2025

Credit Risk Analysis and Classification Model on the German Credit Dataset

Milán Péter

Objective of Analysis

This analysis aims to explore credit risk through advanced machine learning techniques, with the dual purpose of uncovering key risk factors and developing predictive models capable of identifying potential bad creditors. By combining statistical insights with algorithmic modelling, the goal is to support more informed and data-driven decision-making in credit risk management.

A structured approach is applied, beginning with a thorough examination of descriptive statistics to understand the distribution and patterns within the data. This foundational analysis sets the stage for building classification models – such as Decision Trees, Random Forests, and other ensemble methods – that can accurately categorize credit risk. Special emphasis is placed on identifying the most influential features and ensuring the models are both robust and interpretable.

Python serves as the primary programming language due to its powerful ecosystem for data science and machine learning. Libraries including Pandas, NumPy, and Scikit-learn will be leveraged to support data preprocessing, model development, and evaluation.

To complement the analysis, effective data visualization is employed using tools such as Excel and Python's visualization libraries. Graphs, heatmaps, and dashboards are used to simplify complex results and enhance communication of key insights.

Ultimately, this project seeks to strengthen both conceptual understanding and practical application of credit risk modelling. By combining technical precision with clear visual storytelling, it demonstrates the value of machine learning in addressing real-world financial challenges.

About the Dataset

In the **German Credit Data** dataset, each entry represents an individual applying for a loan from a bank. When processing a loan application, the bank must assess the applicant's profile to determine whether to approve or reject the request. This decision carries inherent risks:

Risk of Rejecting a Good Applicant
If an applicant is a good credit risk – meaning they are likely to repay the loan – rejecting their application results in a loss of potential business for the bank.

Risk of Accepting a Bad Applicant
Conversely, if an applicant is a bad credit risk – indicating a high likelihood of default – approving their loan can lead to financial losses for the bank.

To mitigate these risks, banks need a robust decision-making framework to determine loan approvals effectively. A loan manager evaluates an applicant's demographic and socio-economic profile before making a final decision, balancing business opportunities with risk management.

Acknowledgement

The original dataset was prepared by Prof. Hans Hofmann in 1994. It is a publicly available dataset on UCI Machine Learning Repository.
Citation: Hofmann, H. (1994). Statlog (German Credit Data) [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77.

Institut für Statistik und Ökonometrie
Universität Hamburg
FB-Wirtschaftswissenschaften
Von-Melle-Park 5 / 2000
Hamburg 13

Content

The dataset contains 20 feature and 1 target variables, with 1000 instances.
Out of the 20 feature variables, 13 are **qualitative** attributes while 7 are **numerical** ones. The target variable is **binary**.

Attribute 1 *(qualitative)*
Status of existing checking account
A11: less, than 0 DM
A12: between 0 & 200 DM
A13: more, than 200 DM
A14: no checking account

Attribute 3 *(qualitative)*
Credit history
A30: no credits taken/all credits paid back duly
A31: all credits at this bank paid back duly
A32: existing credits paid back duly till now
A33: delay in paying off in the past
A34: critical account/other credits existing (not at this bank)

Attribute 19 *(qualitative)*
Telephone
A141: none
A142: yes, registered under the customer's name

Attribute 2 *(numerical)*
Duration in month

Attribute 5 *(numerical)*
Credit amount in DM

Attribute 11 *(numerical)*
Present residence since

Attribute 8 *(numerical)*
Instalment rate in percentage of disposable income

Attribute 13 *(numerical)*
Age in years

Attribute 6 *(qualitative)*
Savings account/bonds
A61: less, than 100 DM
A62: between 100 & 500 DM
A63: between 500 &1000 DM
A64: more, than 1000 DM
A65: unknown/no savings account

Attribute 17 *(qualitative)*
Job
A171: unemployed/unskilled & non-resident
A172: unskilled & resident
A173: skilled employee/official
A174: management/self-employed/highly qualified employee/officer

Attribute 18 *(numerical)*
Number of people being liable to provide maintenance for

Attribute 16 *(numerical)*
Number of existing credits at this bank

Attribute 4 *(qualitative)*
Purpose
A40: car (new)
A41: car (used)
A42: furniture/equipment
A43: radio/television
A44: domestic appliances
A45: repairs
A46: education
A47: vacation
A48: retraining
A49: business
A410: others

Attribute 15 *(qualitative)*
Housing
A151: rent
A152: own
A153: for free

Attribute 9 *(qualitative)*
Personal status and sex
A91: male & divorced/separated
A92: female & divorced/separated/married
A93: male & single
A94: male & married/widowed
A95: female & single

Attribute 14 *(qualitative)*
Other instalment plans
A141: bank
A142: stores
A143: none

Target *(binary)*
Debtor status
1: good
2: bad

Attribute 7 *(qualitative)*
Present employment since
A71: unemployed
A72: less, than 1 year
A73: between 1 & 4 years
A74: between 4 & 7 years
A75: more, than 7 years

Attribute 10 *(qualitative)*
Other debtors/guarantors
A101: none
A102: co-applicant
A103: guarantor

Attribute 12 *(qualitative)*
Property
A121: real estate
A122: building society savings agreement/life insurance
A123: car or other, not in Attribute 6
A124: unknown/no property

Attribute 20 *(qualitative)*
Foreign worker
A201: yes
A202: no

Files

The original dataset files contain 4 different files.

“german.data-numeric”
A file without extension.

Converted into readable .csv file.

For algorithms that need numerical attributes, Strathclyde University produced this file. It has been edited and several indicator variables added to make it suitable for algorithms which cannot cope with categorical variables. Several attributes that are ordered categorical have been coded as integer.

However, this dataset will not be used, as this analysis requires custom data preparation process.

“german.data”
A file without extension.

Converted into readable .csv file.

The original dataset, in the form provided by Prof. Hofmann, contains categorical/symbolic attributes.

This dataset will be the basis of this analysis.

“german.doc”
A file with outdated world extension.

Converted into readable .txt file.

It contains the description of the German Credit Dataset, including source information, attribute description, and cost matrix.

“Index”
A file without extension.

Converted into readable .txt file.

It contains some additional information about the sources. It is unnecessary for the analysis.

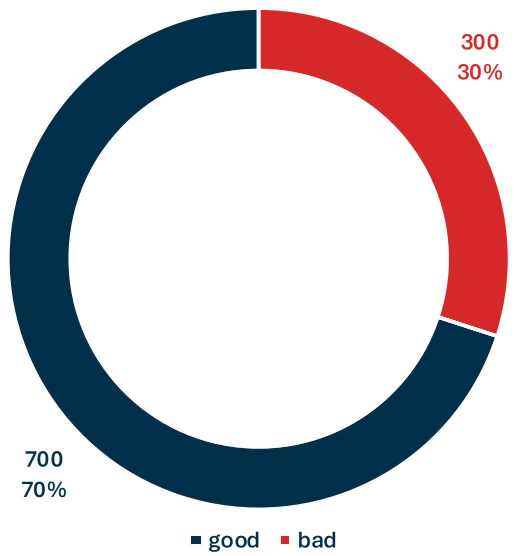
Initial Transformations in Python

- importing as a data frame with 21 columns (variables) and 1000 rows (observations)
- naming columns, as it is missing in the file
- replacing values code (e.g. A11) with more describing names
- exporting as **“credit_data.csv”** to work with henceforth

Script: “dataprep.ipynb”

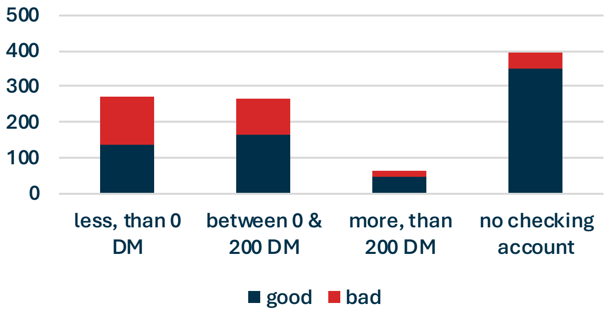
Exploratory Data Analysis

Target: Distribution of good and bad debtors



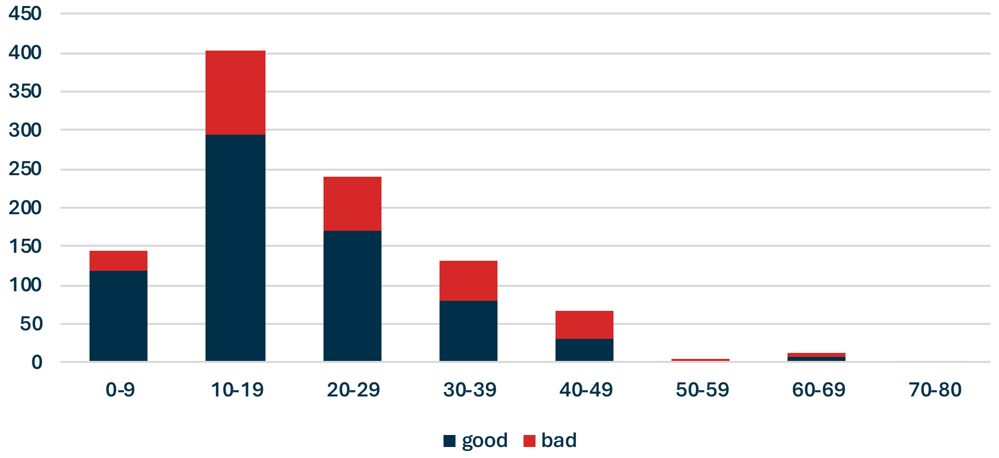
The sample consists of **1000** observations, with 700 classified as good debtors and 300 as bad debtors, resulting in a 70-30 split. From a probabilistic perspective, the likelihood of a new customer being a bad debtor is 30%.

Attribute 1: Status of existing checking account in relation to debtor status



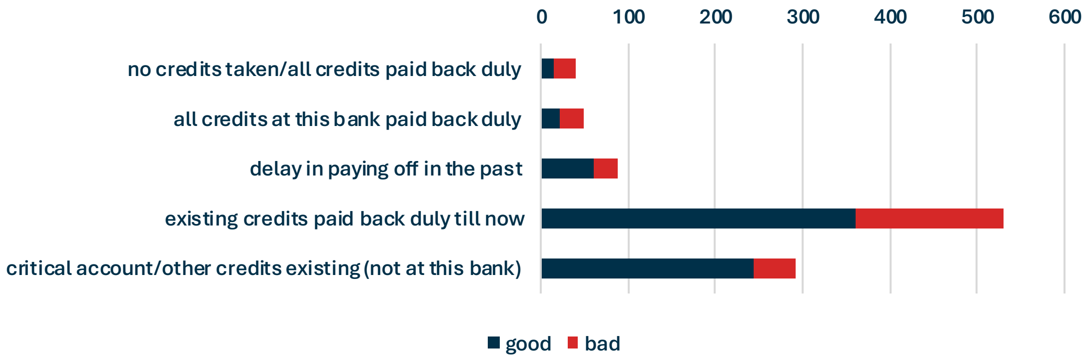
The status of existing checking accounts varies among debtors. Approximately 25% have less, than 0 Deutsche Mark, while another 25% hold an account with less than 200 Deutsche Marks. Fewer than 10% maintain a balance exceeding 200 Deutsche Marks, and the remaining 40% do not have a checking account at all.

Attribute 2: Distribution of loan duration in months in relation to debtor status



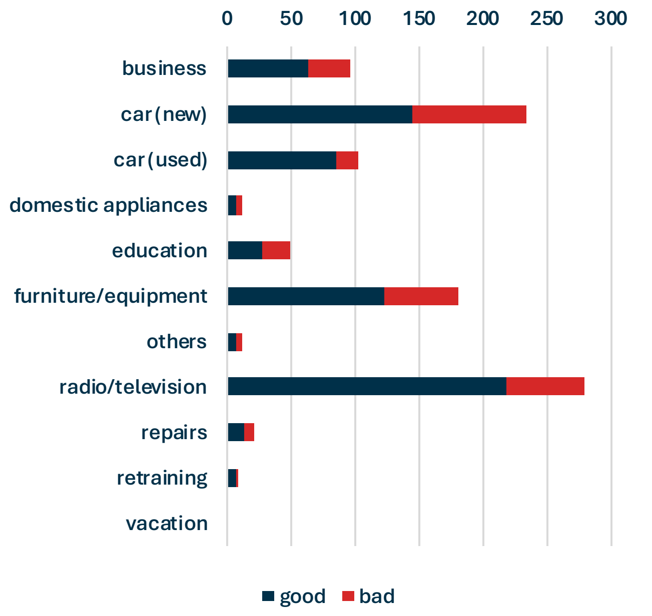
The majority of loans fall within the 10–19-month range, with a significant decline in frequency as the loan duration increases. Notably, bad debtors are more concentrated in longer loan durations (30+ months), while good debtors dominate the shorter-term loans.

Attribute 3: Credit history of debtor in relation to debtor status



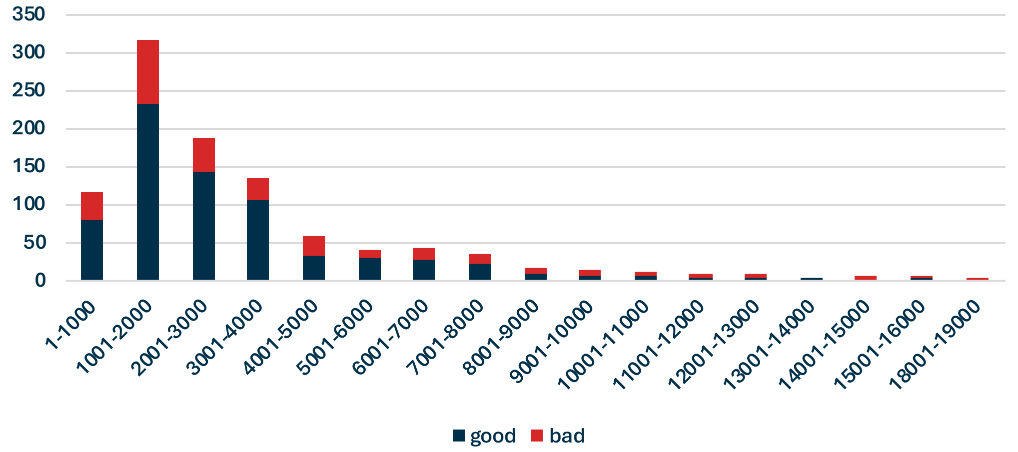
Most debtors have existing credits that have been duly repaid. Good debtors are more likely to have multiple credit lines, while bad debtors are more common among those with no prior credit or only bank-related credit. Delayed payments appear in both groups, but critical accounts are more prevalent among good debtors.

Attribute 4: Purpose of taking a loan in relation to debtor status



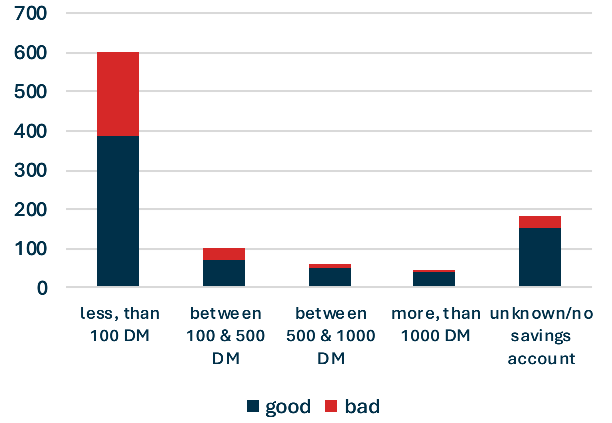
This data, collected in 1990's Germany, shows that most loans were taken for radio/television purchases (218 good, 62 bad) and new cars (145 good, 89 bad), reflecting consumer priorities at the time. Furniture/equipment and business loans were also common. Bad debtors were more concentrated in education and furniture/equipment loans, while used car and retraining loans had a lower bad debtor ratio. Notably, no loans were recorded for vacations, indicating different spending habits compared to today.

Attribute 5: Credit amount taken in relation to debtor status



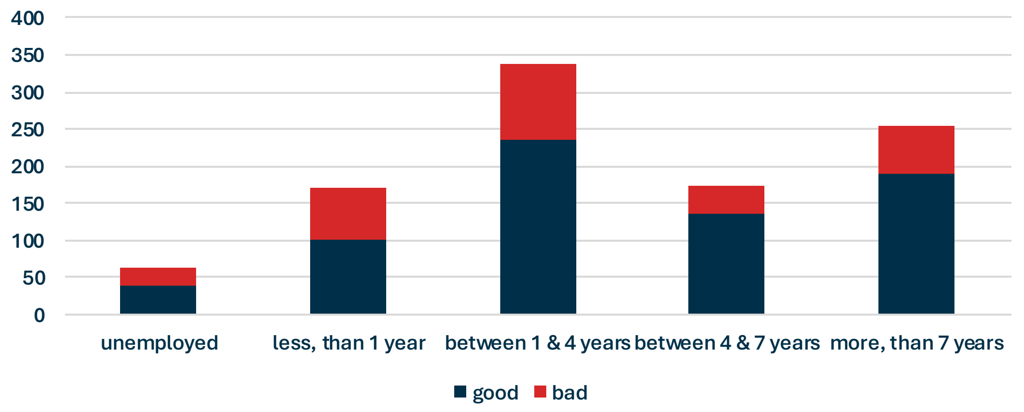
Most loans were relatively small, with the majority falling between 1 000-4 000 DM. Larger loans became less common, but bad debtors were more frequent in higher credit amounts, especially above 4 000 DM. Loans exceeding 15 000 DM were rare.

Attribute 6: Debtor's savings account/bonds in relation to debtor status

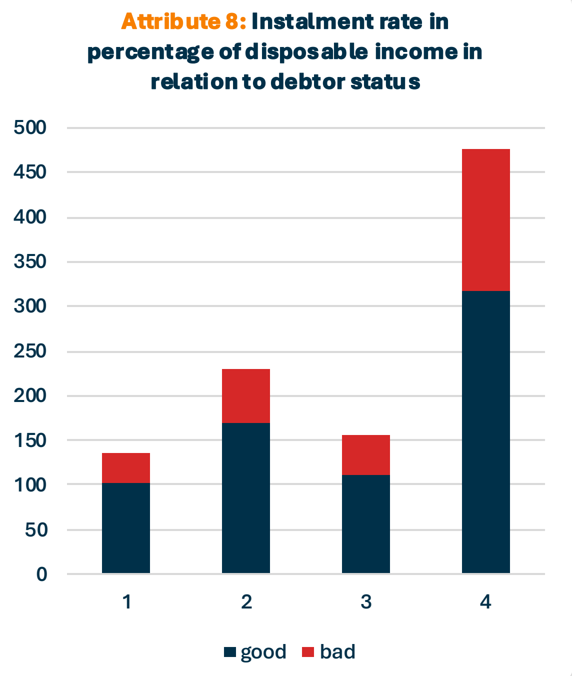


Most debtors had less than 100 DM in savings, with being more bad debtors in this category. Higher savings amounts were associated with fewer bad debtors, especially for those with over 1 000 DM. A notable portion had no savings account or unknown savings status.

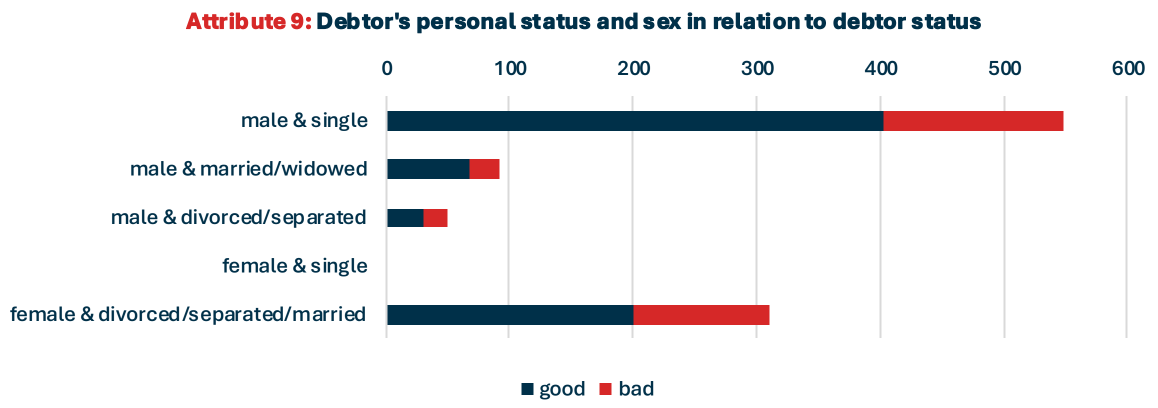
Attribute 7: Debtor's present employment since, in relation to debtor status



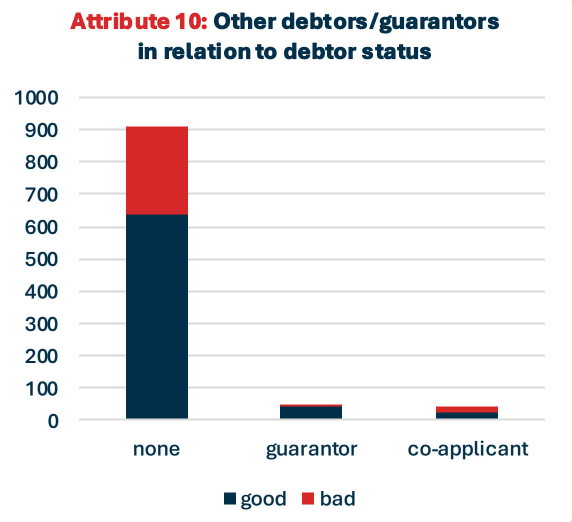
Debtors with longer employment history were generally more likely to be good debtors. Those employed for over 7 years had a lower proportion of bad debtors, while the 1 to 4-year group was the most common among both good and bad debtors. Unemployed and short-term employed individuals had a higher likelihood of being bad debtors.



A higher instalment rate – the percentage of a debtor's disposable income allocated to loan repayments – was associated with a greater number of bad debtors. Most debtors had an instalment rate of 4 (highest category), where bad debtors were most concentrated. Lower instalment rates (1 or 2) had relatively fewer bad debtors, suggesting that a higher repayment burden increased credit risk.

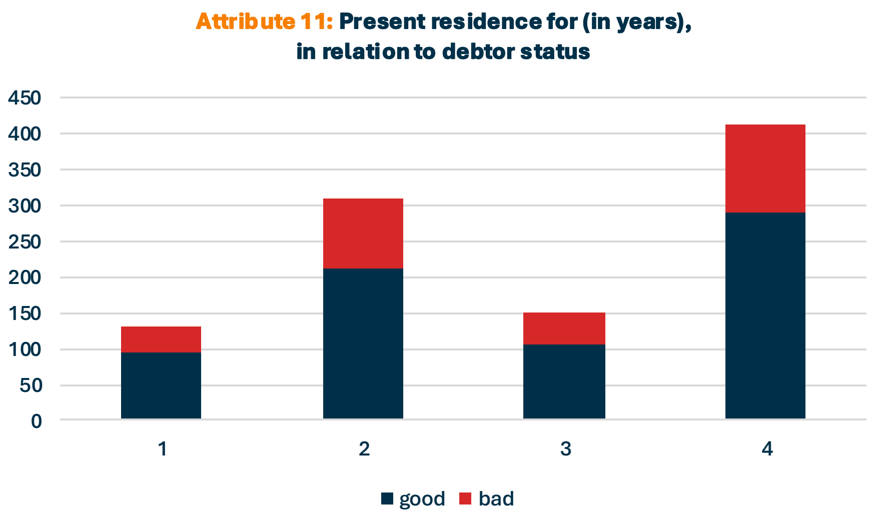


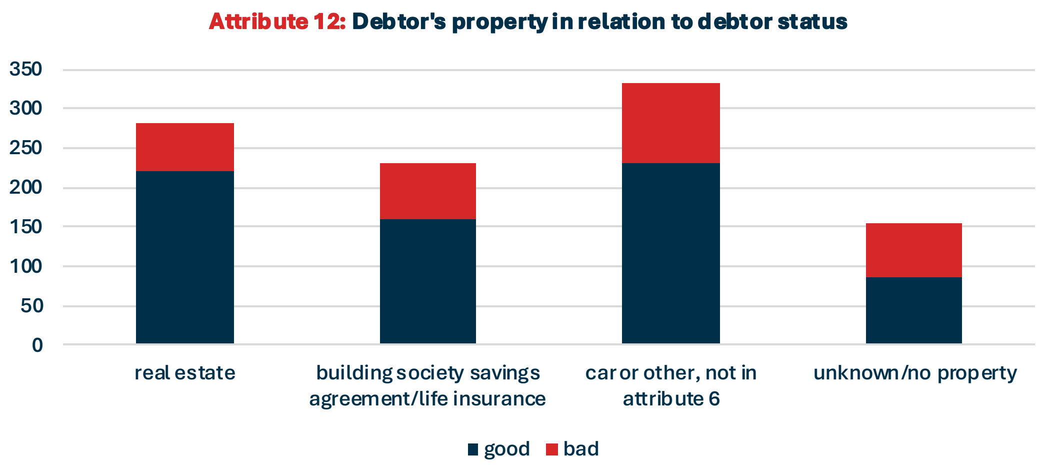
Single males were the largest debtor group, with a relatively high number of bad debtors. Married/widowed and divorced/separated males had fewer loans overall but still had some bad debtors. Interestingly, no single females had loans, while married, divorced, or separated females accounted for a significant portion of both good and bad debtors. This suggests possible gender-based differences in credit accessibility or financial independence at the time.



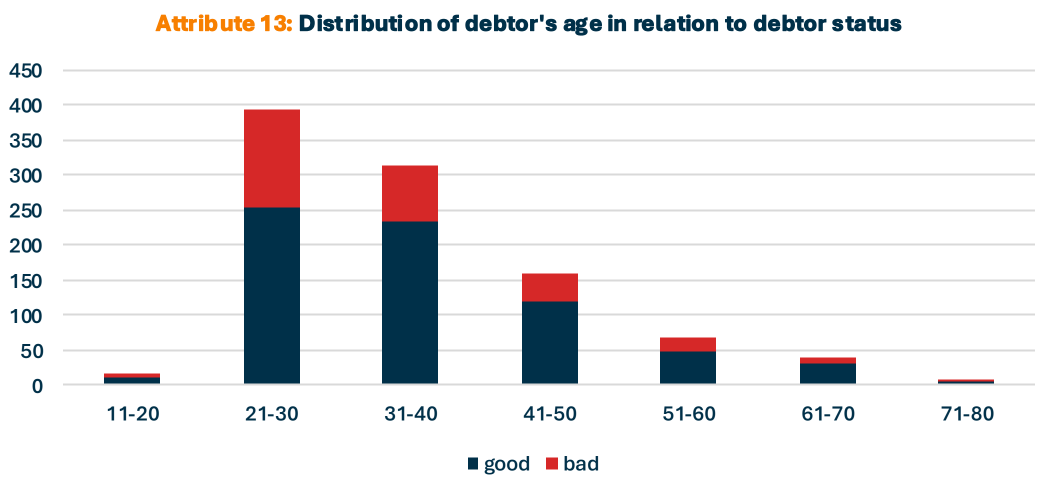
Most debtors had no guarantors or co-applicants, with bad debtors making up a significant portion of this group. Loans with a guarantor had a lower bad debtor ratio, suggesting added security. However, loans with a co-applicant had a relatively higher proportion of bad debtors, indicating possible financial risk-sharing among less creditworthy applicants.

Debtors who had lived at their current residence for longer periods tended to have fewer bad debtors. The 4-year group was the largest, but it also had the highest number of bad debtors in absolute terms. Shorter residence durations (1–2 years) showed a relatively higher proportion of bad debtors, suggesting that stability in residence might be linked to better creditworthiness.



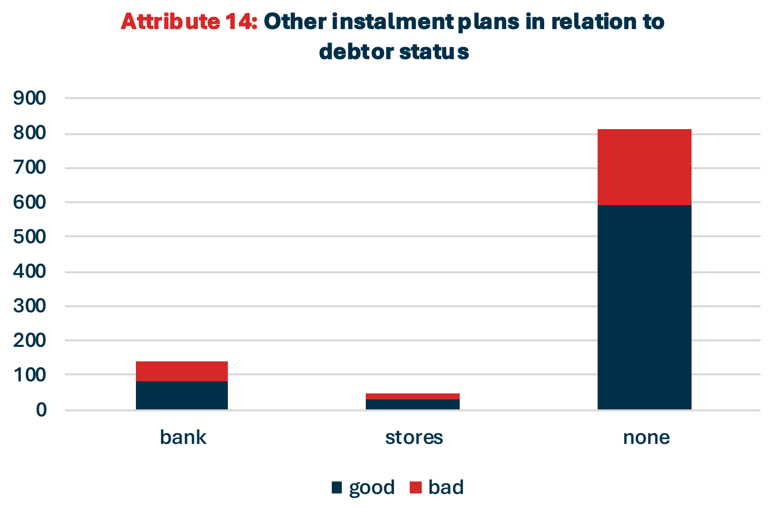


Debtors with real estate or life insurance savings had a lower proportion of bad debtors, suggesting greater financial stability. Those with cars or other property were also common but had a higher number of bad debtors. The highest bad debtor ratio was among those with no property, indicating a higher credit risk for applicants without significant assets.

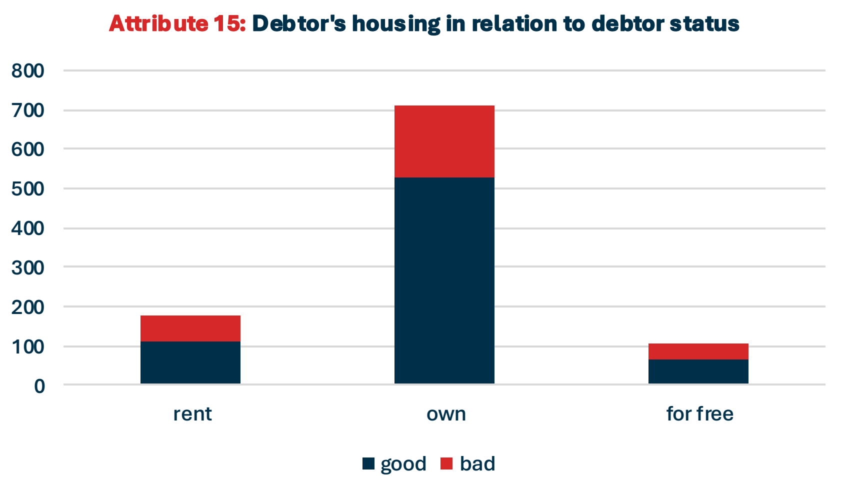


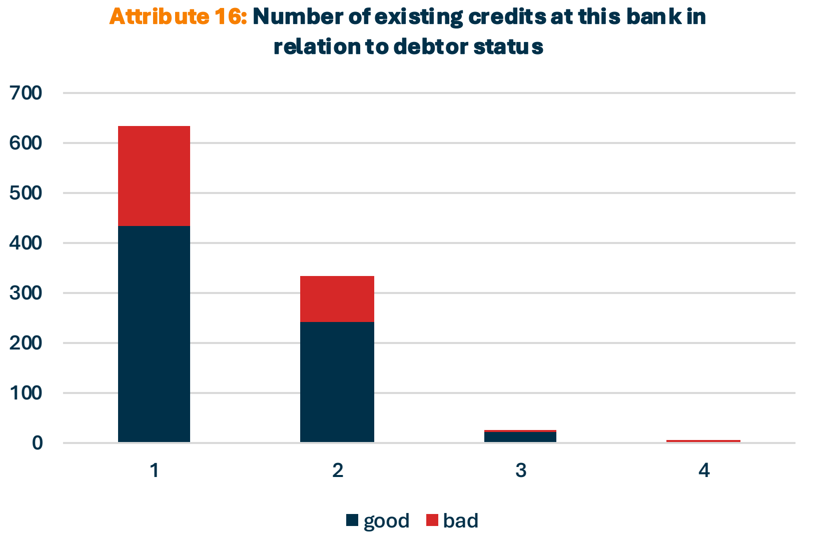
Most debtors were between 21 and 40 years old, with this group also containing the highest number of bad debtors. The proportion of bad debtors generally decreased with age, suggesting that older individuals were more creditworthy, likely due to greater financial stability. Very few loans were given to those under 20 or over 70.

Most debtors had no other instalment plans, though this group still included a significant number of bad debtors. Those with bank instalment plans had the highest proportion of bad debtors, suggesting a higher credit risk. Store instalment plans were less common but still had a notable share of bad debtors. This indicates that additional financial commitments, especially through banks, may have contributed to repayment difficulties.



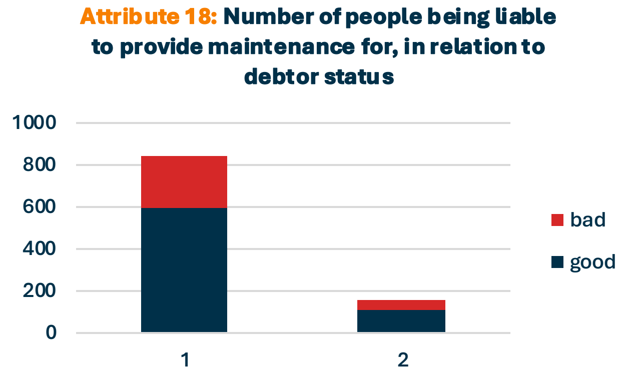
Debtors who owned their homes had the lowest proportion of bad debtors, indicating greater financial stability. Those who rented had a significantly higher bad debtor ratio, while individuals living for free (e.g., with family) had a relatively small but notable number of bad debtors, suggesting potential financial dependence.





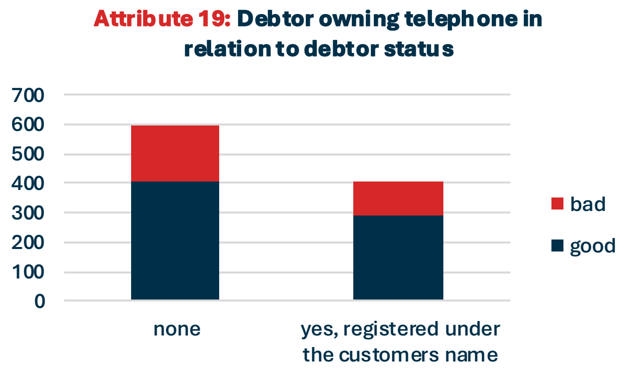
Most debtors had only one credit, but this group also contained the highest number of bad debtors. Having two credits was also common, with a lower proportion of bad debtors. Three or more credits were rare, but bad debtors were still present, suggesting that multiple loans did not necessarily indicate higher credit risk, possibly due to stricter lending criteria for repeat borrowers.

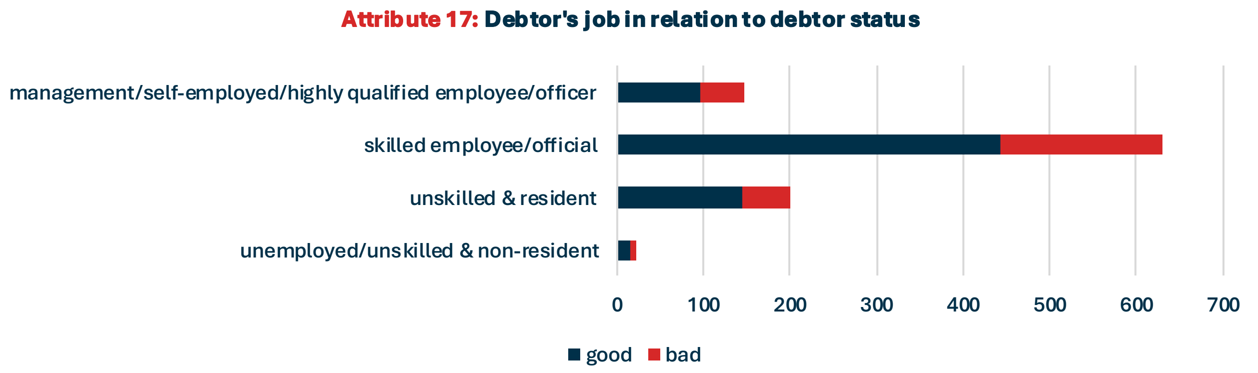
Most debtors had one dependent, with a significant number of bad debtors in this group. Those with two dependents were fewer in number but still had a notable share of bad debtors. This suggests that having more financial responsibilities may have contributed to higher credit risk, though the overall effect appears moderate.



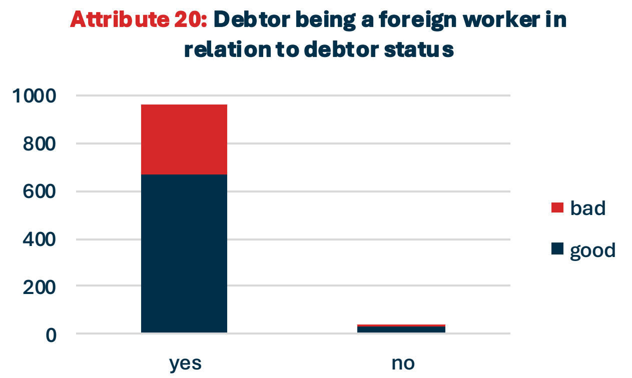
Skilled employees made up the largest debtor group but also had a significant number of bad debtors. Management and highly qualified professionals had fewer loans but a relatively high proportion of bad debtors. Unskilled workers (both residents and non-residents) had fewer loans overall, yet they still faced credit risk. This suggests that higher job qualifications did not always guarantee lower credit risk, though employment stability likely played a role in loan approvals.

Debtors with a registered telephone had a lower proportion of bad debtors, suggesting a possible link between having a telephone and financial stability. Meanwhile those without a telephone made up a larger share of bad debtors, indicating that limited access to communication might correlate with higher credit risk.





The vast majority of debtors were foreign workers, making up both the good and bad debtor categories. However, the proportion of bad debtors among foreign workers was slightly higher compared to local workers. Non-foreign (local) workers were a small group with very few bad debtors, suggesting they had a lower credit risk overall.



Encoding Variables

Data Processing in Python

Script: "dataprep.ipynb"

Most statistical analysis and modelling algorithms needs data to be numerical to be processed. Currently, the database consist of text values that are describing – which are easier to understand – instead of numerical ones. This part shows variables encoding process.

Binary Features

- T: debtor status
- A19: debtor owning telephone
- A20: debtor being a foreign worker

A19: "none" → 0 / "yes, registered under the customers name" → 1
A20: "no" → 0 / "yes" → 1

Target variable:
"good" → 0
"bad" → 1

It is a common practice to encode the target variable this way in classification tasks. Encoding the "bad" class (default) as 1 and "good" class (non-default) as 0 aligns with many financial modelling standards, where a default event is often treated as the "positive" outcome (i.e., the event we're trying to predict).

Discrete Features

- A1: status of existing checking account
- A3: credit history of debtor
- A4: purpose of taking a loan
- A6: debtor's savings account/bonds
- A7: debtor's present employment since
- A9: debtor's personal status and sex
- A10: other debtors/guarantors
- A12: debtor's property
- A15: debtor's housing
- A17: debtor's job

In case of discrete features, we have to differentiate nominal and ordinal variables. When dealing with nominal categorical variables meaning that they have no inherent order, One-Hot encoding is the preferred method. It creates a new binary column for each category. Meanwhile, when the categorical variable has an ordinal relationship, Label encoding is more useful as it converts each unique category into a number.

One-Hot Encoding

A3: 5 different values
A4: 11 different values
A9: 5 different values
A12: 4 different values
A14: 3 different values
A15: 3 different values
A17: 4 different values

Label Encoding

A1: 0 for no checking account,
then from 1 to 3 for the ascending categories
A6: 0 for unknown or no savings,
then from 1 to 4 for the ascending categories
A7: 0 for unemployed,
then from 1 to 4 for the ascending categories
A10: 0 for unemployed, 1 for unskilled,
2 for skilled, 3 for highly qualified

Continuous Features (that were suggested as discrete ones)

- A8: instalment rate in percentage of disposable income
- A11: present residence for (in years)
- A16: number of existing credits at this bank
- A18: number of people being liable to provide maintenance for

Continuous Features

- A2: loan duration in months
- A5: credit amount taken in Deutsche Marks
- A13: debtor's age in years

Continuous variables do not need encoding because they are already numerical and can be directly used in machine learning models. However, transforming or scaling them might be needed to depending on the model and data distribution. In this analysis, tree-based models (Decision Trees, Random Forests, AdaBoost, Gradient Boosting) will be used. These type of machine models work differently; they are not sensitive to feature magnitudes.

Tree models split data based on threshold values, not distances between points. They make decisions based on feature rankings and thresholds, not numerical distances. Because of this, feature scaling (e.g., Standardization, Min-Max Scaling) does not impact their performance.

Finally, all variables and column names have been renamed to enhance readability and facilitate their reference while writing scripts.

Correlation Matrix & Feature Importance

Measuring correlation is a fundamental step in credit risk modelling because it helps understand how features relate to each other and to the target variable. Ignoring correlation can lead to redundant features, unstable models, and poor interpretability. In this database, there are different kind of variables and these need to be measured with different correlation computing methods.

Script: "dataprep.ipynb"

Phi Coefficient correlation of target and binary feature variables

		DEBTOR_STATUS
	TELEPHONE	-0,04
	FOREIGN	0,08
CREDIT HISTORY	CRITICAL ACCOUNT	-0,18
	DELAYED REPAYMENT	0,01
	NO_CREDITS	0,14
	PAID_BACK OTHER_BANK	0,04
	PAID_BACK THIS_BANK	0,13
PURPOSE	BUSINESS	0,04
	CAR_NEW	0,10
	CAR_USED	-0,10
	DOMESTIC APPLIANCES	0,01
	EDUCATION	0,07
	FURNITURE EQUIPMENT	0,02
	OTHERS	0,03
	RADIO TELEVISION	-0,11
	REPAIRS	0,02
	RETRAINING	-0,04

		DEBTOR_STATUS
PERSONAL STATUS	FEMALE MARRIED	0,08
	MALE DIVORCED	0,05
	MALE MARRIED	-0,02
	MALE SINGLE	-0,08
GUARANTORS	CO-APPLICANT	0,06
	GUARANTOR	-0,06
	NONE	0,00
PROPERTY	LIFE INSURANCE	0,01
	OTHER	0,01
	REAL ESTATE	-0,12
	UNKNOWN	0,13
OTHER INSTALLMENTS	BANK	0,10
	NONE	-0,11
	STORES	0,05
HOUSING	FREE	0,08
	OWN	-0,13
	RENT	0,09

Spearman correlation of target and ordinal feature variables

	DEBTOR_STATUS	ACCOUNT_STATUS	SAVINGS	EMPLOYMENT_SINCE	EXISTING_CREDIT	JOB	OBLIGORS
DEBTOR_STATUS	1	0,23	0,02	-0,12	-0,05	0,04	0,00
ACCOUNT_STATUS	0,23	1	0,03	-0,12	-0,11	-0,04	-0,04
SAVINGS	0,02	0,03	1	-0,02	-0,01	-0,04	-0,03
EMPLOYMENT_SINCE	-0,12	-0,12	-0,02	1	0,14	0,08	0,10
EXISTING_CREDIT	-0,05	-0,11	-0,01	0,14	1	-0,02	0,10
JOB	0,04	-0,04	-0,04	0,08	-0,02	1	-0,10
OBLIGORS	0,00	-0,04	-0,03	0,10	0,10	-0,10	1

Pearson correlation of target and numerical feature variables

	DEBTOR_STATUS	DURATION	CREDIT_AMOUNT	INSTALLMENT_RATE	RESIDENCE_SINCE	AGE
DEBTOR_STATUS	1	0,21	0,15	0,07	0,00	-0,09
DURATION	0,21	1	0,63	0,07	0,03	-0,04
CREDIT_AMOUNT	0,15	0,63	1	-0,27	0,03	0,03
INSTALLMENT_RATE	0,07	0,07	-0,27	1	0,05	0,06
RESIDENCE_SINCE	0,00	0,03	0,03	0,05	1	0,27
AGE	-0,09	-0,04	0,03	0,06	0,27	1

The correlation analysis reveals that loan duration, account status, and credit amount are the strongest predictors of default, suggesting that longer loan terms, certain account conditions, and higher credit amounts increase risk. Employment stability, homeownership, and real estate ownership are associated with lower default probability, highlighting financial stability as a protective factor. Credit history plays a mixed role, with critical account history lowering default likelihood, while having no prior credits or repaying at other banks slightly increases risk. Renting and foreign status show mild associations with default, while instalment rate and purpose-related factors exhibit weaker influence. Several features, including residence duration, owning telephone, and existing credit lines, show little to no correlation, suggesting limited direct impact on default risk.

Baseline Model

A baseline model is an intentionally simple but meaningful model , serving as a reference point for assessing more complex models. It establish a minimum performance threshold, and any advanced models must significantly outperform them to justify their added complexity. The baseline should be easy to interpret while still capturing some patterns in the data. In this analysis, two models will be used: a majority class model, and a shallow decision tree.

Majority Class Model

This simple model predicts the majority class (e.g., the most frequent outcome in the target variable, debtor status) for all instances in the dataset. It provides a straightforward benchmark by assuming that the target class is predictable solely by its prevalence. While this model is expected to perform poorly in distinguishing between the classes, it helps highlight the baseline accuracy that any more complex model should exceed.

Confusion Matrix

Actual Values \ Predicted Values	Positive	Negative
Positive	700	0
Negative	300	0

Classification Metrics

Accuracy	0.70
Precision	0.49
Recall	0.70
ROC-AUC	0.50

Shallow Decision Tree

A decision tree with a shallow depth (limited to a depth of 1 or 2) was used as the second baseline model. This simple tree can create basic decision rules based on one or two features, offering a more interpretable, yet still simple, model compared to the Majority Class Model. The shallow decision tree helps to assess whether even basic decision-making logic can outperform the majority prediction model.

Confusion Matrix

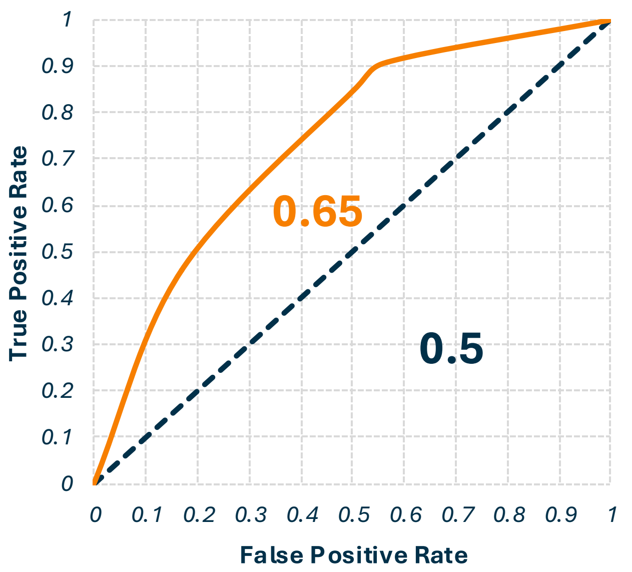
Actual Values \ Predicted Values	Positive	Negative
Positive	581	119
Negative	162	138

Classification Metrics

Accuracy	0.72
Precision	0.71
Recall	0.72
ROC-AUC	0.65

Explanation of Classification Metrics

ROC Graphs with AUC Values



The **ROC-AUC** (Receiver Operating Characteristic - Area Under the Curve) is a metric used to evaluate the performance of binary classification models. The ROC curve itself plots the true positive rate against the false positive rate at different classification thresholds, showing how well the model separates the two classes. The AUC, or area under this curve, quantifies the model's ability to distinguish between positive and negative cases. A value of 1.0 indicates perfect classification, while 0.5 suggests the model is no better than random guessing. If the AUC is below 0.5, the model is performing worse than random, meaning it systematically misclassifies cases. A higher AUC generally indicates a better-performing model.

Accuracy: The overall proportion of correctly classified instances across all classes.
Precision: The proportion of correctly predicted positive instances out of all predicted positive instances. It measures the model's ability to avoid false positives.
Recall: The proportion of correctly predicted positive instances out of all actual positive instances. It reflects the model's ability to capture all relevant cases.

Confusion Matrix

Actual Values \ Predicted Values	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

Cost-Sensitive Learning

In a typical classification problem, the model minimizes errors **equally** without considering the real-world impact of different misclassifications.

Cost-sensitive learning integrates the real-world **consequences of misclassification** into model training. Instead of just minimizing the number of errors, it aims to minimize total misclassification costs based on a predefined cost matrix.

Cost-Matrix

Predicted Values

Actual Values	Good Debtor	Bad Debtor
Good Debtor	0	1
Bad Debtor	5	0

The cost matrix defines how costly each type of misclassification is. This dataset requires use of a cost matrix as it was defined in the dataset description.

If a bad debtor is wrongly classified as a good debtor, the cost is **5** → This means granting a loan to a risky customer, which could lead to non-repayment, financial losses, and increased default rates.
If a good debtor is wrongly classified as a bad debtor, the cost is **1** → This means denying a loan to a creditworthy customer, which is a loss of potential business, but much less severe than a default.

Misclassification Cost of the Baseline Models

Misclassification cost refers to the penalty incurred when a model incorrectly classifies an instance. In many real-world applications, not all errors are equal.

Cost-Aware Evaluation
Computes the total misclassification cost using the cost-weighted confusion matrix after predictions. The theoretical maximum of misclassification cost in this case is 5000 – if the sample of 1000 would consist of only bad debtors and every one of them would be classified as good debtor.

Cost Matrix of the Majority Class Model

Misclassification Cost is 1500
Relative Cost is 30%

Predicted Values

Actual Values	Good Debtor	Bad Debtor
Good Debtor	0	0
Bad Debtor	1500	0

Cost Matrix of the Shallow Decision Tree

Misclassification Cost is 929
Relative Cost is 18.6%

Predicted Values

Actual Values	Good Debtor	Bad Debtor
Good Debtor	0	119
Bad Debtor	810	0

To make comparison easier, the relative misclassification cost is calculated as a percentage of the theoretical maximum. The Majority Class Model incurs 30% of the worst-case cost, while the Shallow Decision Tree reduces this to just 18.6%, highlighting a significant improvement in financial risk management.

Class Weights in Model Training – Adjusts the classification models to be more sensitive to costly misclassifications. Class weights help decision trees handle misclassification costs by making the model more sensitive to errors that carry higher penalties. In a credit risk model, misclassifying a bad debtor as good is far worse than the other way around. By assigning higher weights to the "bad debtor" class, the decision tree adjusts its splits to minimize these costly mistakes. This means the model becomes more cautious when predicting "good" debtors, prioritizing the reduction of false negatives. While this might slightly lower overall accuracy, it ensures that high-risk cases receive stricter scrutiny, aligning the model's decisions with financial risk priorities.

In this analysis, minimising the lowest possible misclassification cost while maintaining a satisfactory level of accuracy will be the primary objective.

In a real banking environment

These misclassification cost values represent the financial impact of incorrect credit decisions in a bank's lending model. They are measured in relative cost units, reflecting the penalties associated with misclassifying debtors.

In a real banking environment, these costs translate into loan loss provisions, capital reserves, and overall profitability impact. A high misclassification cost means more non-performing loans (NPLs), lower returns, and increased financial instability. By minimizing these costs through cost-sensitive learning, banks can make better lending decisions, reducing default rates while maintaining a profitable and sustainable loan portfolio.

By incorporating cost-sensitive learning, banks can strategically weigh financial risks, prioritizing the reduction of costly misclassifications. This leads to more accurate credit risk assessments, lower default rates, and a stronger, more resilient financial position.

Modelling Framework

Objective

The primary goal of this modelling framework is to minimize misclassification cost – with special attention to costly false negatives – and at the same time maximize overall classification accuracy. This dual focus ensures that the model is both financially efficient and statistically reliable in predicting credit risk.

Models

To capture various aspects of data complexity and learning patterns, four machine learning models are tested. Each offers unique strengths in handling nonlinear relationships, interactions between variables, and varying levels of noise in the data.

Hyperparameter Tuning

To optimize each model's performance, grid search was applied to systematically explore combinations of key hyperparameters (e.g., max depth, learning rate, number of estimators). This helps in finding the best settings that balance bias and variance, and more importantly, reduce costly misclassifications.

Validation

The models were validated using k-fold cross-validation, a robust method that splits the data into k equal parts. Each model is trained on k−1 parts and tested on the remaining fold, repeated k times. This approach provides a reliable estimate of out-of-sample performance, ensures stability, and avoids overfitting to any one train-test split.

Feature Importance

Feature importance scores indicate how much each input feature contributes to a model's predictions by measuring its impact on reducing error or impurity during training. These scores are commonly used in tree-based models, and they help identify which features the model relies on most. The values are typically normalized to sum to 1, highlighting the relative influence of each feature.

Decision Tree

A simple, interpretable model that splits data based on feature thresholds. It provides clear, rule-based decisions and is useful for understanding basic data structure and feature influence.

Random Forest

An ensemble of multiple decision trees trained on bootstrapped samples with feature randomness, which improves robustness and reduces overfitting. It handles noisy data well and provides more stable predictions than a single tree.

AdaBoost

A boosting method that sequentially builds weak learners, focusing on correcting previous mistakes by increasing the weight of misclassified observations. It is effective for improving accuracy, especially in the presence of bias.

Gradient Boosting (XGBoost)

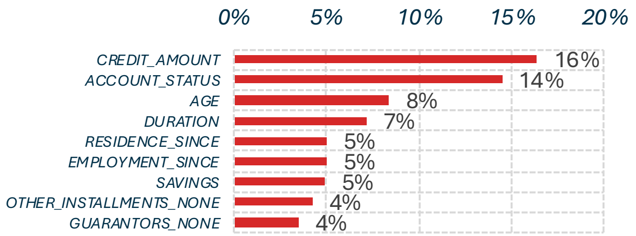
An advanced boosting algorithm that builds trees sequentially to minimize a loss function using gradient descent. XGBoost is known for its high predictive performance, regularization capabilities, and flexibility, making it particularly suitable for structured, tabular data such as financial datasets.

Decision Tree

Classification Metrics

Accuracy	0.71
Precision	0.75
Recall	0.71
ROC-AUC	0.69

Feature Importace



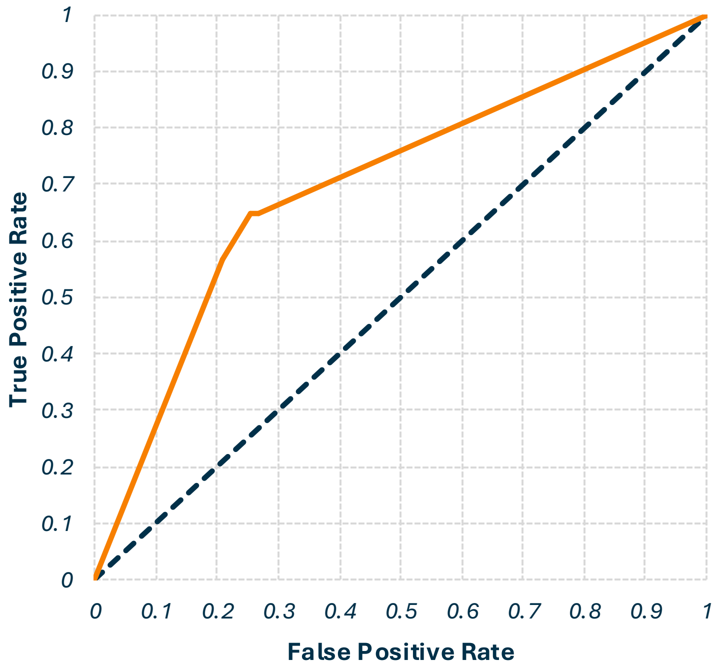
Cost Matrix of the Majority Class Model

Misclassification Cost is
130
(out of 1000)

Relative Cost is
13%

Actual Values	Predicted Values: Good Debtor	Predicted Values: Bad Debtor
Good Debtor	109	40
Bad Debtor	18	33

ROC Curve with AUC Values

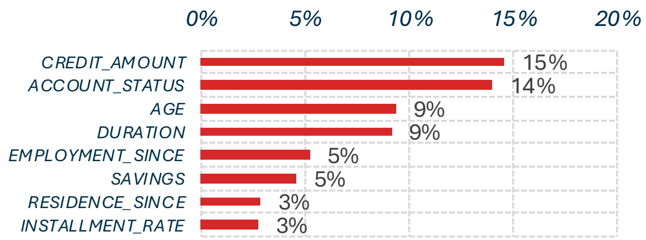


Random Forest

Classification Metrics

Accuracy	0.80
Precision	0.79
Recall	0.80
ROC-AUC	0.68

Feature Importance



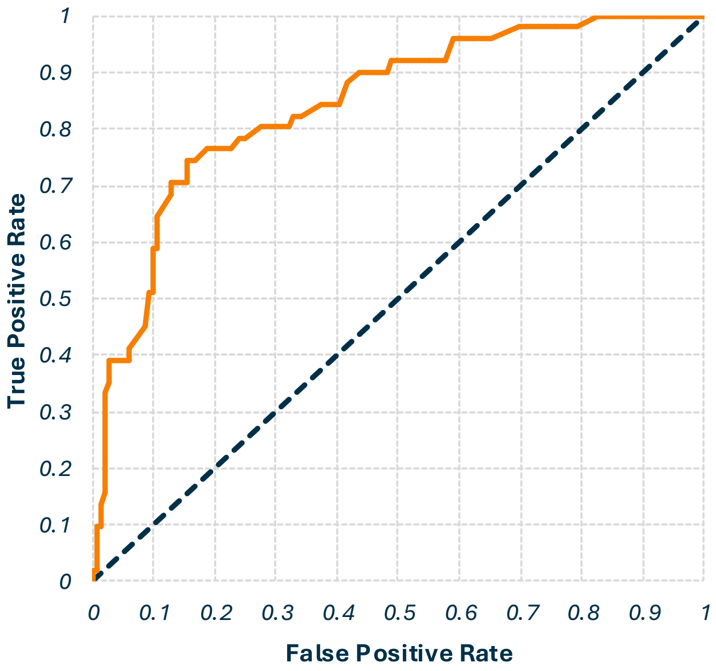
Cost Matrix of the Majority Class Model

Misclassification Cost is
156
(out of 1000)

Relative Cost is
15.6%

Actual Values	Predicted Values: Good Debtor	Predicted Values: Bad Debtor
Good Debtor	138	11
Bad Debtor	29	22

ROC Curve with AUC Values

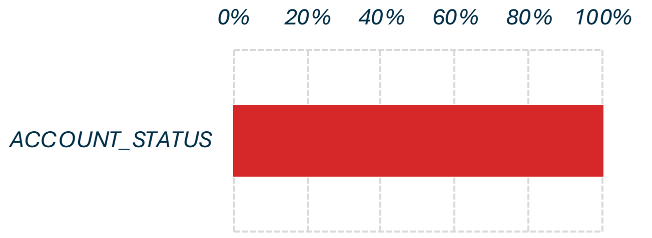


AdaBoost

Classification Metrics

Accuracy	0.60
Precision	0.79
Recall	0.60
ROC-AUC	0.70

Feature Importace



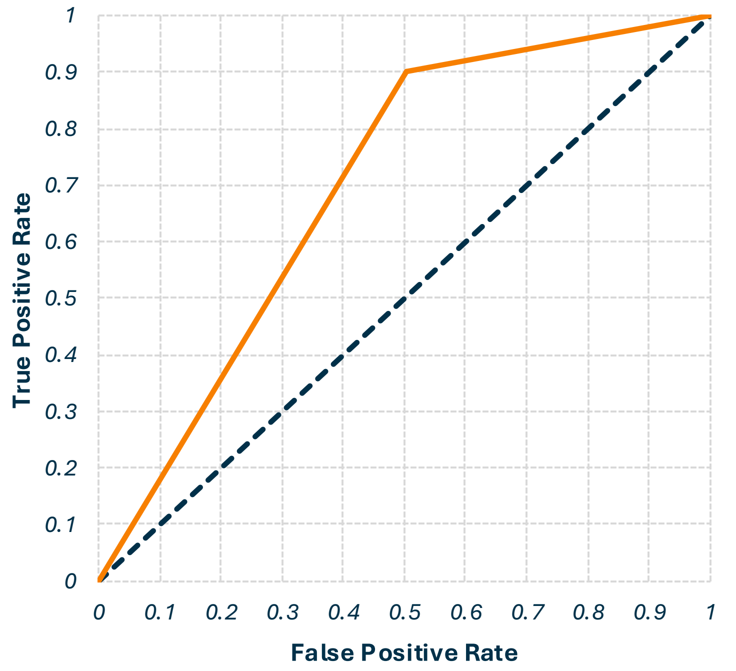
Cost Matrix of the Majority Class Model

Misclassification Cost is
100
(out of 1000)

Relative Cost is
10%

Actual Values \ Predicted Values	Good Debtor	Bad Debtor
Good Debtor	74	75
Bad Debtor	5	46

ROC Curve with AUC Values

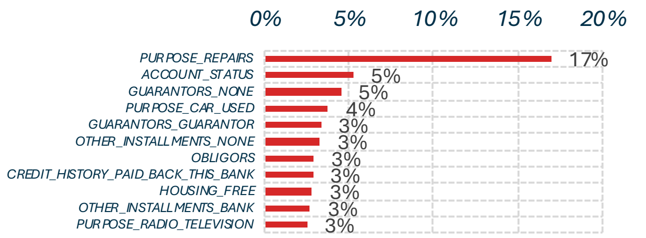


Gradient Boost

Classification Metrics

Accuracy	0.80
Precision	0.81
Recall	0.80
ROC-AUC	0.76

Feature Importace



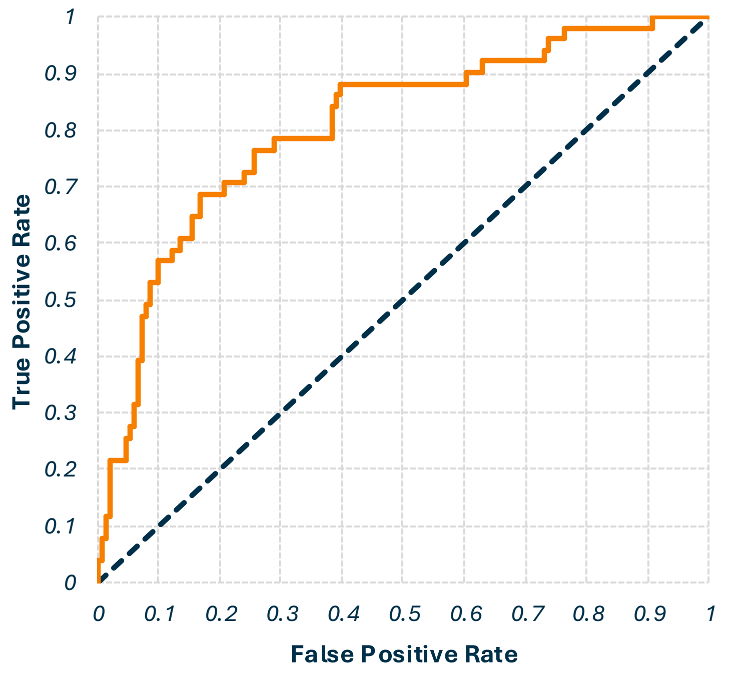
Cost Matrix of the Majority Class Model

Misclassification Cost is
105
(out of 1000)

Relative Cost is
10.5%

Actual Values \ Predicted Values	Good Debtor	Bad Debtor
Good Debtor	124	25
Bad Debtor	16	35

ROC Curve with AUC Values



Model Comparison

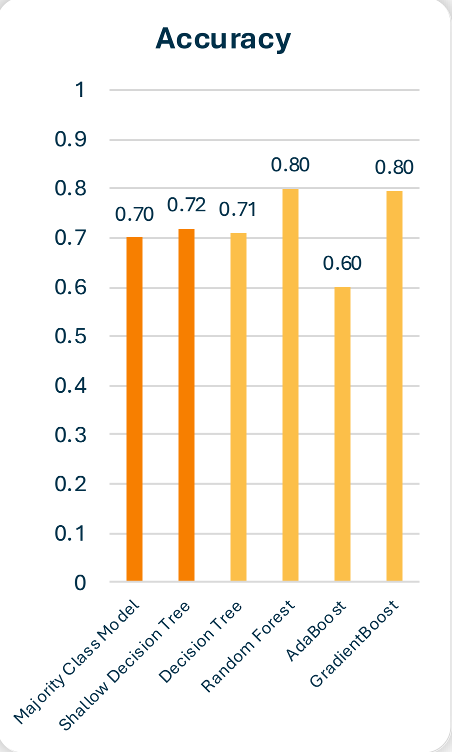
The primary objective of this modelling framework is to minimize misclassification cost – particularly costly false negatives – while maximizing overall classification accuracy. This ensures the model is both financially efficient and statistically reliable in predicting credit risk. The two baseline and four advanced models tested demonstrate different strengths in serving this dual goal.

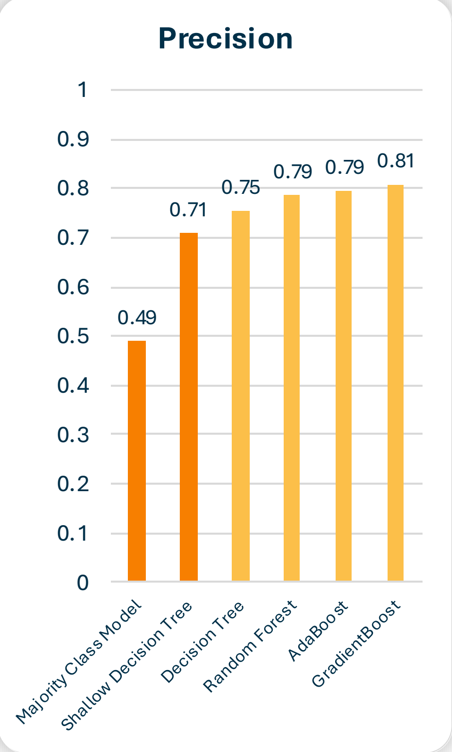
Gradient Boosting: with high accuracy (0.795), the highest precision (0.8086), strong recall (0.795), and the best ROC-AUC (0.7592), Gradient Boosting achieves the best overall balance between predictive power and error minimization. Its relative misclassification cost (0.105) is slightly less, that the lowest one. This makes it the most suitable model when both financial risk and classification reliability are critical, particularly in credit scoring scenarios where both false positives and false negatives carry substantial cost.

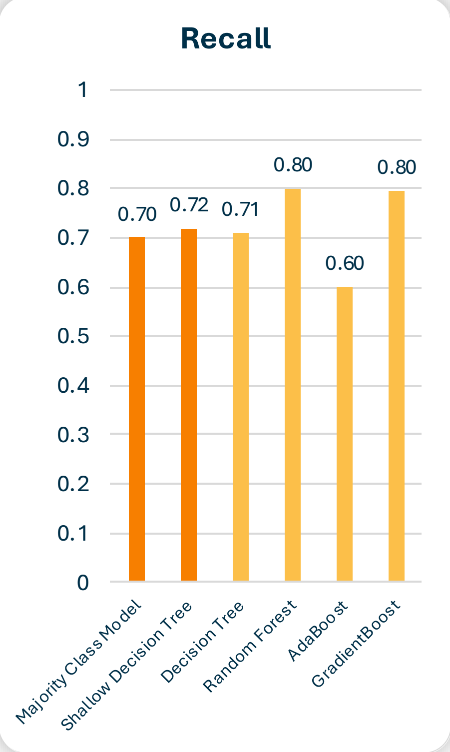
Random Forest: delivering the highest accuracy (0.80) and recall (0.80), Random Forest effectively captures true positives, reducing the risk of false negatives. Although its misclassification cost (0.156) is slightly higher than Gradient Boosting, it remains a strong candidate where sensitivity to high-risk cases is prioritized, such as default risk detection.

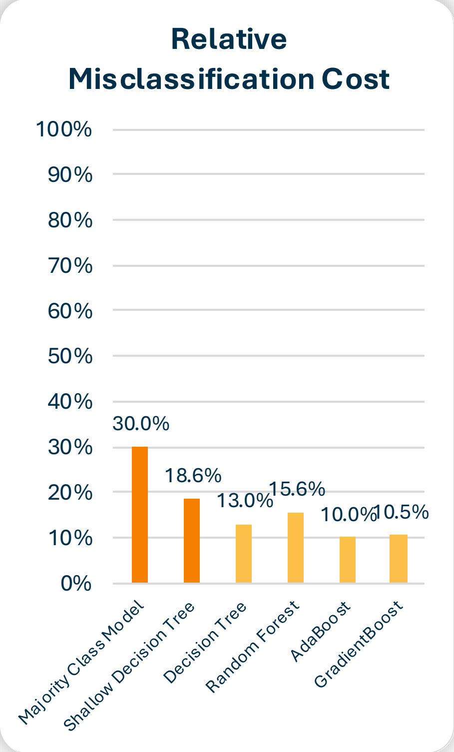
AdaBoost: despite lower accuracy (0.60) and recall (0.60), AdaBoost shows high precision (0.7948) and achieves the lowest misclassification cost (0.10). This model is best suited when financial cost reduction is the overriding priority, and minimizing false positive rates is less critical – such as in early-stage credit screening or pre-filtering processes.

Decision Tree: as a simple and interpretable model, the Decision Tree offers moderate performance across all metrics and a relatively low misclassification cost (0.13). While it does not outperform ensemble methods, it may be preferred in contexts where model transparency and ease of interpretation are essential, especially in regulatory environments.









Further Considerations & Real-World Extensions

Broader Modelling Approaches

While my analysis focused on classification models like Decision Tree, Random Forest, AdaBoost, and Gradient Boosting, other model families are also widely used in credit risk modelling, such as:

- **Logistic Regression:** a classical, interpretable model still heavily used in production
- **Support Vector Machines & k-NN:** applied in some scoring use cases
- **Neural Networks:** used in large-scale consumer scoring with richer datasets
- **Bayesian Models:** for uncertainty estimation

Risk Parameters Beyond Classification

In practice, financial institutions assess credit risk using more granular risk metrics:

- **PD (Probability of Default):** likelihood that a borrower will default
- **LGD (Loss Given Default):** proportion of the exposure the lender expects to lose if a default occurs
- **EAD (Exposure at Default):** expected outstanding amount at the time of default

In this analysis, only the classification (default vs non-default) was modelled. Predicting PD, LGD, and EAD would require different methodologies and additional data, such as:

Risk Metric	Additional Data Needed
PD	Historical default rates, time dimension, borrower risk profiles
LGD	Recovery amounts, collateral details, collection outcomes
EAD	Credit limits, utilization trends, payment behaviour

Default Reason Analysis

In practice, understanding *why* a customer defaults is critical.
This would require:

- Detailed transaction history
- Behavioural scoring (e.g., changes in balance usage, late payments)
- Macroeconomic indicators (e.g., unemployment, inflation)

Regulatory Context & Model Governance

Credit scoring models often operate under strict regulatory standards (such as Basel II/III, IFRS 9), requiring:

- Transparency and interpretability
- Robust validation and backtesting
- Fairness and bias monitoring

Conclusion

This project showcases essential techniques in credit risk classification, such as data preprocessing, model development, and performance evaluation. While the analysis focused on classification outputs (default vs. non-default), real-world credit risk modelling also involves PD, LGD, and EAD.

Although this project was based on a limited dataset, I'm familiar with the practical side of credit risk modelling through my 3-month internship as a Credit Risk Modelling Intern, where I gained experience with model monitoring in Power BI, supported the development of regulatory credit risk models in SAS, and assisted in building data infrastructure using SQL. This hands-on exposure has allowed me to bridge the gap between academic concepts and industry practices.