Assessing Credit Risk for Bank Loan Approvals

Zoe Dawkins → zoe.dawkins@torontomu.ca

Introduction

Problem:

- The bank requires an optimized system for its prospective customer loan approval process by predicting their credit risk based on the given German Credit Dataset.
- We used a data-backed machine learning approach that identifies where the customer has **good (1) or bad (0)** creditworthiness so the bank can increase profitable lending while decreasing the risk of defaults.
- The dataset includes 1000 records and 20 customer attributes (such as account balance, credit history, employment duration, etc.) along with a target variable — **Creditability** — that indicates a customer's credit risk

Solution:

- We used Weka and Python to implement 2 popular machine learning techniques (Decision Tree and Naïve Bayes)
- Each technique had 2 method of experimental implementation: baseline v.s. selected features. Our aim was to compare these 4 models based on **accuracy and recall** metrics.

- Key Results
  - Decision Tree Accuracy: 72.35 % baseline vs 72.94% (selected feature)
  - Naive Bayes Accuracy: 73.82% (selected features) vs 75% baseline
  - Critical Recall: 93.72% (identifies 94% of creditworthy applicants)
  - Most Relevant Predictors: Account Balance, Credit Duration, Payment History

Workload Distribution

Member Name	**List of Tasks Performed**
Abdalla	Train-Test Split, Data Preparation, Model Comparison
Gurleen	Decision Tree, Data Preparation, Conclusion, Model Comparison
Zoe	Naïve Bayes, Data Preparation, Recommendations, Summary/abstract

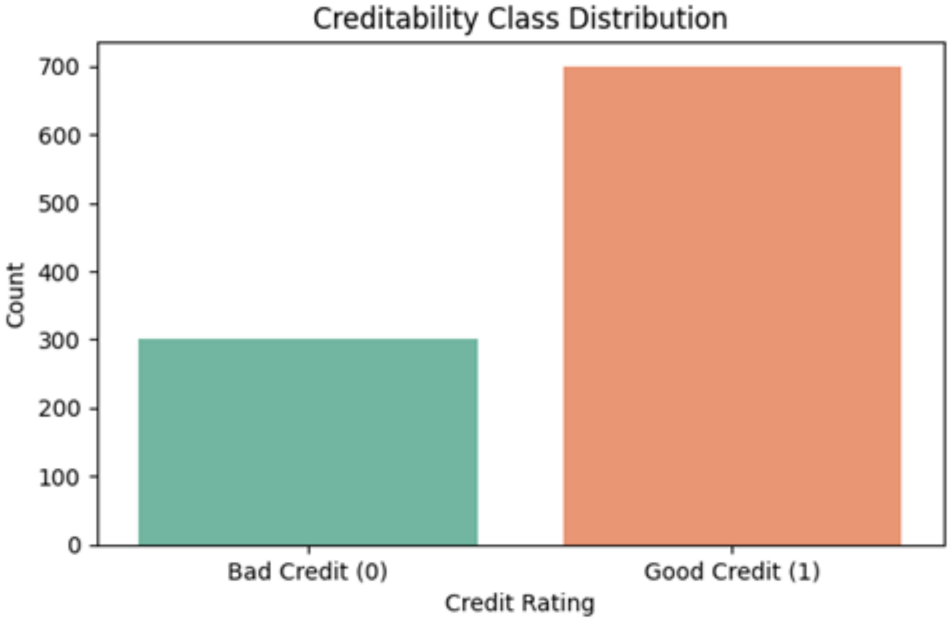
Data Preparation

1. Dataset Overview & Cleaning
   a. Loaded the dataset using Pandas
   b. Used .head(), .info(), and .describe() to review structure and summary stats
   c. No missing values detected → imputation not needed

2. Data Splitting
   a. Applied 66/34 Train-Test Split
   b. 660 records used for training, 340 for testing
   c. Balanced split ensures enough data for training without sacrificing test reliability – ideal for a 1,000-record dataset

Class Distribution

Class Distribution



The bar for 'Good Credit (1)' is much taller than the bar for 'Bad Credit (0)', it means we have more instances of 'Good Credit' in our dataset. This could indicate a class imbalance, and that we might need to consider strategies to address it before training our machine learning model.

With the imbalanced dataset, it can bias the model towards predicting the majority class more often, so metrics like recall and precision become more meaningful than just accuracy.

Dataset Attributes

- Our dataset had the following attributes types, with **nominal values being the most frequent**
- All given values were appropriate for the attribute type in a real world context so **no data validation steps had to be made**

Index	Attribute	Type	Values/Range	Description
0	Creditability (**Class Attribute**)	Nominal	{0,1}	The class attribute indicating credit risk. 1 = good credit, 0 = bad credit risk.
1	Account Balance	Ordinal	{1,2,3,4}	Checking account status: 1 = < 0 DM, 2 = 0-200 DM, 3 = ≥ 200 DM, 4 = No checking account.
2	Duration of Credit (month)	Quantitative	Numeric	Duration of credit in months. Values range from 4 to 72 months.
3	Payment Status of Previous Credit	Ordinal	{0,1,2,3,4}	Credit history: 0 = no credits taken, 1 = all credits paid back duly, 2 = existing credits paid back duly till now, 3 = delay in paying off in the past, 4 = critical account.
4	Purpose	Nominal	{0,1,2,3,4,5,6,8,9,10}	Purpose of the loan: 0 = New car, 1 = Used car, 2 = Furniture/Equipment, 3 = Radio/Television, 4 = Domestic Appliances, 5 = Repairs, 6 = Education, 7 = Vacation, 8 = Retraining, 9 = Business, 10 = Others.
5	Credit Amount	Quantitative	Numeric	Amount of credit in Deutsche Mark (DM). Values range from 250 to 18,424 DM.
6	Value Savings/Stocks	Ordinal	{1,2,3,4,5}	Average balance in savings and stocks: 1 = < 100 DM, 2 = 100-500 DM, 3 = 500-1000 DM, 4 = ≥ 1000 DM, 5 = unknown/no savings account.
7	Length of current employment	Ordinal	{1,2,3,4,5}	Employment duration: 1 = unemployed, 2 = < 1 year, 3 = 1-4 years, 4 = 4-7 years, 5 = ≥ 7 years.
8	Instalment per cent	Quantitative	Numeric	Installment rate as percentage of disposable income.
9	Sex & Marital Status	Nominal	{1,2,3,4}	Gender and marital status: 1 = male divorced/separated, 2 = female divorced/separated/married, 3 = male single, 4 = male married/widowed.
10	Guarantors	Nominal	{1,2,3}	Guarantors and co-applicants: 1 = none, 2 = co-applicant, 3 = guarantor.

Index	Attribute	Type	Values/Range	Description
11	Duration in Current address	Ordinal	{1,2,3,4}	Time at current address: 1 = < 1 year, 2 = 1-2 years, 3 = 2-3 years, 4 = ≥ 3 years.
12	Most valuable available asset	Nominal	{1,2,3,4}	Most valuable assets: 1 = real estate, 2 = savings agreement/life insurance, 3 = car or other, 4 = unknown/no property.
13	Age (years)	Quantitative	Numeric	Age of applicant in years.
14	Concurrent Credits	Nominal	{1,2,3}	Installment plans: 1 = bank, 2 = stores, 3 = none.
15	Type of apartment	Nominal	{1,2,3}	Housing type: 1 = rent, 2 = own, 3 = for free.
16	No of Credits at this Bank	Quantitative	Numeric	Number of existing credits at the bank.
17	Occupation	Nominal	{1,2,3,4}	Job category: 1 = unemployed/unskilled - non-resident, 2 = unskilled - resident, 3 = skilled employee/official, 4 = management/self-employed/highly qualified employee.
18	No of dependents	Quantitative	Numeric	Number of people financially dependent on the applicant.
19	Telephone	Nominal	{1,2}	Whether applicant has a telephone: 1 = yes, 2 = no.
20	Foreign Worker	Nominal	{1,2}	Whether applicant is a foreign worker: 1 = yes, 2 = no.

Statistical Summary Report

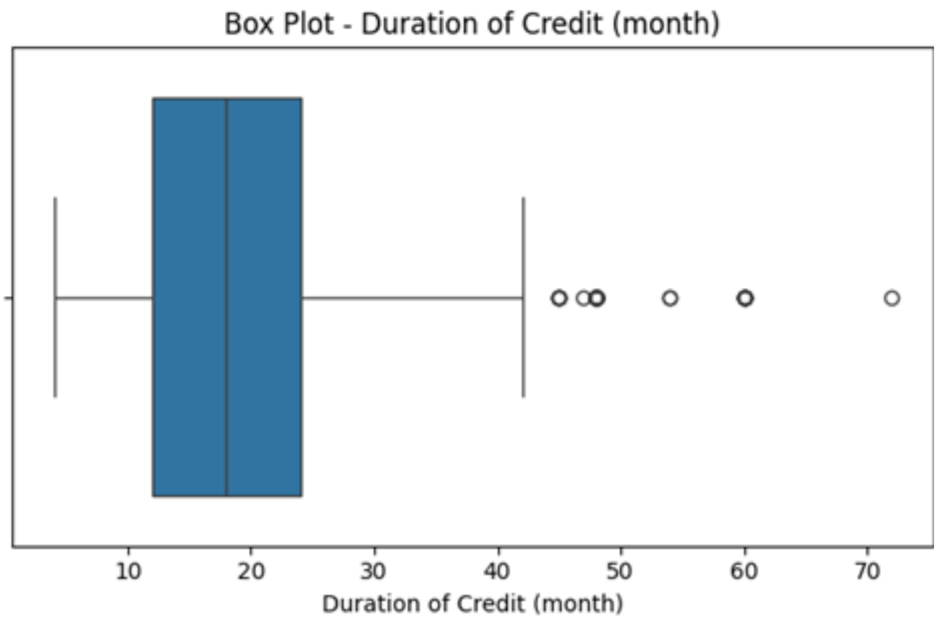
Feature	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Creditability	1000	0.7	0.458	0	0	1	1	1
Account Balance	1000	2.577	1.258	1	1	2	4	4
Duration of Credit (months)	1000	20.903	12.059	4	12	18	24	72
Payment Status of Previous Credit	1000	2.545	1.083	0	2	2	4	4
Purpose	1000	2.828	2.744	0	1	2	3	10
Credit Amount	1000	3271.25	2822.75	250	1365.5	2319.5	3972.3	18424
Value Savings/Stocks	1000	2.105	1.58	1	1	1	3	5
Length of Employment	1000	3.384	1.208	1	3	3	5	5
Instalment Percent	1000	2.973	1.119	1	2	3	4	4
Sex & Marital Status	1000	2.682	0.708	1	2	3	3	4
Duration in Current Address	1000	2.845	1.104	1	2	3	4	4
Most Valuable Asset	1000	2.358	1.05	1	1	2	3	4
Age (years)	1000	35.542	11.353	19	27	33	42	75
Concurrent Credits	1000	2.675	0.706	1	3	3	3	3
Type of Apartment	1000	1.928	0.53	1	2	2	2	3
No. of Credits at Bank	1000	1.407	0.578	1	1	1	2	4
Occupation	1000	2.904	0.654	1	3	3	3	4
No. of Dependents	1000	1.155	0.362	1	1	1	1	2
Telephone	1000	1.404	0.491	1	1	1	2	2
Foreign Worker	1000	1.037	0.189	1	1	1	1	2

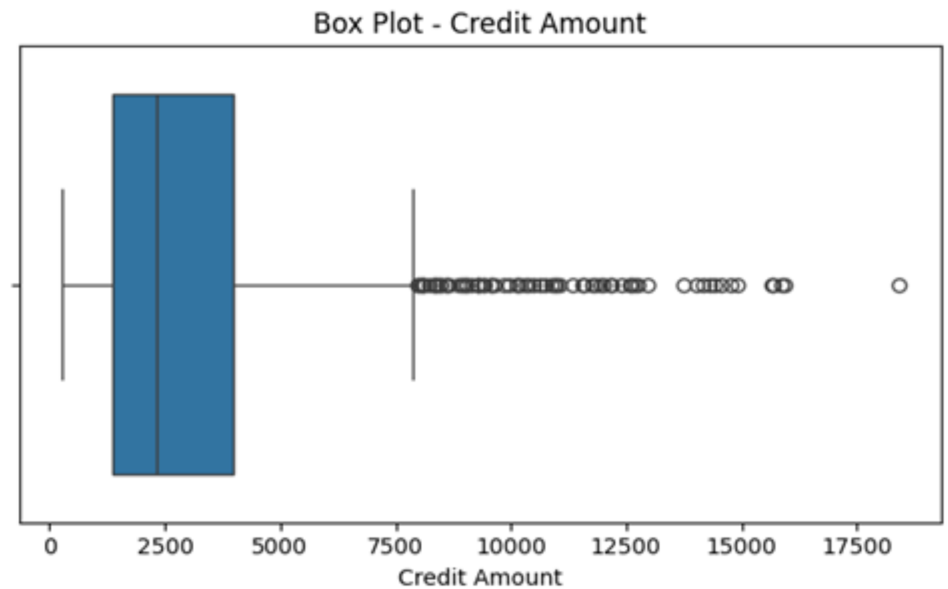
Key Attributes:

- **Creditability (Target Variable):**
  - Mean: 0.7 → **70% of the applicants are classified as good credit risks**.
  - Indicates a class imbalance leaning toward "good" credit customers.
- **Credit Amount:**
  - Mean: **3,271 DM***, but max is **18,424 DM**, and the distribution is **right-skewed**.
  - 75% of loans are **≤ 3,972 DM**, suggesting that **higher-value loans are less common.**
- **Duration of Credit:**
  - Median is 18 months.
  - Mean with the longest term being 72 months.
  - The mean (21) and median (18) suggest that the bank **typically issues short-to-mid-term loans of 1.5-2 years**

*DM - Deutsche Mark (German currency)

Outlier Detection





Most of the duration of credit amounts are **between 5 and 40 months** (whisker ends)

The central value is the median which is about 18 roughly 20 months that is what most applicants are typically requesting from the bank

The dots represent unusually high duration of credit card months (outliers).

Most credit amounts are **small to medium**, but a few people are requesting **very large loans**.

The central value (median) is about 2500 DM ie. this is the typical loan amount of the average borrower

The dots represent unusually high credit amounts. There are a great number of credit amt. outliers compared to the duration of credit

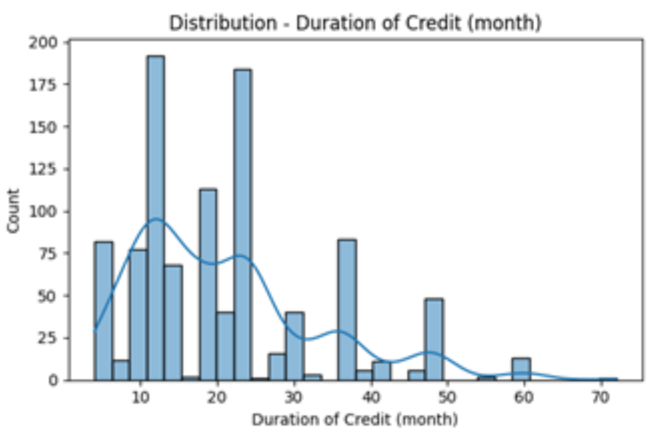
Distribution Visualizations

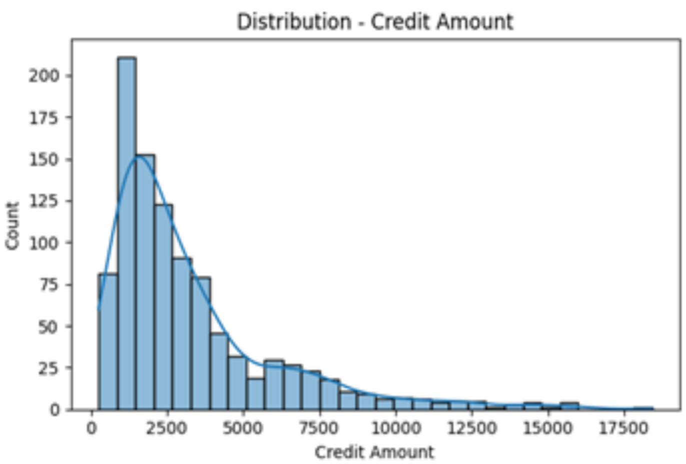
Credit Amount Interpretation

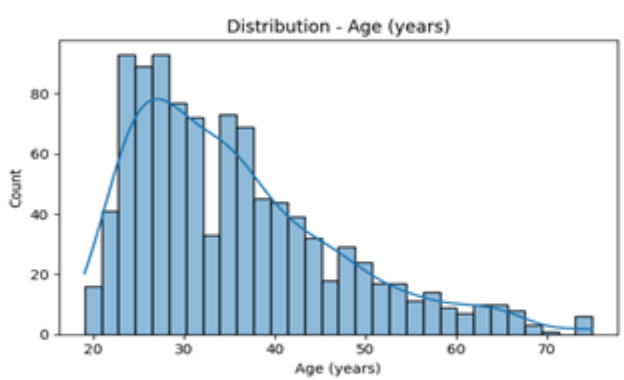
Right skewed distribution therefore **most applicants** are taking **smaller to moderate loans**.

A **few applicants** are applying for **very large loans**, which pull the **mean** to the right.

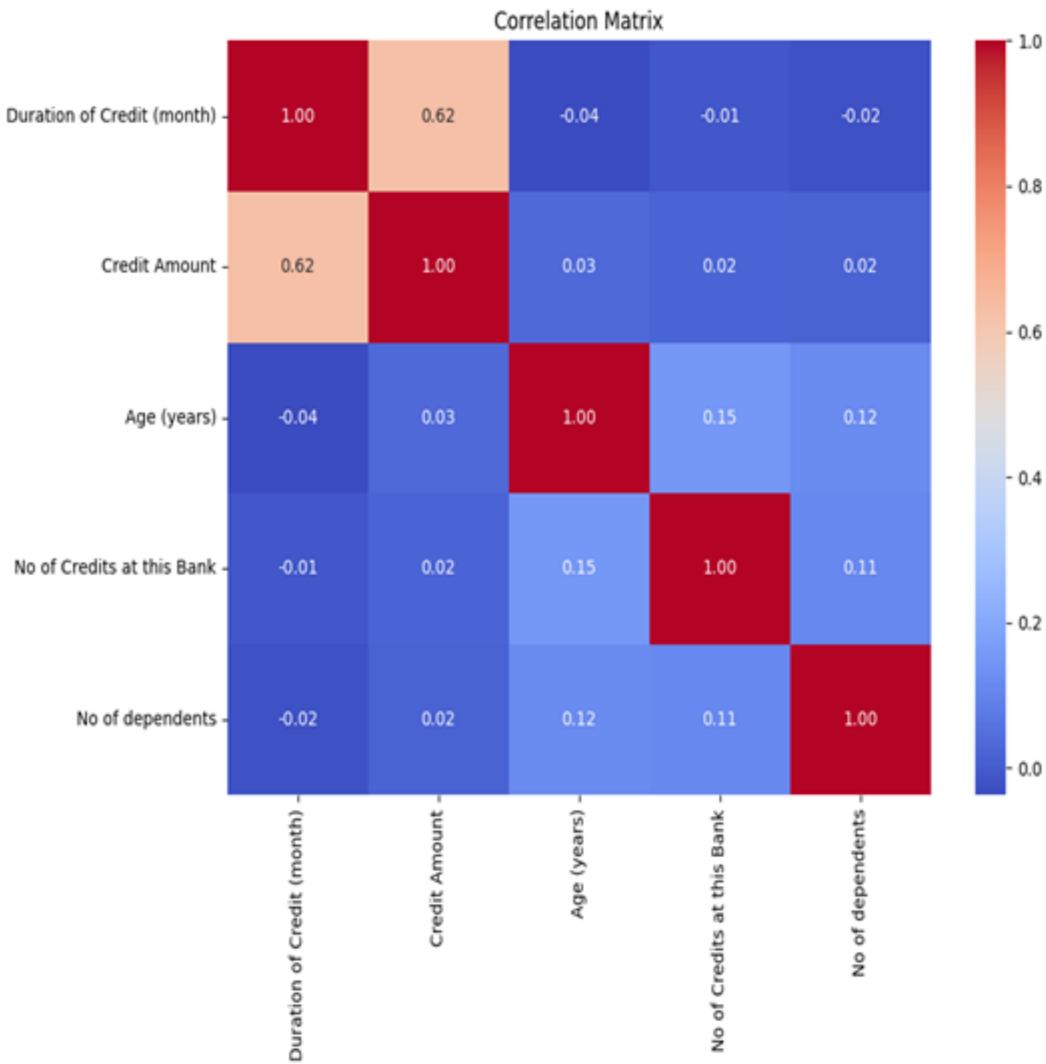
The **mean** is higher than the **median**, and the **tail** extends toward the **larger amounts**.







Correlation Matrix



Key Finding: Credit Duration and Amount have the strongest correlation

- Longer loans are associated with larger borrowed amounts

Strong Correlations:

- Duration of Credit and Credit Amount (0.62): This moderate positive correlation suggests that larger loan amounts are typically associated with longer repayment periods

Weak Correlations:

- Age shows minimal correlation with credit duration (-0.04) and credit amount (0.03), indicating age has little impact on borrowing patterns
- Number of Credits at this Bank has a slight positive correlation with Age (0.15), suggesting older customers might have slightly more credit products
- Number of dependents shows weak positive correlations with Age (0.12) and Number of Credits (0.11)

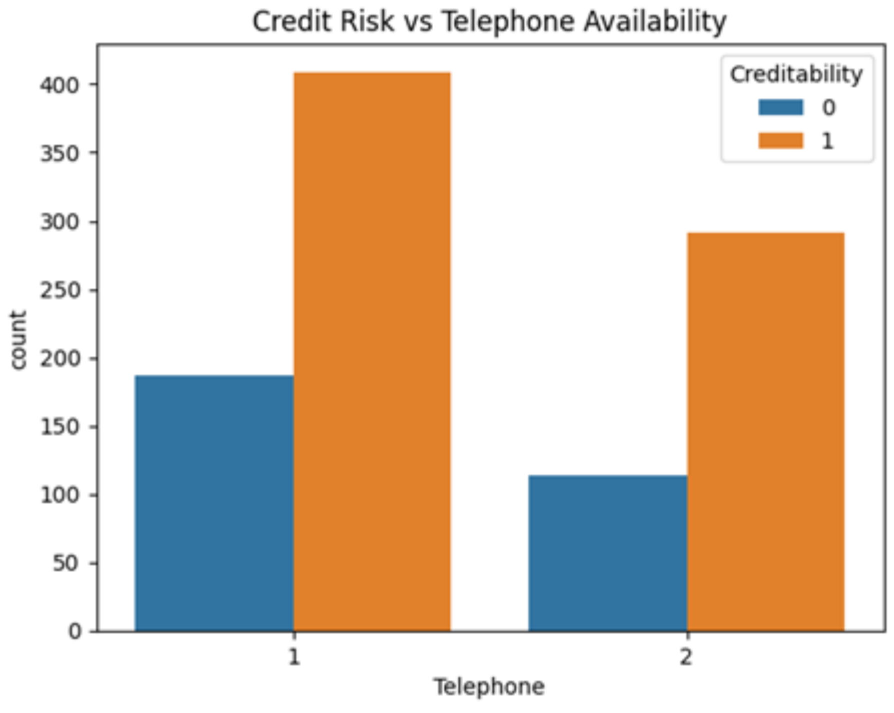
Negligible Correlations:

- Most other relationships in the matrix show correlations close to zero, indicating no meaningful relationships between those variables

Data Preparation

- **Early** in the process, **we considered dropping the Telephone column** as it initially appeared to have limited predictive value. Creditability rates were plotted against telephone status; there was a marked distinction between customers with telephone access and those without.

- That showed a good opportunity for the Telephone variable for meaningful correlations — as a socio-economic indicator and a low-cost/ high-impact attribute when used in concert with other determinants to better the model. Hence, **the Telephone attribute was retained** in the file.



Feature Selection

- We used the **BestFirst search algorithm** with the CFS subset evaluator in Weka to evaluate the following features as most impactful attributes that affect creditworthiness in both the Decision Tree and Naïve Bayes models

1. Account Balance
2. Duration of Credit (month)
3. Payment Status of Previous Credit

- *Creditability → retained because it is the class attribute

- No modifications needed to be made to the data within each selected attribute
- These 3 were chosen because they had the highest merit score (0.0761)

Benefits

1. Highlights critical signals while minimizing redundancies
2. Reduces model overfitting risk by removing noise from less relevant attributes
3. Focuses stakeholder attention for decision-making

Decision Tree Classification Results

Baseline Model

- **Test Set Size:** 340 instances
- **Classification Accuracy:** 72.35% (246 correctly classified, 85 misclassified)
- **Kappa Statistic:** 0.3214
- **Error Metrics:**
  – Mean Absolute Error: 0.3367
  – Root Mean Squared Error: 0.4776
  – Relative Absolute Error: 78.156%
  – Root Relative Squared Error: 99.5787%

Class 0 (Creditability = 0)

- **True Positive Rate (Recall):** 0.39
- **False Positive Rate:** 0.10
- **Precision:** 0.66
- **Recall:** 0.39

Class 1 (Creditability = 1)

- **True Positive Rate:** 0.89
- **False Positive Rate:** 0.60
- **Precision:** 0.73
- **Recall:** 0.89

Confusion Matrix

	Positive	Negative
Positive	46	71
Negative	23	200

Selected Feature Model

- **Test Set Size:** 340 instances
- **Classification Accuracy:** 72.94% (248 correctly classified, 92 misclassified)
- **Kappa Statistic:** 0.3417
- **Error Metrics:**
  -Mean absolute error 0.3581
  -Root mean squared error 0.4392
  -Relative absolute error 83.1278 %
  -Root relative squared error 91.5696 %

Class 0 (Creditability = 0)

- **TP 0.4188034188034188**
- **FP 0.10762331838565023**
- **Precision 0.6712328767123288**
- **Recall 0.4188034188034188**

Class 1 (Creditability = 1)

- **TP 0.8923766816143498**
- **FP 0.5811965811965812**
- **Precision 0.7453183520599251**
- **Recall 0.8923766816143498**

Confusion Matrix

	Positive	Negative
Positive	49	68
Negative	24	199

Naïve Bayes Classification Results

Baseline Model

- **Test Set Size:** 340 instances
- **Classification Accuracy:** 75% (255 correctly classified, 85 misclassified)
- **Kappa Statistic:** 0.3959
- **Error Metrics:**
  – Mean Absolute Error: 0.2972
  – Root Mean Squared Error: 0.4181
  – Relative Absolute Error: 68.99%
  – Root Relative Squared Error: 87.17%

Class 0 (Creditability = 0)

- **True Positive Rate (Recall):** 0.4615
- **False Positive Rate:** 0.0987
- **Precision:** 0.7105
- **Recall:** 0.4615

Class 1 (Creditability = 1)

- **True Positive Rate:** 0.9013
- **False Positive Rate:** 0.5385
- **Precision:** 0.7614
- **Recall:** 0.9013

Confusion Matrix

	Positive	**Negative**
Positive	54	63
Negative	22	201

Selected Feature Model

- **Test Set Size:** 340 instances
- **Classification Accuracy:** 73.82% (251 correctly classified, 89 misclassified)
- **Kappa Statistic:** 0.3381
- **Error Metrics:**
  – Mean Absolute Error: 0.3418
  – Root Mean Squared Error: 0.4279
  – Relative Absolute Error: 79.36%
  – Root Relative Squared Error: 89.21%

Class 0 (Creditability = 0)

- **True Positive Rate (Recall):** 0.4615
- **False Positive Rate:** 0.0987
- **Precision:** 0.7105
- **Recall:** 0.4615

Class 1 (Creditability = 1)

- **True Positive Rate:** 0.9013
- **False Positive Rate:** 0.5385
- **Precision:** 0.7614
- **Recall:** 0.9013

Confusion Matrix

	Positive	**Negative**
Positive	42	75
Negative	14	209

Predictive Modeling

- Evaluation method: 66-34 Train-test Split
- Results (perspective of Class 1)

Model	Accuracy	Precision	Recall
Baseline Decision Tree	72.35%	73.80%	89.69%
Baseline Naïve Bayes	75.00%	76.14%	90.13%
Decision Tree on Selected Features	72.94%	74.53%	89.23%
Naïve Bayes on Selected Features	73.82%	73.59%	93.72%

Conclusion

- Naïve Bayes with selected features performed best with highest recall (93.72%) and strong overall accuracy (73.82%).
- Feature selection improved model focus by using only 3 key predictors:
  - Account Balance, Duration of Credit, and Payment Status of Previous Credit.
- Naïve Bayes is recommended due to its simplicity, speed, and strong performance, especially in identifying high-creditability clients.
- If the main goal is to identify as many trustworthy (Credible = 1) customers as possible, the selected-feature model performs slightly better.
- But if the bank needs a more balanced approach—especially to better catch risky applicants (Credible = 0)—then the baseline model is the safer choice.

Business Recommendation

- While this model is more than sufficient, we recommend investing in improving the model with better features, more advanced algorithms, and using it as a **decision support tool rather than an automatic approval system.**
- With some tuning and additional data, the model can become a valuable asset in reducing risk and streamlining the approval process.
- Do **not** use the model alone for approval decisions or replace human judgment yet

Business Opportunities:

- Adding additional variables for better credit behaviour prediction
- Opportunity to develop more nuanced customer segmentation models
- Potential for tailored product offerings based on duration-amount relationship