Sefirot
Financial Research

Supervised Machine Learning for Credit Default Prediction

Author:
Mariani Federico

January, 2026

Index

1 Objective Description 1

2 Dataset Selection 1

3 Literature and Context of the Case 1

4 Dependent Variable 2

5 Independent Variables 2

6 Exploratory Data Analysis (EDA) 3
6.1 Data Structure 3
6.2 Descriptive Statistics 5
6.2.1 Numerical Variables 5
6.2.2 Categorical Variables 6
6.2.3 Bivariate Analysis 12
6.2.4 Correlation Analysis 14
6.2.5 Correlation categorical variables (Cramer's V) 14

7 Model Selection 15
7.1 Distribution and Link Function 15
7.2 Logistic Regression Assumptions 15
7.3 Transformations 15

8 Predictor Selection 18

9 Split the Data 20

10 Model Fitting and Interpretation 20
10.1 Stepwise Regression 20
10.2 Interpretations of Results 21
10.3 Model Diagnostics 23
10.3.1 Multicollinearity - VIF 23
10.3.2 Cook's Distance 24
10.3.3 Residual Analysis 25
10.3.4 Linearity check 26
10.3.5 Overdispersion 27

11 Model Fit and Performance 28
11.1 Goodness of Fit 28
11.2 Classification Performance 29
11.3 ROC Curve and AUC 29

12 Conclusion 30

1 Objective Description

The primary objective of this analysis is to model, explain, and predict credit risk for loan applicants using the German Credit dataset. Credit risk is defined as the probability that a customer will be a good payer versus a bad payer. The aim of this analysis is to identify which variables (financial, demographic and credit history) are most strongly associated with customer creditworthiness. Based on this, we are going to apply a statistical model to better predict the binary outcome of Credit Risk. In the end, the model's quality is assessed by evaluating both its goodness of fit and its predictive performance through appropriate diagnostic tools.

2 Dataset Selection

We decided to use the German Credit Data, which is a standard dataset employed in machine learning for credit scoring tasks (you can find the dataset here: Dataset). This dataset is composed of: 1.000 observations, 20 predictors (independent variables) and 1 target variable (dependent variable). The data types are mixed (categorical and numerical). We decided to use this dataset because it's well-suited for this kind of analysis, as the target variable is binary, making it useful to apply a generalized linear model with a binomial link function. On top of that, there are a lot of different variables that can allow a global understanding of assessing the creditworthiness.

3 Literature and Context of the Case

Credit risk modeling is an important problem in the finance industry. Logistic regression is still strongly used to assess this kind of problem, despite there being a lot more modern machine learning models, thanks to its interpretability, regulatory acceptance and robustness.

4 Dependent Variable

The dependent variable is Credit Risk. It's defined as: 1 = good payer; 2 = bad payer. The measurement scale is binary and the class distribution is: 70% good payer; 30% bad payer. Classes, as we can see here, are not equally balanced.

Table 1: Data Type - Dependent variable

Variable	Type	Levels	Role
Credit_Risk	Categorical	2	Dependent

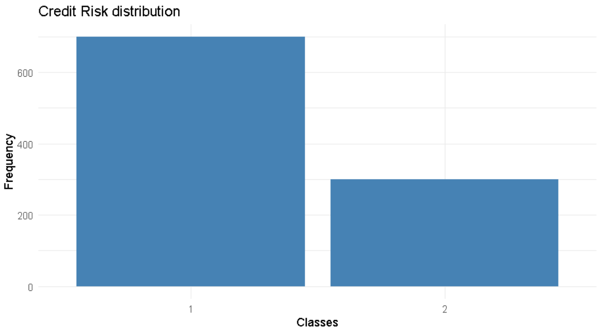

Figure 1: Bar plot of Credit_Risk frequencies

5 Independent Variables

We decided to use all the variables within this dataset, except for the variable ”Telephone”, as it's considered non relevant to assess the creditworthiness. These are the selected variables with their respective data types:

Table 2: Data Type - Independent variables

Variable	Type	Levels	Role
Status_Checking_Acc	Categorical	4	Independent
Duration	Numeric	NA	Independent
Credit_History	Categorical	5	Independent
Purpose	Categorical	10	Independent
Credit_Amount	Numeric	NA	Independent
Savings	Categorical	5	Independent
Employment	Categorical	5	Independent
Installment_Rate	Numeric	NA	Independent
Personal_Status	Categorical	4	Independent
Debtors	Categorical	3	Independent
Residence_Since	Numeric	NA	Independent
Property	Categorical	4	Independent
Age	Numeric	NA	Independent
Other_Installments	Categorical	3	Independent
Housing	Categorical	3	Independent
Existing_Credits	Numeric	NA	Independent
Job	Categorical	4	Independent
People_Liable	Numeric	NA	Independent
Foreign_Worker	Categorical	2	Independent
Credit_Risk_num	Numeric	NA	Independent

6 Exploratory Data Analysis (EDA)

EDA is conducted to understand the structure of the data, identify potential issues, and inform decisions regarding model specification and predictor selection.

6.1 Data Structure

As we said before, this dataset is composed of 1.000 observations and 19 selected independent variables. We only removed the variable ”Telephone”, as considered non relevant to assess the creditworthiness. We observe no missing values. We transformed data type because R needs this kind of data type in order to proceed with the code.

Table 3: *Number of unique values for each variable*

Variable	Type	Unique values
Credit_Risk	Categorical	2
Status_Checking_Acc	Categorical	4
Duration	Numeric	33
Credit_History	Categorical	5
Purpose	Categorical	10
Credit_Amount	Numeric	921
Savings	Categorical	5
Employment	Categorical	5
Installment_Rate	Numeric	4
Personal_Status	Categorical	4
Debtors	Categorical	3
Residence_Since	Numeric	4
Property	Categorical	4
Age	Numeric	53
Other_Installments	Categorical	3
Housing	Categorical	3
Existing_Credits	Numeric	4
Job	Categorical	4
People_Liable	Numeric	2
Foreign_Worker	Categorical	2

Table 4: Data types after their transformation

Variable	Type
Credit_Risk	Factor
Status_Checking_Acc	Factor
Duration	Integer
Credit_History	Factor
Purpose	Factor
Credit_Amount	Integer
Savings	Factor
Employment	Factor
Installment_Rate	Integer
Personal_Status	Factor
Debtors	Factor
Residence_Since	Integer
Property	Factor
Age	Integer
Other_Installments	Factor
Housing	Factor
Existing_Credits	Integer
Job	Factor
People_Liable	Integer
Foreign_Worker	Factor
Credit_Risk_num	Numeric

6.2 Descriptive Statistics

6.2.1 Numerical Variables

These are the statistics of the numerical variables (integer in R):

Table 5: Summary statistics for numeric variables

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Duration	4	12	18	20.9	24	72
Credit_Amount	250	1366	2320	3271	3972	18424
Age	19	27	33	35.55	42	75
Installment_Rate	1	2	3	2.973	4	4
Residence_Since	1	2	3	2.845	4	4
Existing_Credits	1	1	1	1.407	2	4
People_Liable	1	1	1	1.155	1	2

About Duration, we see that Mean > Median, so there's positive asymmetry (skewed to the right). 75% of loans lasts 24 months (3rd Quartile). Long durations exist (up to 72 months) but it's only the 25% of the total loans. About Credit_Amount we see that also here that Mean > Median, so there's positive asymmetry (skewed to the right). As that the Max. are farther from the Median than the Min. to the Median, and as that the Mean is way bigger than the Median, we deduce that Credit_Amount is influenced by big events. Variable Age is skewed to the right and that the most part of the clients are from 27 to 42 years (50% of total clients), while there is a 25% of probability that a client is from 19 to 26 years and a 25% possibility that a client is from 43 to 75 years. Installment_rate ranges from 1 to 4, with a median of 3. Most clients (50%) fall in the middle categories, indicating moderate repayment burdens. Extreme rates (1 or 4) are relatively uncommon. Length of residence (Residence_Since) varies from 1 to 4 years, with a median of 3. This suggests that most clients have moderate residential stability, with a smaller proportion living less than 2 years or more than 4 years at their current address. Number of existing credits (Existing_Credits) ranges from 1 to 4, with most clients holding only 1 active credit. A small minority has multiple concurrent credits, which could slightly increase repayment burden. In the end, number of people financially dependent (People_Liable) on the client is mostly 1, with few clients supporting 2 dependents. This variable is highly concentrated, suggesting limited variation in household financial burden.

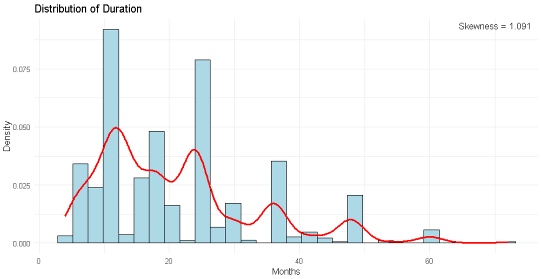

Figure 2: Distribution - Duration

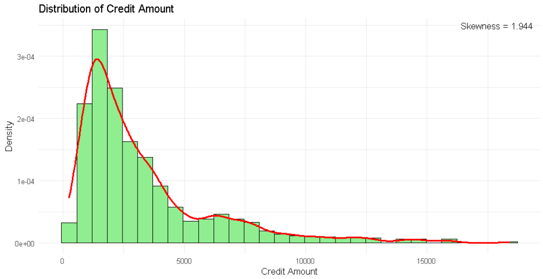

Figure 3: Distribution - Credit_Amount

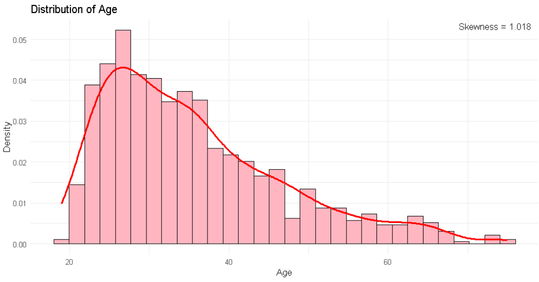

Figure 4: Distribution - Age

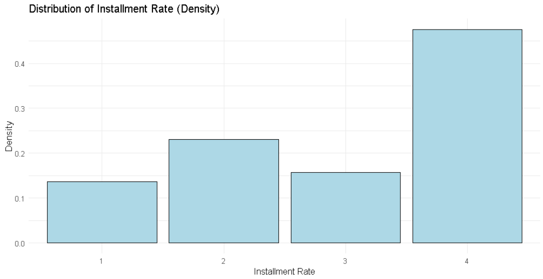

Figure 5: Distribution - Installment_Rate

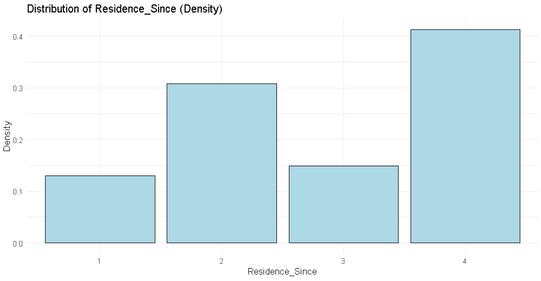

Figure 6: Distribution - Residence_Since

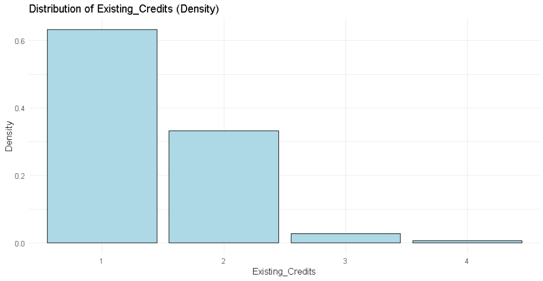

Figure 7: Distribution - Existing_Credits

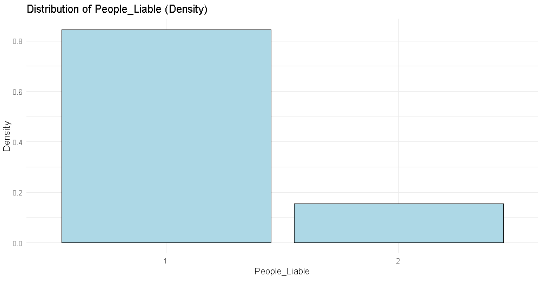

Figure 8: Distribution - People_Liable

6.2.2 Categorical Variables

Here we have the frequency tables for the categorical variables:

Table 6: Frequency Tables - Independent Categorical variables

Variable	Level	Frequency
Status_Checking_Acc	A11	274
	A12	269
	A13	63
	A14	394
Credit_History	A30	40
	A31	49
	A32	530
	A33	88
	A34	293
Employment	A71	62
	A72	172
	A73	339
	A74	174
	A75	253
Housing	A151	179
	A152	713
	A153	108
Purpose	A40	234
	A41	103
	A410	12
	A42	181
	A43	280
	A44	12
	A45	22
	A46	50
	A48	9
	A49	97
Savings	A61	603
	A62	103
	A63	63
	A64	48
	A65	183

Table 7: Frequency Tables - Independent Categorical variables

Variable	Level	Frequency
Personal_Status	A91	50
	A92	310
	A93	548
	A94	92
Debtors	A101	907
	A102	41
	A103	52
Property	A121	282
	A122	232
	A123	332
	A124	154
Other_Installments	A141	139
	A142	47
	A143	814
Job	A171	22
	A172	200
	A173	630
	A174	148
Foreign_Worker	A201	963
	A202	37

According to the source of the dataset:

Table 8: Description of categorical variables used in the analysis

Variable	**Description**	**Categories / Levels**
Status_Checking_Ac	Status of existing checking account	A11: balance < 0 DM A12: $0 \leq$ balance < 200 DM A13: balance ≥ 200 DM / salary assignments A14: no checking account
Credit_Risk	Creditworthiness outcome	1: good credit risk 2: bad credit risk
Credit_History	Past credit repayment behavior	A30: no credits / all paid back duly A31: all credits at this bank paid back duly A32: existing credits paid back duly A33: delay in paying off A34: critical account / other credits
Employment	Present employment duration	A71: unemployed A72: < 1 year A73: 1–4 years A74: 4–7 years A75: ≥ 7 years
Housing	Housing status	A151: rent A152: own A153: for free
Purpose	Purpose of the loan	A40: car (new) A41: car (used) A42: furniture/equipment A43: radio/television A44: domestic appliances A45: repairs A46: education A48: retraining A49: business A410: others
Savings	Savings account / bonds	A61: < 100 DM A62: 100–499 DM A63: 500–999 DM A64: ≥ 1000 DM A65: unknown / no savings
Personal_Status	Personal status and sex	A91: male, divorced/separated A92: female, divorced/separated/married A93: male, single A94: male, married/widowed

Table 9: Description of categorical variables used in the analysis

Variable	**Description**	**Categories / Levels**
Status_Checking_Acct	Status of existing checking account	A11: balance < 0 DM A12: $0 \leq$ balance < 200 DM A13: balance ≥ 200 DM / salary assignments A14: no checking account
Credit_Risk	Creditworthiness outcome	1: good credit risk 2: bad credit risk
Credit_History	Past credit repayment behavior	A30: no credits / all paid back duly A31: all credits at this bank paid back duly A32: existing credits paid back duly A33: delay in paying off A34: critical account / other credits
Employment	Present employment duration	A71: unemployed A72: < 1 year A73: 1–4 years A74: 4–7 years A75: ≥ 7 years
Housing	Housing status	A151: rent A152: own A153: for free
Purpose	Purpose of the loan	A40: car (new) A41: car (used) A42: furniture/equipment A43: radio/television A44: domestic appliances A45: repairs A46: education A48: retraining A49: business A410: others
Savings	Savings account / bonds	A61: < 100 DM A62: 100–499 DM A63: 500–999 DM A64: ≥ 1000 DM A65: unknown / no savings
Personal_Status	Personal status and sex	A91: male, divorced/separated A92: female, divorced/separated/married A93: male, single A94: male, married/widowed

Here we see the distributions of the categorical variables:

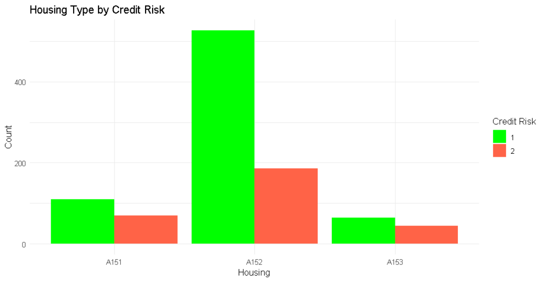

Figure 9: Barplot - Housing_Type by Credit_Risk

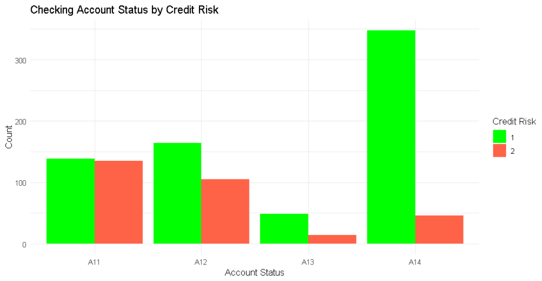

Figure 10: Barplot - Status_Checking_Acc by Credit_Risk

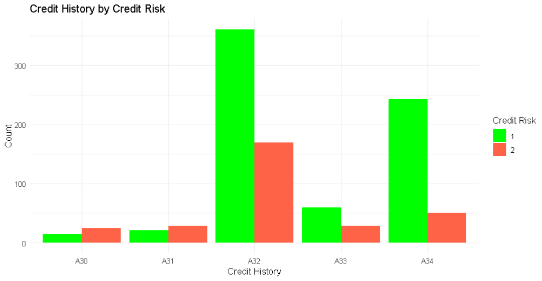

Figure 11: Barplot - Credit_History by Credit_Risk

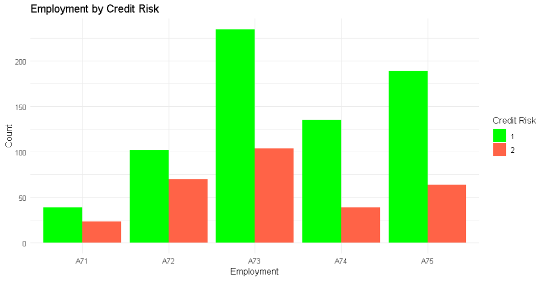

Figure 12: Barplot - Employment by Credit_Risk

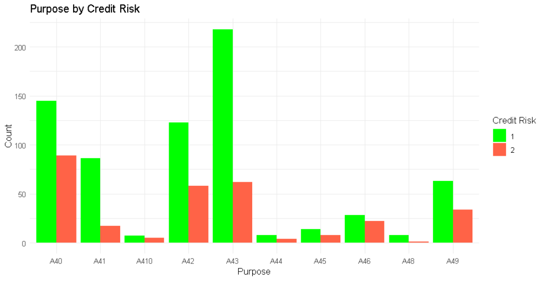

Figure 13: Barplot - Purpose by Credit_Risk

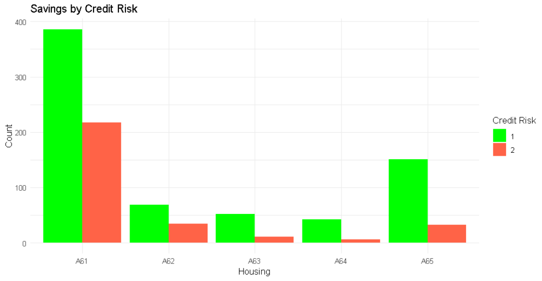

Figure 14: Barplot - Savings by Credit_Risk

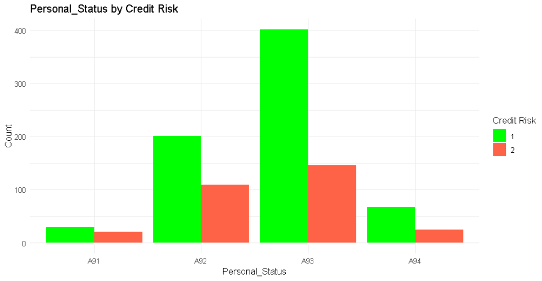

Figure 15: Barplot - Personal_Status by Credit_Risk

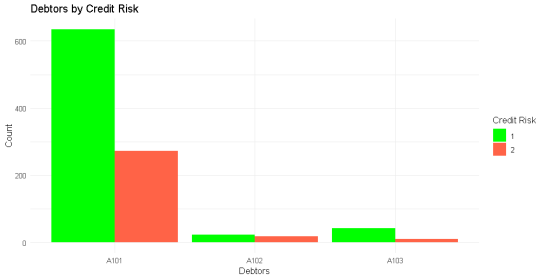

Figure 16: Barplot - Debtors by Credit_Risk

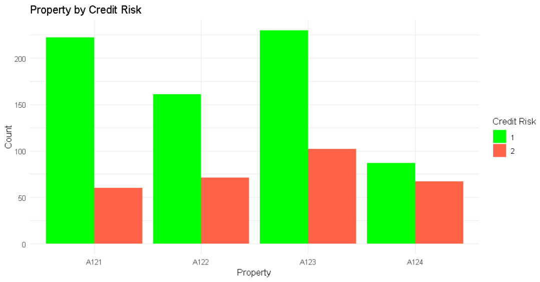

Figure 17: Barplot - Property by Credit_Risk

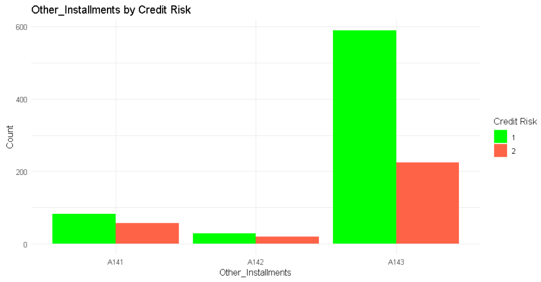

Figure 18: Barplot - Other_Installments by Credit_Risk

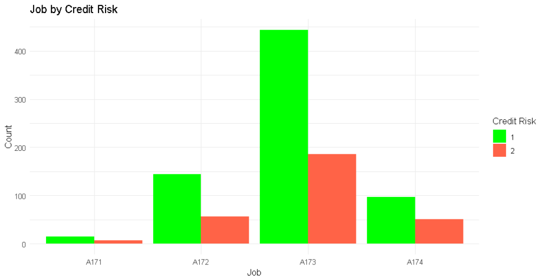

Figure 19: Barplot - Job by Credit_Risk

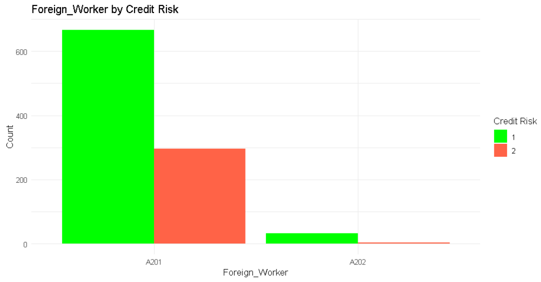

Figure 20: Barplot - Foreign_Worker by Credit_Risk

6.2.3 Bivariate Analysis

Bivariate analysis highlights relationships between predictors and the outcome. As we can see from the scatterplot below, credit risk is very dependent on the duration and the credit amount of the loan, as that we see more red points in the rightmost part of the graph.

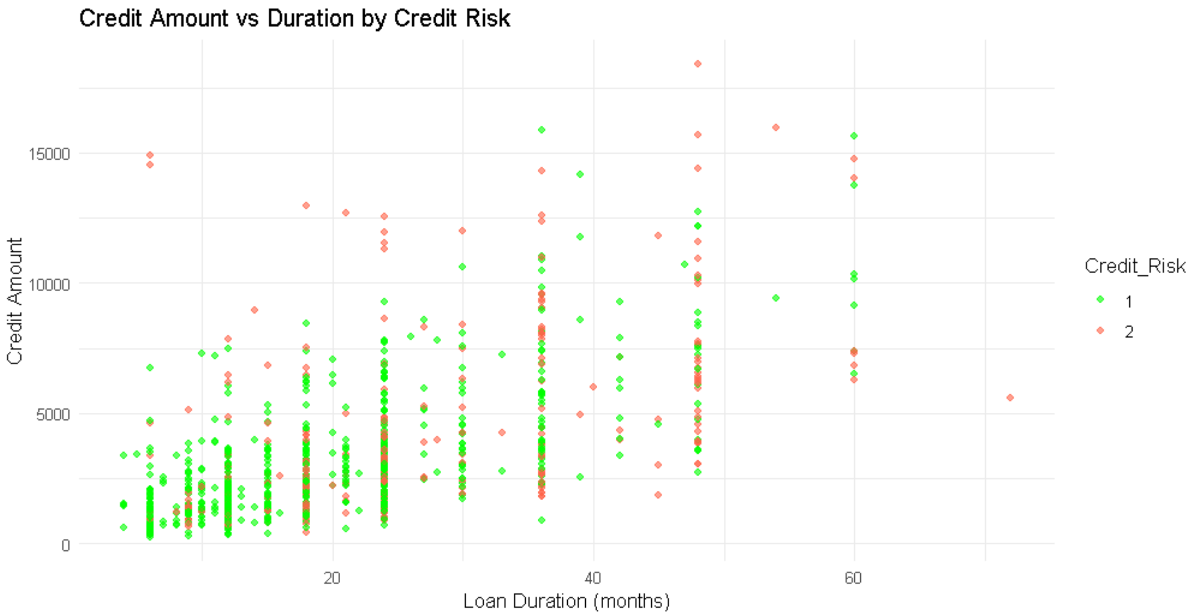

Figure 21: Scatterplot - Credit_Amount vs Duration by Credit_Risk

Boxplot of each numeric variable stratified by Credit_Risk:

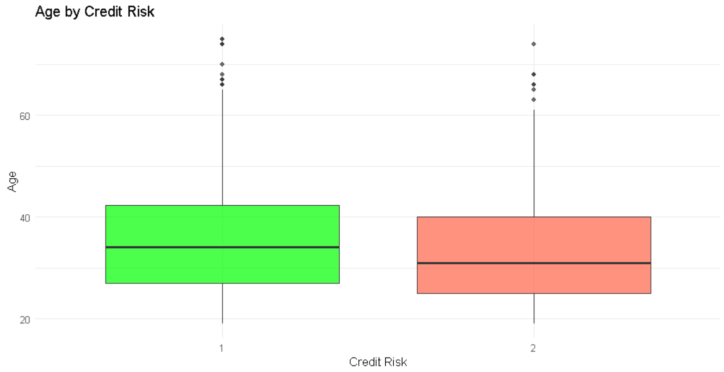

Figure 22: Boxplot - Age vs Credit_Risk

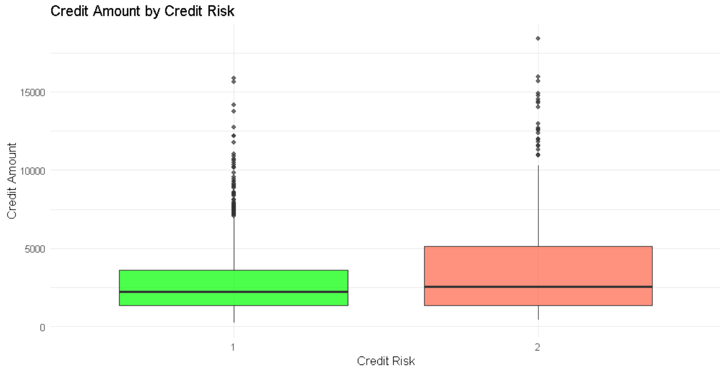

Figure 23: Boxplot - Credit_Amount vs Credit_Risk

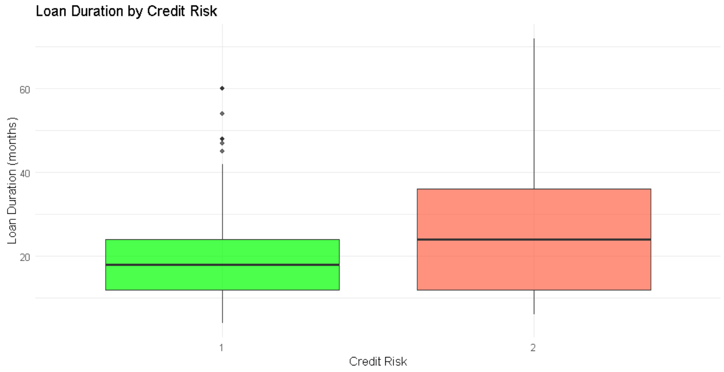

Figure 24: Boxplot - Duration vs Credit_Risk

6.2.4 Correlation Analysis

Among the selected independent variables, we can see from the heatmap that there are low levels of correlation, but the one between Duration and Credit_Amount that shows a higher level of correlation with respect to the others.

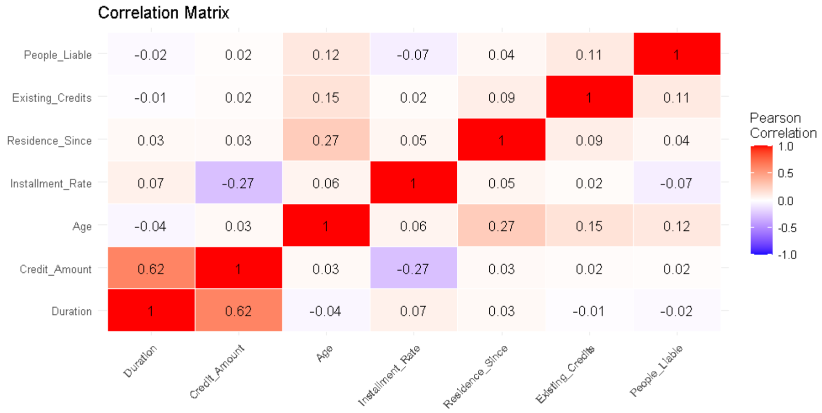

Figure 25: Correlation Heatmap

6.2.5 Correlation categorical variables (Cramer's V)

Associations among categorical predictors were assessed using Cramér's V. Overall, the results indicate weak associations for most variable pairs, suggesting low multicollinearity among categorical predictors. The strongest association was observed between Property and Housing (V = 0.55), reflecting their conceptual overlap in describing housing and asset ownership. A moderate association was also found between Employment and Job (V = 0.31), indicating partial redundancy between employment stability and job type. All remaining associations were weak (V < 0.30), supporting the inclusion of these variables in the same model.

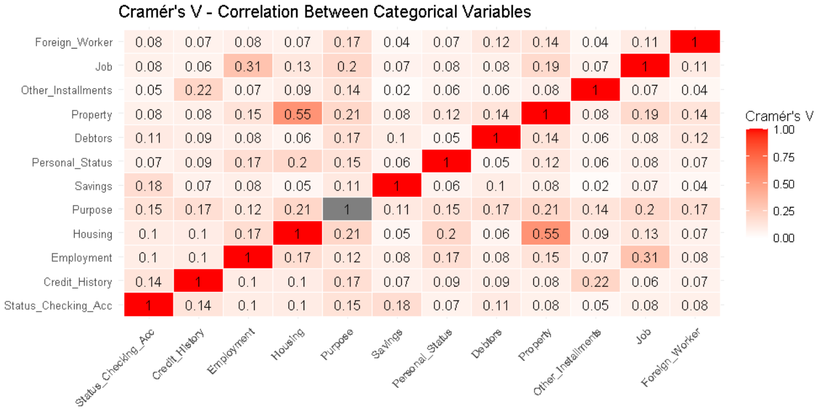

Figure 26: Cramer's V - Correlation between Categorical Variables

7 Model Selection

7.1 Distribution and Link Function

The dependent variable is binary, so we have a Binomial distribution. So, a Generalized Linear Model (GLM) with logit link is therefore appropriate.

7.2 Logistic Regression Assumptions

These are the assumptions that we have to satisfy in order to apply the Logistic Regression model:

1. Dependent Variable is binary (0 and 1): as we have seen above, we satisfy this assumption;

2. Independence of observations: we can verify this by summing duplicated ID rows & columns, which should be equal to 0. Also this assumption is satisfied;

3. Logit linearity for numeric predictors: Box–Tidwell tests indicate non-linear relationships between numeric predictors and the logit of the outcome. In this method, the null hypothesis H_0 is that all the = 1. says how the variable should be transformed to become linear in the model (e.g., if = 0, a log(X) transformation is needed). Results from the Box–Tidwell test indicated significant departures from linearity for all continuous predictors. Specifically, Duration (= 0.09, p = .034), Credit Amount (= 5.41, p < .001), and Age (= −2.86, p = .011) showed statistically significant deviations from the assumption of linearity in the logit. The global score test further rejected the null hypothesis that all parameters equal 1, F(3, 993) = 8.46, p < .001, indicating that at least one continuous predictor violates the linearity assumption. Taken together, these results suggest that the linearity assumption is not satisfied and that transformations of the continuous variables are required.

Table 10: MLE of λ and Score test results

Variable	MLE of λ	Score Statistic (t)	p-value
Duration	0.0928	-2.1202	0.0342*
Credit_Amount	5.4077	4.0967	4.53×10^{-5}***
Age	-2.8637	2.5469	0.0110*

Score test for the null hypothesis that all $\lambda = 1$: $F(3, 993) = 8.4613$, $p = 1.487 \times 10^{-5}$.

7.3 Transformations

As we have seen in the previous chapter, we have to transform the variables in order to apply the logistic regression model. Based on the values of the lambda: polynomial terms of degree 2 were introduced for Duration and Age, while a logarithmic transformation was applied to Credit_Amount. Now we confront each variable with the logic before and after the transformations:

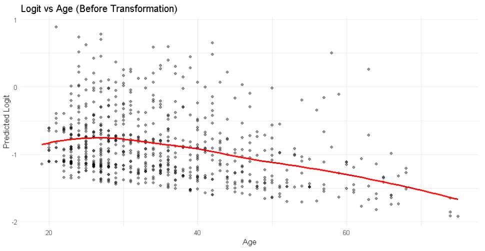

Figure 27: Logit vs Age (Before Transformation)

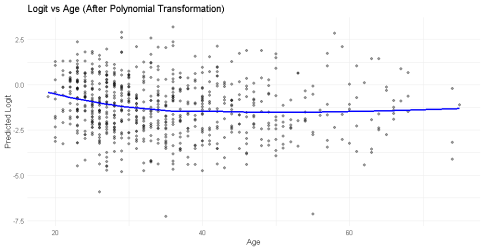

Figure 28: Logit vs Age (After Transformation)

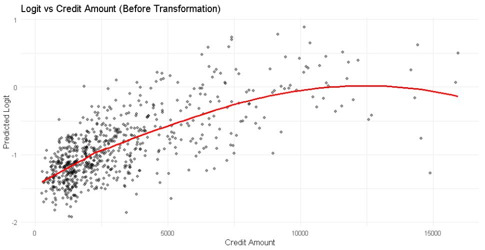

Figure 29: Logit vs Credit_Amount (Before Transformation)

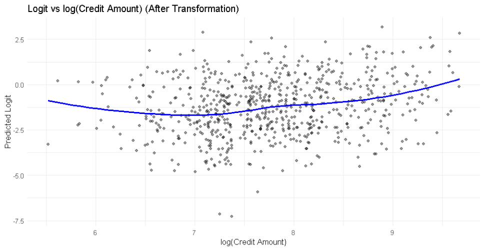

Figure 30: Logit vs Credit_Amount (After Transformation)

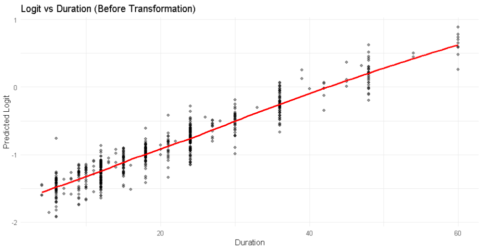

Figure 31: Logit vs Duration (Before Transformation)

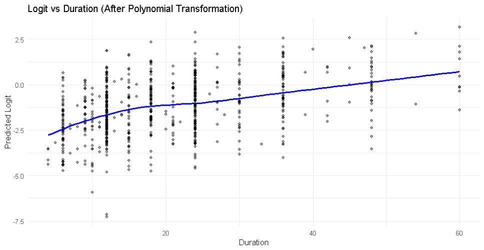

Figure 32: Logit vs Duration (After Transformation)

In the case of age, the graph before the transformation shows a clearly nonlinear relationship with the logit. A peak is observed around age 30, followed by a decline, indicating that the effect of age on the outcome changes direction over the course of a person's life. This behavior suggests that age does not have a monotonous influence on risk (or estimated probability), but rather has a more complex effect. Introducing a polynomial term (Age^2) allows the model to capture this curvature. After the transformation, the estimated line is much closer to the data structure and more stable on the logit scale, indicating that the initially observed nonlinearity has been effectively modeled.

About Credit_Amount, before the transformation, the graph shows a marked nonlinearity in the
relationship between the credit amount and the logit. Specifically, as the amount increases, a rather rapid initial growth of the logit is observed, followed by a stabilization phase or a slight decrease for values above approximately 10,000. This trend clearly suggests that assuming a simple linear relationship between Credit Amount and logit would be inappropriate. After applying the logarithmic transformation, the distribution of points is more balanced and less skewed, a typical characteristic of monetary data. Although the estimated line still shows slight fluctuations, the logarithmic transformation has helped make the relationship more manageable for the model, attenuating the effect

of extreme values.

Duration already shows a nearly linear relationship with the logit, as suggested by the nearly straight red line. This indicates that duration is a predictor with a strong and well-defined linear effect. Even after introducing a polynomial term, the shape of the relationship remains essentially unchanged, confirming that there were no significant nonlinear components to correct. However, a change in the scale of the predicted values or residuals is noted, indicating that the addition of the polynomial term nevertheless contributed to an overall recalibration of the model.

8 Predictor Selection

Table 11: Description of predictors included in the model

Predictor	Type	Explanation
Duration	Numeric (poly)	Loan duration determines the length of the repayment horizon. Longer loans generally increase default risk due to prolonged exposure to income shocks and economic uncertainty. Exploratory analysis and the Box–Tidwell test indicate a nonlinear relationship with the logit, justifying a polynomial specification.
Credit_Amount	Numeric (log)	The loan amount reflects the financial burden imposed on the borrower. Larger credit amounts are associated with higher repayment pressure and increased default risk. The variable is strongly right-skewed, and a nonlinear relationship with the outcome was detected, motivating a log-transformation.
Age	Numeric (poly)	Age serves as a proxy for life-cycle effects, income stability, and financial maturity. Both younger and older borrowers may exhibit higher risk profiles. Nonlinear effects were observed and confirmed through formal testing, supporting a polynomial specification.
Status_Checking_Acc	Categorical	Checking account status captures short-term liquidity and financial discipline. Lower balances or the absence of a checking account are typically associated with higher credit risk, making this a key indicator of financial behavior.
Credit_History	Categorical	Past repayment behavior is one of the strongest predictors of future credit risk. This variable summarizes previous credit outcomes and directly reflects borrower reliability.

Table 12: Description of predictors included in the model

Predictor	Type	Explanation
Purpose	Categorical	The purpose of the loan differentiates between types of expenditures (e.g., car, education, household goods), which may be associated with heterogeneous risk levels due to varying returns and necessity.
Savings	Categorical	Savings account status proxies accumulated wealth and precautionary financial behavior. Higher savings generally indicate a greater ability to absorb income shocks and reduce default risk.
Employment	Categorical	Employment duration reflects job stability and income security. Longer employment histories are typically associated with lower default risk. The variable is retained for its strong economic interpretation, even when statistical effects are moderate.
Installment_Rate	Numeric (ordinal)	The installment rate represents the proportion of income devoted to loan repayment. Higher rates imply tighter budget constraints and increased financial stress, potentially raising default risk.
Personal_Status	Categorical	Personal and marital status captures household composition and financial responsibilities. Differences in risk may arise due to income pooling, dependents, or social support structures.
Debtors	Categorical	The presence of co-applicants or guarantors reflects shared liability. Additional debtors may either reduce risk through risk-sharing or indicate higher underlying credit risk.
Residence_Since	Numeric (ordinal)	Length of residence proxies residential stability and attachment to a location. More stable housing situations are generally associated with lower credit risk.
Property	Categorical	Property ownership indicates long-term wealth and collateral availability. Borrowers owning property are typically considered less risky.
Other_Installments	Categorical	This variable indicates whether the borrower has other installment plans. Multiple concurrent financial obligations may increase repayment burden and default probability.

Table 13: Description of predictors included in the model

Predictor	Type	Explanation
Housing	Categorical	Housing status (renting vs. owning) proxies long-term stability and wealth accumulation, both relevant for creditworthiness assessment.
Existing_Credits	Numeric (ordinal)	The number of existing credits captures current indebtedness. A higher number of outstanding loans may signal increased financial strain and elevated default risk.
Job	Categorical	Job category reflects skill level and income stability. Higher-skilled or permanent employment positions are generally associated with lower credit risk.
People_Liable	Numeric (binary)	The number of people financially dependent on the borrower measures household burden. More dependents may reduce disposable income and increase default risk.
Foreign_Worker	Categorical	This variable may capture labor market vulnerability or institutional barriers. Its interpretation should be handled cautiously and framed as a socioeconomic proxy rather than a causal factor.

9 Split the Data

We decided to split the data in: 70% of the data = train_data; 30% of the data = test_data. So in total we have 700 observations for train_data and 300 observations for test_data. We also divided the dataset in a way to maintain the same classes proportion of the dependent variable (1 = good payer (70%); 2 = bad payer (30%)).

10 Model Fitting and Interpretation

We fitted a logistic regression model using transformed numeric predictors and categorical variables. In a logistic regression (logit) model, we estimate the logit function:

$$\text{logit}(P) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

where $P(Y=1)$ represents the probability that the event occurs. In our case, $Y=1$ corresponds to a good payer. The coefficients β_j represent the estimated effect of each predictor X_j on the log odds of the outcome.

10.1 Stepwise Regression

To build the optimal logistic regression model, we decided to use the Stepwise regression, in order to delete the removes one variable at a time, choosing the one that reduces the AIC the most. The Akaike Information Criterion (AIC) is a statistical metric used to compare and select the best-fitting model from a set of candidates, balancing model accuracy (goodness-of-fit) with simplicity (parsimony) to prevent overfitting.

Table 14: Stepwise regression: comparison of AIC at start and end of procedure

Variable	AIC if removed (Start)	AIC if removed (End)
poly(Duration, 2)	738.66	729.08
log(Credit_Amount)	728.71	-
poly(Age, 2)	726.94	-
Installment_Rate	728.99	-
Residence_Since	728.28	-
Existing_Credits	728.42	-
People_Liable	728.89	-
Status_Checking_Acc	730.26	745.59
Credit_History	735.76	711.61
Employment	729.74	704.93
Housing	729.57	704.74
Purpose	741.29	716.05
Savings	736.20	710.44
Personal_Status	726.30	-
Debtors	734.05	709.07
Property	725.49	-
Other_Installments	736.57	711.49
Job	725.05	-
Foreign_Worker	735.81	709.76

Note: The table shows the impact of removing each predictor on the AIC at the start and end of the stepwise regression procedure. Variables not present in the final model are indicated with ”-”. The procedure reduced the AIC from 730.26 (start) to 703.89 (end), indicating improved model fit.

10.2 Interpretations of Results

Interpretation is supported using odds ratios, which quantify the multiplicative effect of predictors on the odds of being a good payer. Odds ratios indicate how much a predictor is associated with a result. Example: we have two groups, the first one doesn’t smoke, while the second one does and we want to understand how much the variable ”smoke” is associated with a lungs cancer. If OR (Odds Ratio) is > 1, we assume that smoking is associated with a percentage increase in facing a lungs cancer. If OR is for example 3, we see that smokers are 3 times more likely to get lungs cancer than non-smokers. As we said in the previous paragraph, we have the coefficients β_j that amplify the effect of each predictor X_j on the logit outcome. From these coefficients, we can calculate the odds ratios as:

$$\text{OR}_j = e^{\beta_j}$$

An OR > 1 indicates that an increase in the predictor X_j increases the odds of being a good payer, while an OR < 1 indicates that an increase in X_j decreases the odds. Now we can interpret the results of our model.

Table 15: Odds ratios and 95% confidence intervals from the logistic regression model

Variable	Odds Ratio	CI Lower	CI Upper
(Intercept)	17.950	4.241	80.459
poly(Duration, 2)$_1$	3.78e+06	14285.916	1.20e+09
poly(Duration, 2)$_2$	0.181	0.001	35.885
Status_Checking_Acc A12	0.713	0.432	1.172
Status_Checking_Acc A13	0.289	0.119	0.653
Status_Checking_Acc A14	0.189	0.110	0.321
Credit_History A31	0.780	0.205	2.886
Credit_History A32	0.447	0.155	1.218
Credit_History A33	0.353	0.110	1.070
Credit_History A34	0.213	0.071	0.608
Employment A72	1.280	0.531	3.151
Employment A73	1.005	0.435	2.379
Employment A74	0.462	0.177	1.207
Employment A75	0.961	0.408	2.319
Housing A152	0.565	0.339	0.942
Housing A153	0.590	0.274	1.263
Purpose A41	0.266	0.117	0.574
Purpose A410	0.089	0.011	0.505
Purpose A42	0.483	0.260	0.886
Purpose A43	0.426	0.243	0.739
Purpose A44	0.115	0.005	1.200
Purpose A45	2.039	0.528	7.942
Purpose A46	1.259	0.508	3.107
Purpose A48	0.237	0.010	2.081
Purpose A49	0.626	0.289	1.331
Savings A62	0.764	0.388	1.469
Savings A63	0.683	0.287	1.514
Savings A64	0.201	0.042	0.691
Savings A65	0.408	0.218	0.736
Debtors A102	2.587	0.953	7.094
Debtors A103	0.339	0.118	0.873
Other_Installments A142	0.593	0.233	1.487
Other_Installments A143	0.394	0.229	0.676
Foreign_Worker A202	0.148	0.022	0.595

Considering the transformed variables, we can say that: ORs of each single term of the poly variables don't have practical interpretation (we should focus on the Logit vs Duration graph (see Figures 32)). Considering the categorical variables, we can say that: a client who is in the category Status_Checking_Acc A14 has a reduction of approximately 81% (1 - 0.189) in the odds of being a good payer relative to its category; a client who is in the category Credit_History A34 has a reduction of approximately 79% (1 - 0.213) in the odds of being a good payer relative to its category; a client who is in the category

Employment A74 has a reduction of approximately 54% (1 - 0.462) in the odds of being a good payer relative to its category; a client who is in the category Housing A152 has a reduction of approximately 43% (1 - 0.565) in the odds of being a good payer relative to its category. But what would happen if we consider the ORs > 1? It'd happen the opposite thing: for example, a client who is in the category Purpose A45 increases the likelihood of being a good payer by 100%. But what about the other element CI? CI is the confidence interval of the odds ratio and represents the range of values within which the true odds ratio is expected to lie with 95% confidence (p-value < 0,05). If the CI includes 1, the effect is not statistically significant at the 5% level (this means that we can't say with certainty that the variable affects the probability of being a good payer, as that the effect observed in the sample could be due to chance). If CI includes 1, there's no significant effect; if CI_Lower is > 1, there's a positive significant effect; if CI_Higher is < 1, there's a negative significant effect. As for the OR, we don't interpret the singular CI of each poly variable, as that the interval is too large (see Figure 32). For example:

- Debtors A102: OR = 2.587; CI = [0.953 : 7.094] means that it's not significant and it doesn't improve the odds because CI contains 1;
- Savings A62: OR = 0.764, CI = [0.388 : 1.469] means that it's not significative.

Overall, the results suggest that credit risk is mainly driven by liquidity conditions and past credit behavior. Variables such as checking account status, credit history, and savings exhibit strong and statistically significant effects, highlighting the importance of financial discipline and repayment experience. Loan duration shows a clear non-linear relationship with credit risk and should therefore be interpreted through the overall shape of the estimated function rather than individual coefficients. Employment status, housing conditions, and loan purpose play a weaker and less consistent role, acting as secondary factors that nonetheless provide complementary economic information.

10.3 Model Diagnostics

10.3.1 Multicollinearity - VIF

We evaluated the adequacy of the logistic regression model using standard diagnostic procedures. Multicollinearity was assessed through the Variance Inflation Factor (VIF) for each predictor. All VIF values were found to be well below commonly used critical thresholds (< 5), indicating that multicollinearity is not a concern and that the predictors provide independent information.

Table 16: Generalized Variance Inflation Factors (GVIF)

Variable	GVIF	Df	$\text{GVIF}^{1/(2\cdot Df)}$
poly(Duration, 2)	1.325	2	1.073
Status_Checking_Acc	1.290	3	1.043
Credit_History	1.552	4	1.056
Employment	1.391	4	1.042
Housing	1.335	2	1.075
Purpose	2.143	9	1.043
Savings	1.368	4	1.040
Debtors	1.242	2	1.056
Other_Installments	1.262	2	1.060
Foreign_Worker	1.081	1	1.040

10.3.2 Cook's Distance

Influence diagnostics based on Cook's distance indicate that no single observation has an excessive impact on the model estimates. In particular, all Cook's distance values remain below commonly used cutoff thresholds (we used $4/n$), which are standard criteria for identifying influential points. Although a small number of observations exhibit relatively high leverage, they are not simultaneously associated with large residuals; as a result, their Cook's distance values remain low. Since influential observations typically combine both high leverage and large residuals, the absence of such cases supports the robustness of the logistic regression model and indicates that parameter estimates are not driven by individual data points.

Table 17: Summary of influential observations and diagnostics

Diagnostic	Threshold / Rule	Influential Observations	Total Observations
DFBETAS (all coefficients)	$\lvert DFBETA\rvert > 2/\sqrt{n}$	See detailed table	700
DFFITS	$\lvert DFFITS\rvert > 2\sqrt{p/n}$	10	700
Covariance ratio (cov.r)	$> 1 \pm 3p/n$	30	700
Cook's distance	> 1	0	700
Hat values (leverage)	$> 2p/n$	19	700

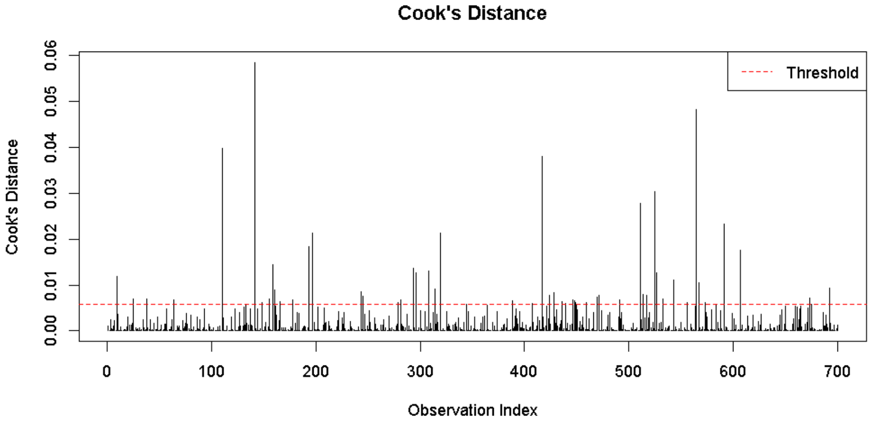

Figure 33: Cook's distance

Table 18: Summary of Cook's distance for logistic regression model

Statistic	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Cook's Distance	1.90e-08	6.888e-05	4.458e-04	1.890e-03	1.766e-03	5.849e-02

Note: The maximum Cook's distance is 0.0585, well below the common threshold of 1, indicating that no single observation has a problematic influence on the logistic regression estimates.

10.3.3 Residual Analysis

The binned residual plot shows the relationship between the probabilities predicted by the logistic model and the average residuals calculated for groups of observations with similar probabilities. The X-axis shows the estimated probabilities, while the Y-axis shows the average residuals (i.e., the difference between the observed and predicted values). The two gray lines indicate the 95% confidence intervals (±2 standard deviations), providing a reference for assessing significant discrepancies between data and predictions. The graph shows that most data points fall within the confidence intervals, with no obvious outliers, and that the average residuals oscillate around the zero line without showing systematic patterns, U-shaped curves, or monotonic trends. This behavior indicates that the model doesn't systematically overestimate or underestimate the predicted probabilities and that the

transformations applied to the variables (logarithm of the credit_amount or the polynomial term on age and on duration) have effectively linearized the relationships with the logit. In the end, we can say that the graph confirms that the model is well-specified and reliable: the residuals behave like random noise, with no evidence of structural errors, systematic discrepancies, or influence points that could compromise the estimates.

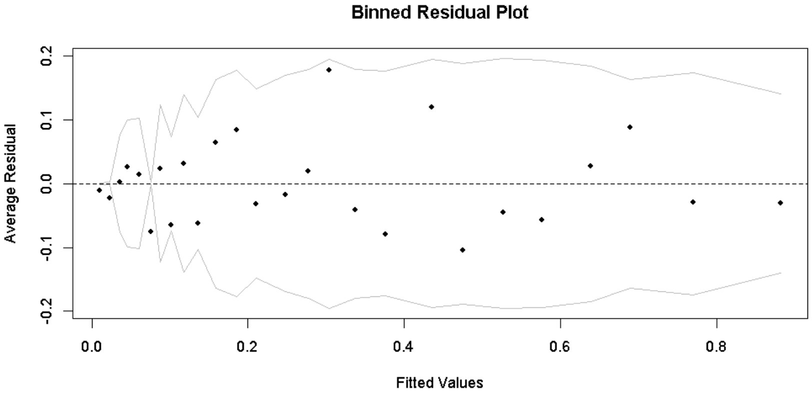

Figure 34: Binned residual plot

10.3.4 Linearity check

The goal of these graphs is to verify that the blue line (representing the estimated relationship) is as straight as possible.

Duration vs. Logit: This variable shows the strongest linear relationship. The blue line rises steadily with a defined slope, indicating that as the loan duration increases, the logit increases proportionally.

Credit Amount vs. Logit: A constant positive slope is noted. This suggests that the logit (risk) increases linearly with the credit amount. The transformation has removed the curvature seen in the previous graphs, making the relationship predictable and stable.

Age vs. Logit: The blue line is almost completely flat and horizontal. This indicates that, after the transformation (likely polynomial), the effect of age has been correctly normalized within the model. The gray band (the confidence interval) widens towards age 75 because there is less data for the very elderly.

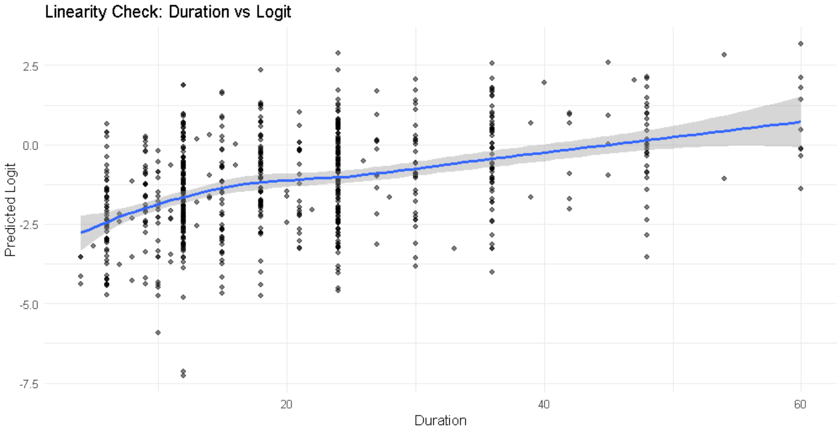

Figure 35: Linearity check - Duration vs Logit

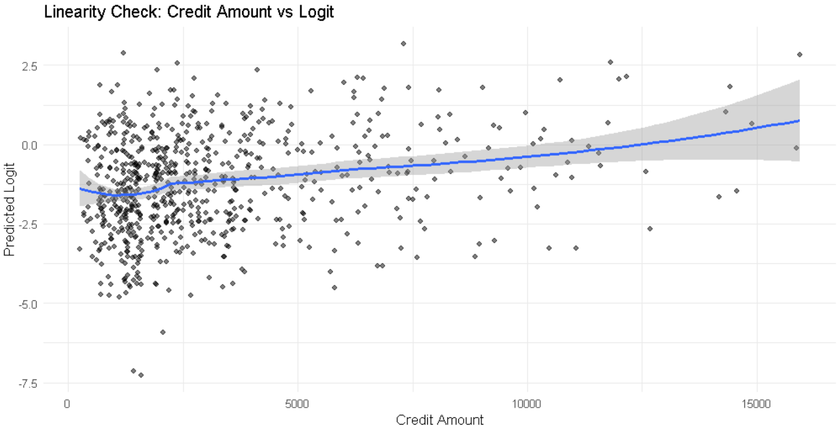

Figure 36: Linearity check - Credit_Amount vs Logit

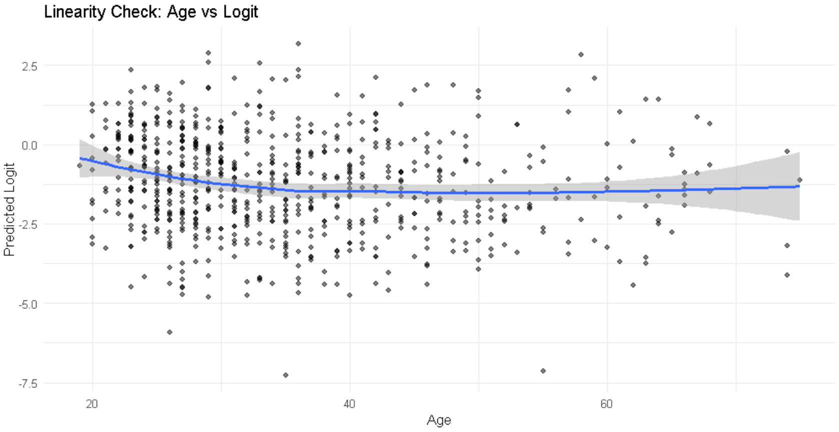

Figure 37: Linearity check - Age vs Logit

10.3.5 Overdispersion

Overdispersion is a measure of the discrepancy between the observed variability in the data and the variability expected under the model. In logistic regression models, it is assumed that the response variable follows a binomial distribution, with variance linked to the predicted probability p according to:

$$\text{Var}(Y) = n \cdot p \cdot (1 - p)$$

If the data exhibit greater variability than expected ($\hat{\phi} > 1$), we have overdispersion. Overdispersion can make the estimated coefficients appear more precise than they actually are, leading to underestimated standard errors. Conversely, if the data exhibit less variability than expected ($\hat{\phi} < 1$), we have underdispersion, which may result in overestimated standard errors.

Table 19: Overdispersion statistic for the logistic regression model

Statistic	Value
Overdispersion ($\hat{\phi}$)	0.9945

Note: The overdispersion statistic is very close to 1, indicating that the logistic regression model does not suffer from over- or under-dispersion and the binomial variance assumption is reasonable.

11 Model Fit and Performance

11.1 Goodness of Fit

We used a null model to see the differences between the logistic regression model and the null model. The likelihood ratio test shows a significant improvement over the null model and also the Pseudo R^2 shows that the logistic regression model is better than the null one. We executed the anova test that helps to calculate the residual deviance of each model, the difference in deviance between the two models and the p-value of the chi-square test, which suggests whether adding the predictors significantly improves the fit compared to the null model. How to interpret the p-value: if it is $< 0{,}05$, predictors definitely improve the model; if it is $> 0{,}05$, predictors are not so useful. Based on the obtained results, we can see that Residual Variance went from 855,21 (null model) to 635,89 (our model) (see Table 19), suggesting that our model is better, as the residual variance the lower it is, the better it is. While the Pseudo R^2 (see Table 20):

- llh = log-likelihood (the higher it is, the better it is). It went from -427,60 to -317,94, so our model is better;
- G2 = Likelihood Ratio Statistic (it measures the same thing as ANOVA above);
- McFadden = Pseudo R-squared (if >0.40 it is perfect). We have 0.26, so the model with good goodness of fit;
- r2ML = Maximum Likelihood R-squared. We have 0.27, so the model with good goodness of fit;
- 2CU = Cragg & Uhler / Nagelkerke R-squared (this can be = 1). We have 0.38, so the model with good goodness of fit.

Table 20: ANOVA - Analysis of deviance for the logistic regression model

Model	Residual Df	Residual Deviance	Df	Deviance	Pr(>Chi)
Null model (Intercept only)	699	855.21	-	-	-
Full model	666	635.89	33	219.32	$< 2.2 \times 10^{-16}$ ***

Note: The comparison of the null and full model shows a highly significant reduction in deviance, indicating that the set of predictors included in the full model significantly improves model fit over the intercept-only model.

Table 21: Pseudo-R^2 measures for the logistic regression model

Statistic	Value
Log-likelihood (full model)	-317.945
Log-likelihood (null model)	-427.605
Deviance (G^2)	219.32
McFadden R^2	0.256
Cox & Snell R^2	0.269
Nagelkerke R^2	0.381

Note: Pseudo-R^2 measures indicate the proportion of improvement in model fit relative to the null model. McFadden R^2 values between 0.2 and 0.4 are generally considered indicative of good model fit. The Nagelkerke R^2 adjusts Cox & Snell R^2 to scale between 0 and 1.

11.2 Classification Performance

After finding an optimal threshold determined by Youden's Index (which in this case is equal to 0,283), we determined the confusion matrix, assessing what the predicted value is and what the actual value is. Thanks to this matrix, we can see the number of: True positives (61), False positives (66), False negatives (149), True negatives (24).

Table 22: Confusion matrix of the logistic regression model

Predicted	Actual	1 (Good)	2 (Bad)
1 (Good)		61	66
2 (Bad)		149	24

Thanks to this confusion matrix, we can calculate: Accuracy, Sensitivity, Specificity, Precision and F1 score. From these results, we see that this model has a strong tendency to predict a good payer (1) when it shouldn't, creating many false negatives and positives. This may be due to the unbalanced classes of the dependent variable (1 = 70% vs 2 = 30%).

Table 23: Classification metrics for the logistic regression model (class 1)

Metric	Value
Accuracy	0.28
Precision	0.14
Sensitivity	0.27
Specificity	0.29
F1 Score	0.18

11.3 ROC Curve and AUC

The ROC curve is a graph that visualizes a binary classification model's performance across all possible thresholds, plotting the True Positive Rate (Sensitivity) (y-axis) against the False Positive Rate (1 - Specificity) (x-axis). The diagonal line (45°) represents a random model guessing whether the payer is bad or good (AUC = 0.5). The more the curve is "pushed towards the upper left corner", the better the model.

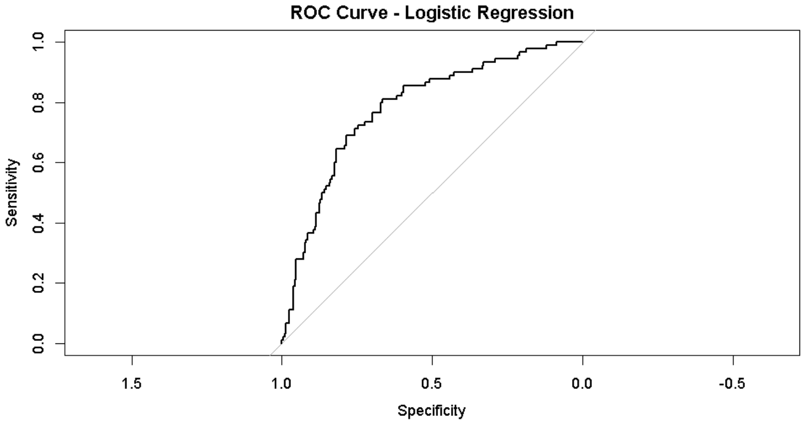

Figure 38: ROC curve - Logistic Regression

Then we go on with the calculation of the AUC, which represents the area under the curve. In our case AUC is equal to 0,7812, meaning that, casually taking a good payer (1) and a bad payer (2), the model assigns a higher probability to the good payer in the 78.12% of the total cases. AUC represents the probability that the model ranks a random positive example higher than a random negative one. We interpret the AUC values in this way: AUC = 1 perfect model (positives rate is = 1, meaning that the model always guesses whether the output is true); AUC = 0,5 model is performing as a random guess. Our AUC's model is closer to 1, indicating that the model is good at guessing the true good payer (1).

12 Conclusion

In this analysis, we developed a logistic regression model to predict credit risk using the German Credit dataset. The primary objective was to identify the key factors associated with a customer being a good or bad payer and to provide a reliable, interpretable model for credit risk assessment. The exploratory data analysis highlighted that both numerical and categorical variables contain meaningful information. Numerical variables such as Duration, Credit Amount, and Age showed skewed distributions and nonlinear relationships with the outcome, which were addressed through appropriate transformations (polynomial terms for Duration and Age, logarithmic transformation for Credit Amount). Categorical variables such as Status_Checking_Acc, Credit_History, Savings, and Employment exhibited strong associations with credit risk, as confirmed by frequency analysis and barplots. The logistic regression model was built using a stepwise selection procedure, optimizing the AIC. The final model indicates that liquidity indicators (checking account status, savings) and past credit behavior (credit history) are the primary determinants of creditworthiness. Nonlinear effects of Age and Duration suggest that risk isn't constant across life stages or loan length, emphasizing the importance of modeling complex relationships. Model diagnostics confirmed the robustness of the results. Multicollinearity was low (all GVIF values < 5), influential points didn't affect estimates (Cook's distance below thresholds), and overdispersion was negligible ($\hat{\phi} = 0,9945$), confirming that the binomial variance assumption is reasonable. The binned residual plot further supported model adequacy, showing no systematic deviations from zero. However, the model also has some limitations. There is a moderate class imbalance in the data (1 = 70% vs 2 = 30%), which may slightly reduce the model's ability to accurately detect bad payers. Overall, the analysis demonstrates that logistic regression, is a reliable tool for credit risk prediction. The model provides interpretable insights into the determinants of creditworthiness, highlighting the importance of financial discipline, repayment history, and liquidity over purely demographic or contractual characteristics.

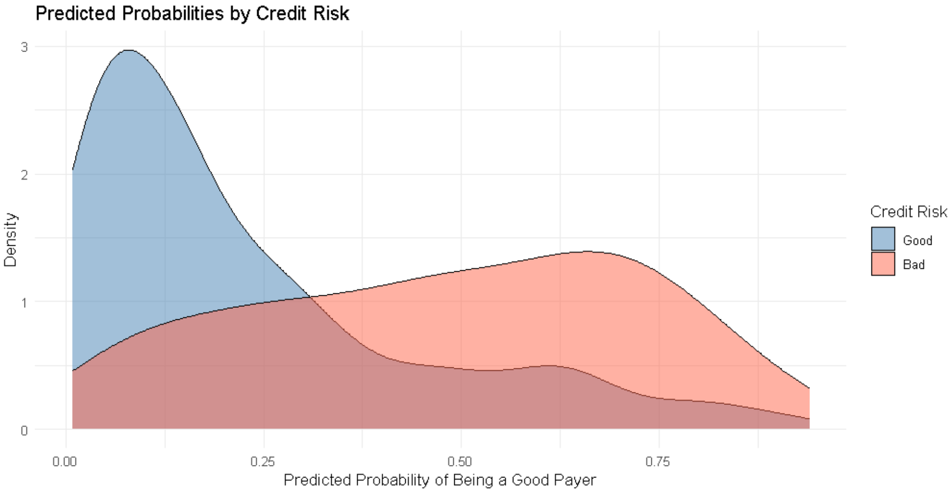

Figure 39: Predicted probabilities by Credit_Risk