train the models to accurately predict the creditworthiness, it is imperative to select predictor variables based on their relevance to the individuals' financial behaviour.

To get a better understanding of the data structure, Table 1 gives an insight into the German dataset.

Table 1: Exemplary View of the Data Table

creditability	status.of.existing.checking.account	duration.in.month
good	... < 0 DM	6
bad	0 <= ... < 200 DM	48
good	no checking account	12
good	... < 0 DM	42
bad	... < 0 DM	24
good	no checking account	36

The selection process focused on variables that reflect past repayment behaviour, financial stability, and loan characteristics. Restricting the model to these input variables aims to produce transparent and robust results that reflect the real world risk modeling practices. To do so, each variable from the German dataset was evaluated on its Information Value (IV) to determine how well they are suited in discriminating between good and bad borrowers (data quality). Variables with high scores were retained for the subsequent modeling and further investigated in their underlying structuring. In practical terms, a higher IV relates to stronger predictive power of the underlying variable, with general thresholds of being good predictors of values larger than 0.3. Table 2 visualizes the calculated IV scores for the selected variables.

The five selected predictor variables are:

- **status.of.existing.checking.account** (categorical)
  This is a categorical variable that describes the applicant's checking account condition using qualitative labels such as "no checking account", "<0 DM", etc.
- **duration.in.month** (numeric)
  This is a numeric variable indicating the length of the loan contract in months, reflecting how long the applicant will take to repay off the credit.
- **credit.history** (categorical)
  This is another categorical variable that describes the applicant's past repayment behaviour, ranging from values of "no credits taken" and "all credits in this bank paid back duly" to "critical account".
- **savings.account.and.bonds** (categorical)
  This is another variable that categorises the applicant's savings level, both in their savings account and bonds. There are several categories such as "unknown/no savings account" to " < 100 DM" to separate those into groups.
- **purpose** (categorical)
  Purpose reflects a categorical variable that gives insight into what the applicant intends to do with the credit. Values range from "used car" and "new car" to "education", etc.

Table 2: Information Value of Variables

Variable	Information Value
status.of.existing.checking.account	0.67
duration.in.month	0.33
credit.history	0.29
savings.account.and.bonds	0.20
purpose	0.17

In the following steps, the data was filtered to exclude rows with missing values and split into train and validation set. The split chosen was at a ratio of .75 to .25. This is being done to have two independent samples for training and validating the model at a later point.

5 Weight-Of-Evidence (WoE) Binning and Transformation

The binning process was conducted using the *woebin()* function from the *scorecard* package in R and applied separately to both training and validation set as to prevent information leakage, in which certain characteristics of the training set might bias the validation set. The used function supports both categorical and numerical variables with an algorithm selecting the most appropriate method based on the underlying data distribution. As for the selected predictor variables, the variable *duration.in.months*, a continuous variable, was binned using the default "width" method, grouping ranges into intervals based on the distribution characteristics. On the other hand, the categorical variable *credit.history* was grouped by merging categories with similar risk profiles until meaningful bins emerged. After determining the structure, both datasets were transformed by replacing the raw predictor values with the corresponding WoE scores. The resulting datasets therefore consisted exclusively of numerical variables with normalized differences across variables, which serve as a robust basis for the further applied methods of logistic regression and scorecard modelling. Furthermore, the transformation results in the beneficial outcome of transparency, as the WoE values are easily interpretable: higher WoE values indicate lower risk.

6 Model Implementation

6.1 Logistic Regression

Following the WoE transformation of the datasets, a logistic regression model was employed to model the log-odds of an invidiual defaulting as a linear function of the selected predictor variables. The WoE transformed values are now used as the predictors $X_1, X_2, ...X_k$ as outlined in equation 2 in the chapter Methodology. As each X_i is expressed in WoE units, the coefficients β_i directly represent the change in log-odds of default for each unit change in WoE, allowing for a more clear interpretation. The model was employed in R using the *glm()* function.

6.2 Scorecard Model

The scorecard model was constructed following the regression analysis. After the variables from the German datasets were binned and transformed into WoE values, each bin was assigned a fixed score that reflect its contribution to the log-odds of default, with the number of points proportional to the product of the corresponding regression coefficient and its WoE value (Equation 4 in Methodology). All WoE binning plots for the selected variables are included in the Appendix for completeness.

7 Reporting of Empirical Results

7.1 Selected Predictors — Structure of Probability of Default

The selected predictors were determined and visualized earlier using the **Information Value (IV)** (Table 3).

Table 3: Information Value of Variables

Variable	Information Value
status.of.existing.checking.account	0.67
duration.in.month	0.33
credit.history	0.29
savings.account.and.bonds	0.20
purpose	0.17

Interpretation:

- `status.of.existing.checking.account` shows the highest IV and is therefore the strongest predictor.
- `duration.in.month` and `credit.history` show medium predictive power.
- `savings.account.and.bonds` and `purpose` remain relevant and are retained in the modelling process.

7.1.1 Structure of PD by Bins (example: duration.in.month)

The WoE binning structure already produced earlier can be referenced directly:

Table 4: Duration-in-Month Bin Structure (Counts, PD, WoE, Points)

variable	bin	count	count_distr	posprob	woe	bin_iv	total_iv
duration.in.month	[-Inf,8)	56	0.088	0.107	-1.247	0.099	0.312
duration.in.month	[8,16)	214	0.337	0.224	-0.367	0.042	0.312
duration.in.month	[16,26)	204	0.321	0.289	-0.026	0.000	0.312
duration.in.month	[26,44)	108	0.170	0.407	0.499	0.046	0.312
duration.in.month	[44, Inf)	53	0.083	0.566	1.139	0.124	0.312

Interpretation:

The PD (positive probability) across bins shows a **monotonic increase** with longer loan durations. This is ideal behaviour for a scorecard variable and supports its inclusion in the model.

7.2 Logistic Regression Model — Parameter Estimates, z-values, Significance

7.2.1 WoE-based Logistic Regression (preferred model)

The regression results have already been produced through:

Table 5: WoE-based Logistic Regression Coefficients (Estimate, Std. Error, z-value, p-value)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8628	0.1023	-8.4351	0e+00
status.of.existing.checking.account_woe	0.8195	0.1310	6.2545	0e+00
duration.in.month_woe	0.9772	0.1897	5.1521	0e+00
credit.history_woe	0.7643	0.1754	4.3565	0e+00
savings.account.and.bonds_woe	0.8944	0.2697	3.3164	9e-04
purpose_woe	0.9840	0.2506	3.9274	1e-04

Key observations:

- All coefficients are statistically significant (p-values shown in the table).
- Signs and magnitudes align with expectations from the IV and bin structures.
- The AIC reported earlier (`data_woe.glm$aic`) confirms model compactness.

7.2.2 Original-variable Logistic Regression (subset)

These results were previously printed via:

Table 6: Original-variable GLM (Selected Predictors)

	Estimate	Prob_z	Stars
(Intercept)	0.3675	0.4879	
status.of.existing.checking.account0 ...	-0.3628	0.1143	
status.of.existing.checking.account.....	-1.2535	0.0061	**
status.of.existing.checking.accountno...	-1.8571	0.0000	***
duration.in.month	0.0335	0.0000	***
credit.historyall credits at this ban...	-0.5436	0.3754	
credit.historyexisting credits paid b...	-1.1827	0.0157	*
credit.historydelay in paying off in ...	-1.0068	0.0709	.
credit.historycritical account/ other...	-1.9179	0.0002	***

The original-variable GLM confirms:

- Significant effects from checking-account categories
- Duration remains highly significant
- Demonstrates consistency with the WoE model

7.3 Scorecard Validation — Gini IS & OoS

Metrics have already been generated by:

Table 7: Table: Probability-Prediction Metrics (RMSE, AUC, Gini)

RMSE	AUC	Gini	RMSE	AUC	Gini
0.3987765	0.8022465	0.6044929	0.4153243	0.7773915	0.554783

Table 8: Table: Score-Prediction Metrics (RMSE, AUC, Gini)

AUC	Gini	AUC	Gini
0.801966	0.6039319	0.7779007	0.5558014

And for scorecard points:

Interpretation:

- **In-sample Gini ~ 0.60**
- **Out-of-sample Gini ~ 0.55**
- The small drop between IS and OoS indicates stable generalization.

7.3.1 Population Stability Index (PSI)

variable	dataset	psi
score	train_validate	0.0424

A PSI around 0.04 indicates **very stable** population behaviour between train and validation samples.

7.4 Summary & Practical Benefits

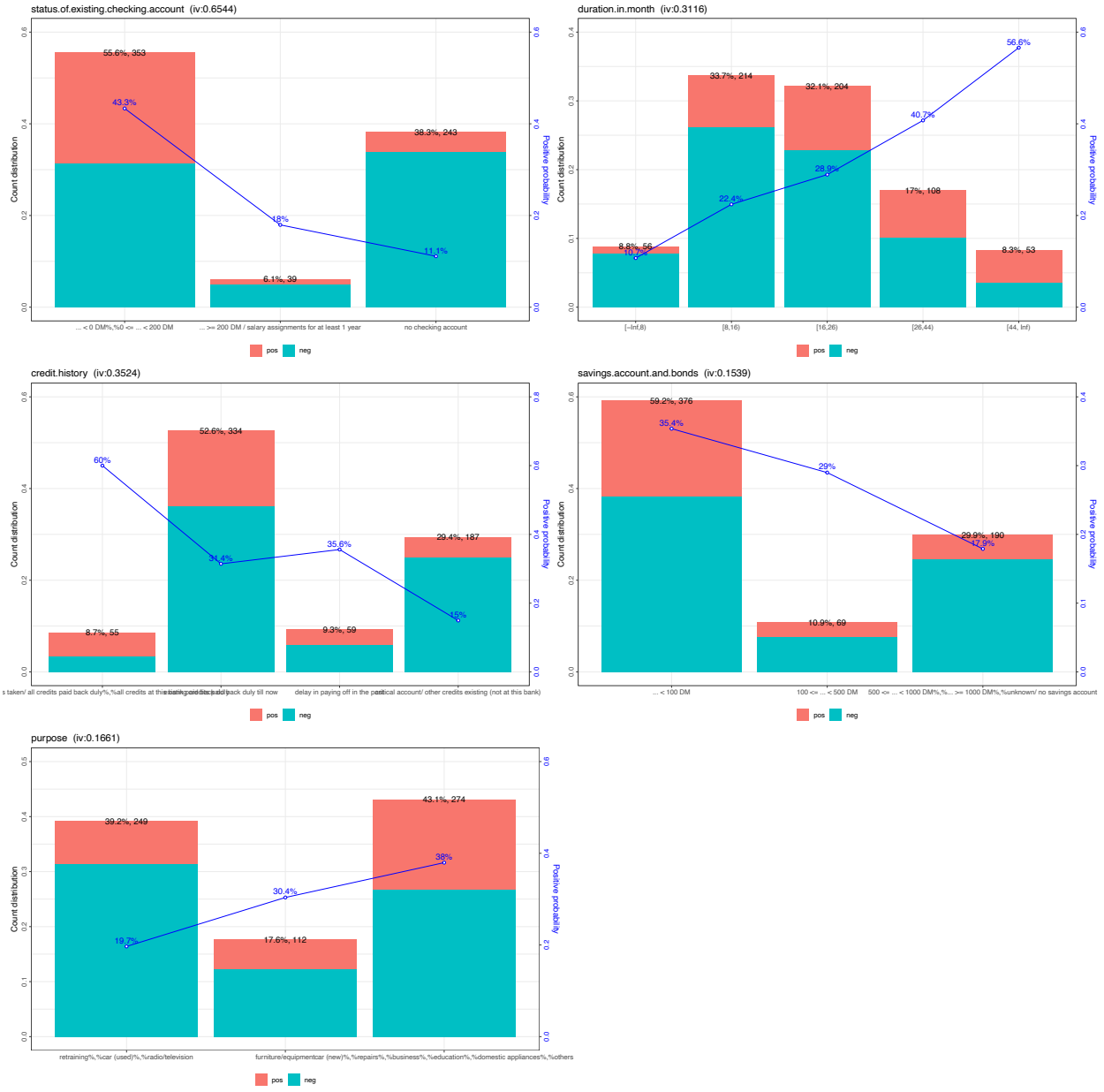
Summary of Findings

- IV analysis confirms five strong and relevant predictors.
- WoE transformation produces monotonic PD patterns and stable coefficients.
- Logistic regression coefficients (all significant) validate predictor usefulness.
- Scorecard performance is strong:
  - IS Gini about 0.60
  - OoS Gini about 0.55
- PSI about 0.04 demonstrates population stability.

Who Benefits

- **Risk management:** receives model stability evidence (Gini, PSI, IV).
- **Underwriting:** gains an interpretable scorecard with robust discriminatory power.
- **Credit policy:** can adjust rules using transparent variable effects and score contributions.
- **Regulatory/validation teams:** benefit from WoE-based transparency and monotonicity.

8 Appendix



References

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

Hand, D. J. and 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.

Siddiqi, N. (2017). *Intelligent Credit Scoring: Building and Implementing Predictive Models*. Wiley, 2 edition.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002). *Credit Scoring and Its Applications*. SIAM.