Akademia Górniczo-Hutnicza
im. Stanisława Staszica
w Krakowie

Corporate Credit Risk

Mateusz Mglej, Łukasz Obrzut, Dominik Sobczyk,
Faustyna Święch

Wydział Matematyki Stosowanej

Kraków 16.06.2025

Contents

1 Motivation

For financial institutions such as banks, accurately assessing credit risk - the likelihood that a borrower will default on their loan is of critical importance. Late payments or complete defaults can lead to significant financial losses, negatively impact a bank's liquidity, and, in extreme cases, threaten its solvency.

Therefore, banks invest substantial resources in developing and maintaining statistical models and machine learning algorithms that help forecast customer behaviour. These models enable better customer segmentation, more precise risk assessment, and more informed credit decisions.

Effective credit risk models also help institutions comply with regulatory requirements, improve credit portfolio management, and optimize loan pricing and conditions. Additionally, such models can detect early warning signs of a borrower's deteriorating financial health, allowing for preventive or restructuring measures to be taken.

2 Structure of the data

The excel spreadsheet provided ("Data Credit IMR.xls") corresponds to a set of qualitative and quantitative information collected from the Credit Risk team in the period 2000-2008 for Middle-Market Wholesale customers. The fields included in the dataset contain the following information:

1. `CUSTOMER_ID` – internal customer identification number.
2. `ASSESSMENT_YEAR` – year of the credit expert assessment.
3. `PRODUCT_DEMAND` – credit expert's opinion on the competitive environment where the company operates, including its market position and the quality of its portfolio. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
4. `OWNERS_MANAGEMENT` – credit expert's opinion on the quality of the management of the company. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
5. `ACCESS_CREDIT` – credit expert's opinion on the ability of the company to obtain funds from different financial entities. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
6. `PROFITABILITY` – credit expert's opinion on the ability of the company to generate profits based on its current portfolio. The variable can show values from 10 to 90, 90 being the best score a company may obtain.
7. `SHORT_TERM_LIQUIDITY` – credit expert's opinion on the ability of the company to generate cash-flows in the short-term to fulfil short-term financial obligations. The variable can show values from 10 to 90, 90 being the best score a company may obtain.

8. MEDIUM_TERM_LIQUIDITY – credit expert's opinion on the ability of the company to generate cash-flows in the medium and long-term. The variable can show values from 10 to 90, 90 being the best score a company may obtain.

9. GROUP_FLAG – categorical variable which indicates whether the customer belongs to a Financial Holding. It may have two values: "0" – The counterparty does not belong to a Financial Holding, "1" – The counterparty belongs to a Financial Holding (i.e. the counterparty is a subsidiary of a holding company).

10. TURNOVER – it contains the value of the Turnover reported in the financial statements available for the assessment.

11. INDUSTRY – it encodes the industry in which the company operates.

12. DEFAULT_FLAG – it contains whether the customer has gone into default (i.e. has failed in its financial obligations with the Bank) in a 12-month period after the credit expert's assessment.

Data set consists of 5804 rows and 12 columns.

3 The goal of analysis

The primary objective of this analysis is to develop a model capable of predicting the probability of default using the information provided in the available dataset (spreadsheet). By leveraging customer and loan-related features, the model aims to estimate the likelihood that a borrower will fail to meet their repayment obligations.

The purpose of this project is to introduce and apply fundamental tools used in credit risk modelling. The central task involves the development and evaluation of a probability of default model, utilizing techniques such as logistic regression, probit regression, or linear regression.

4 Data exploration and preparation

At first we can assume that variable CUSTOMER_ID do not have an influence on the probability of default. Also we can see that we have missing values for three of our customers, so we remove these whole rows. In the rest of our data we do not have missing values. We divide each numerical variable into groups and convert to a factor type. To improve the clarity and readability of plots, long industry names have been replaced with standardized abbreviations (see Table 1).

4.1 Class imbalance in the outcome variable

Before building and evaluating our models, it is crucial to consider the distribution of the outcome variable. In our case, the binary target variable DEFAULT_FLAG is highly imbalanced. Most observations belong to the class "non-default" (label 0), while the "default" cases (label 1) constitute a much smaller proportion of the data.

This imbalance can lead to biased model performance, where metrics such as accuracy might be misleading. For example, a model that predicts all observations

Code	Industry Description
A,L&F	Agriculture, Livestock and Fisheries
EI	Extractive Industries
M	Manufacturing
O	Other
T	Trade
EG&W	Electricity, Gas and Water
H&L	Hotels and Leisure
OM&CI	Office Machinery and Computer Industries
P&CS	Property and Construction Sectors
TS&CI	Transport, Storage and Communications Infrastructure

Table 1: Industry codes and their full descriptions

as "non-default" may still yield a high accuracy but would completely fail to identify actual defaults. Figure **??** illustrates the imbalance in the distribution of the `DEFAULT_FLAG` variable (0: 5243 - non-default, 1: 558 - default).

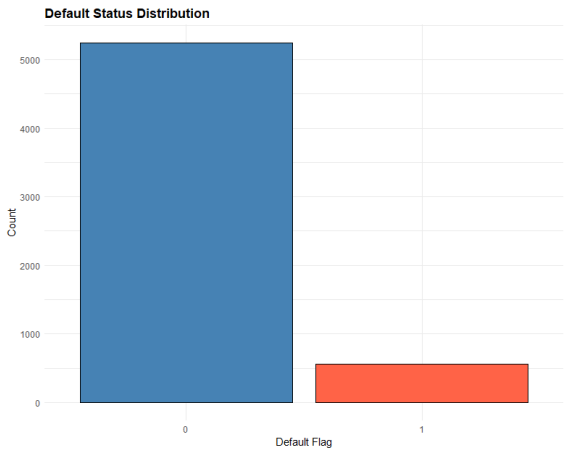


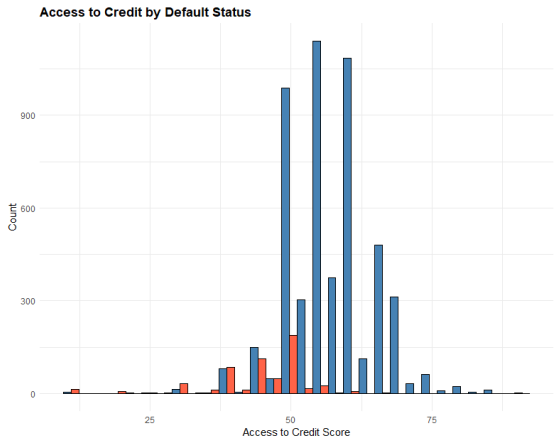
Figure 1: Default Status Distribution

5 Preliminary data analysis

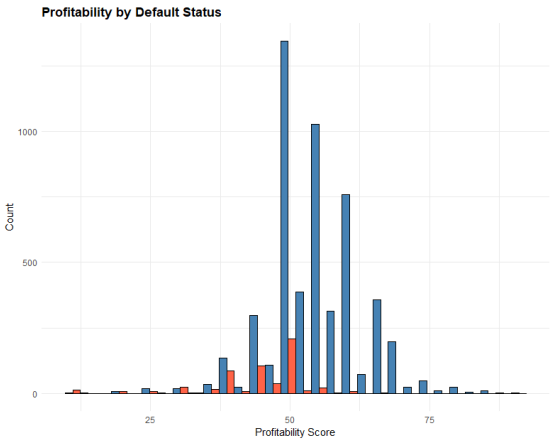
In this section, we present the distribution of the explanatory variables across both default and non-default groups. This analysis helps us understand how frequently each category occurs within the outcome classes.

5.1 Continuous variables

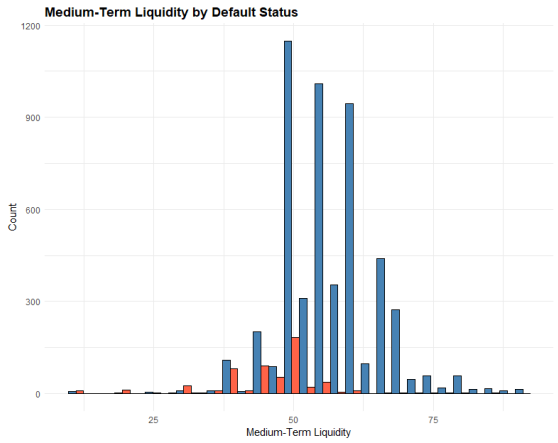
We begin by presenting the distribution of **continuous variables**. These variables offer valuable insight into the relationship between their magnitude and the default status. The figures illustrating their distributions across default and non-default observations are shown below.



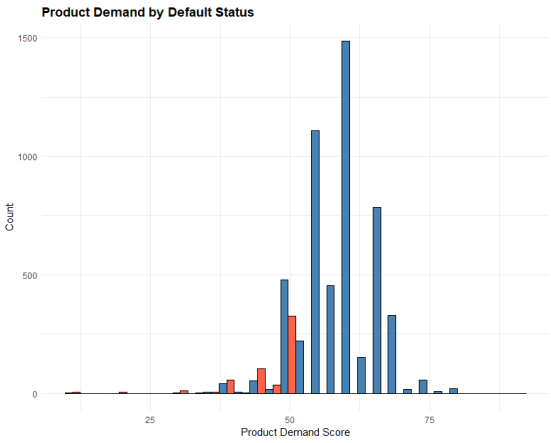
Access to Credit



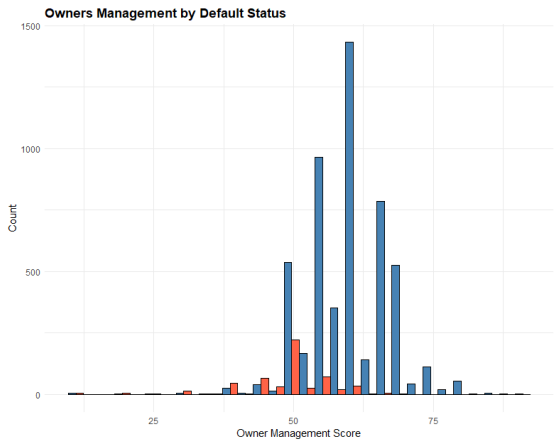
Profitability



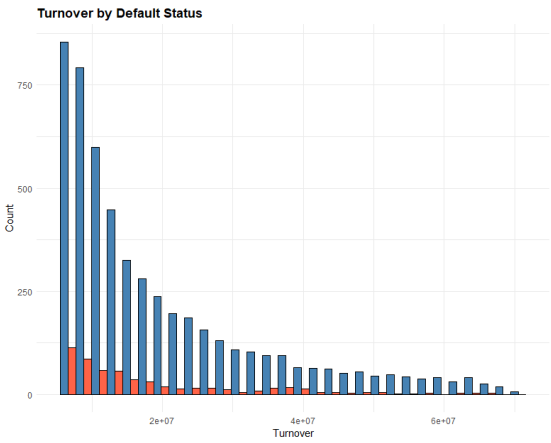
Medium-Term Liquidity



Product Demand



Owners Management



Turnover

Figure 2: Distribution of explanatory variables across default statuses

Based on the presented figures, we can draw preliminary visual insights regarding the distribution of selected explanatory variables across default and non-default

cases. While these observations do not constitute formal statistical conclusions, they provide an initial understanding of how certain variables may relate to loan repayment behaviour.

- In the case of `MEDIUM_TERM_LIQUIDITY`, higher values are more frequently found among non-defaulted loans. A similar pattern can be observed in the distributions of all the aforementioned variables, with the exception of `TURNOVER`. This observation may prove useful in the context of assessing the probability of default.

- For example the `ACCESS_CREDIT` variable, we observe that whenever its value exceeds 60, all corresponding observations belong to non-defaulted loans. This suggests a strong association between high scores and repayment.

Besides these, we will also examine the correlations between these variables, as shown in the figure below.

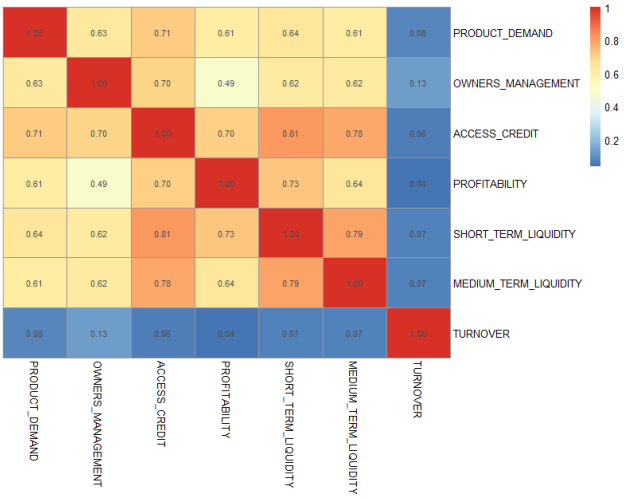
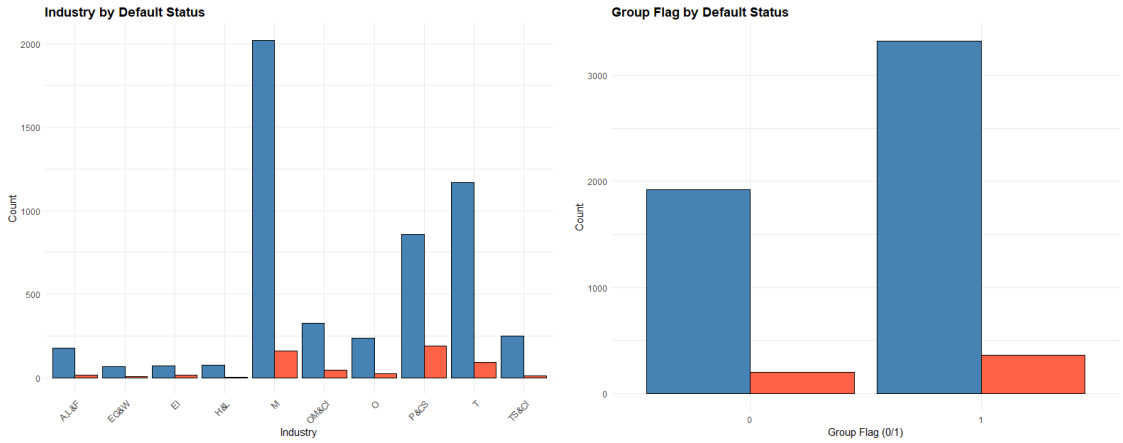


Figure 3: Pearson correlation matrix of numerical variables

5.2 Categorical variables

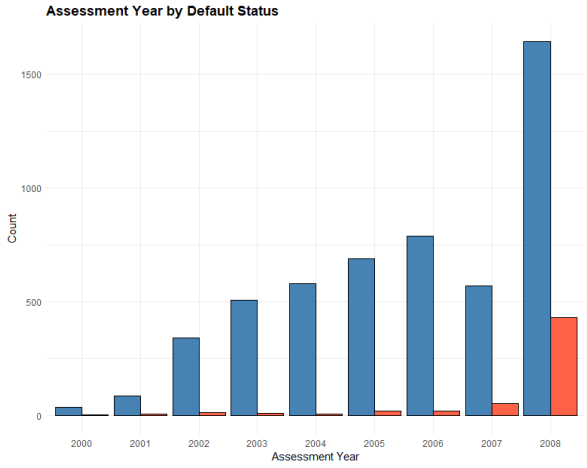
In this section, we focus on variables of a **categorical** nature. As a reminder, these include ASSESSMENT_YEAR, which represents the year in which the credit assessment of a company was performed; GROUP_FLAG, an indicator of whether the company belongs to a capital group; and INDUSTRY, denoting the sector in which the company operates.

The distributions of these variables with respect to credit repayment status are presented in **Figure 4**.



Industry

Group Flag



Assessment Year

Figure 4: Distribution of explanatory categorical variables across default statuses

Below, we also present conclusions drawn from the distributions of the categorical variables.

- The `ASSESSMENT_YEAR` variable shows that most loans were granted in 2008. Additionally, we observe an overall increasing trend in the number of issued loans over time, except for the year 2007, which deviates from this pattern.
- Although the majority of loans in the `INDUSTRY` category belong to "M" (Manufacturing), most of these loans are repaid. However, the highest proportion of defaulted loans within this category is found in the "Property and Construction Sectors."

6 Preliminary variable transformation and selection

6.1 Information value

Information Value (IV) is a widely used metric in the development of credit scoring models and plays a crucial role in variable selection. It measures how well a variable can separate the binary outcome classes, such as default vs. non-default. The formula for calculating Information Value is given below:

$$IV = \sum_{i=1}^{n} (\text{DistributionPositive}_i - \text{DistributionNegative}_i) \cdot \ln\left(\frac{\text{DistributionPositive}_i}{\text{DistributionNegative}_i}\right).$$

The logarithmic component in the IV formula is known as the *Weight of Evidence* (WOE), defined as:

$$WOE_i = \ln\left(\frac{\text{DistributionPositive}_i}{\text{DistributionNegative}_i}\right)$$

Based on the value of IV, we can interpret the predictive strength of each variable as follows:

Information Value	**Predictive Power**
< 0.02	Useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
> 0.3	Strong predictor

Additionally, we examined the **identical rate** of each variable, which represents the proportion of observations sharing the most frequent value. A high identical rate may indicate low variability and limited predictive power.

Results

The table below presents the Information Value for the selected variables:

Variable	**Identical Rate**	**Information Value**
PRODUCT_DEMAND	0.2105	5.0283
ACCESS_CREDIT	0.1791	2.7138
OWNERS_MANAGEMENT	0.2155	2.4537
SHORT_TERM_LIQUIDITY	0.1778	2.2917
MEDIUM_TERM_LIQUIDITY	0.1968	2.0888
PROFITABILITY	0.2333	1.9309
ASSESSMENT_YEAR	0.3568	1.1853
INDUSTRY	0.3794	0.2178
TURNOVER	0.0013	0.0641
GROUP_FLAG	0.6290	0.0001

Table 2: Identical rate and information value for selected variables (sorted by IV)

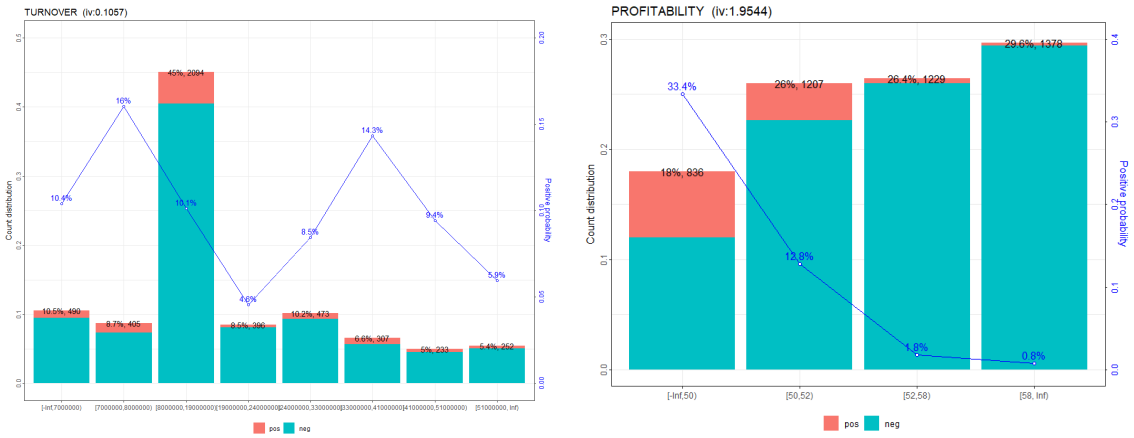
At this stage of the analysis, we decided to **exclude** the variable `GROUP_FLAG` based on its very low Information Value. As previously discussed, an IV below 0.02 indicates that the variable provides little to no predictive power.

6.2 WOE binning

WOE binning is a variable transformation technique that splits a predictor into bins and assigns each bin a WOE value based on the distribution of "good" and "bad" outcomes within it. The *woebin()* function from the scorecard package automatically generates bins that best separate the classes, maximizing Information Value and ensuring monotonicity of WOE where possible.

Results

Below, we present an example of WOE binning for the variables `TURNOVER` and `PROFITABILITY`.



WOE Binning - Turnover WOE Binning - Profitability

Figure 5: Example WOE binning procedure

As illustrated in the figure, we can see why the Information Value (IV) is high for `PROFITABILITY` and low for `TURNOVER`. The IV for `TURNOVER` is low because there is no clear pattern or trend in the distribution of default rates across the bins—the proportion of defaults fluctuates rather than consistently increasing or decreasing, which suggests that the current binning does not effectively separate risk levels. In contrast, for `PROFITABILITY`, there is a clear decreasing trend: the default rate systematically decreases as `PROFITABILITY` increases, indicating strong discriminatory power and a well-aligned binning with credit risk.

Below is a table presenting all the variable binnings generated using the WOE binning.

Variable	Binning intervals
ASSESSMENT_YEAR	2000–2006, 2007, 2008
PRODUCT_DEMAND	[-Inf,50), [50,51), [51,Inf)
OWNERS_MANAGEMENT	[-Inf,50), [50,52), [52,62), [62,Inf)
ACCESS_CREDIT	[-Inf,50), [50,52), [52,58), [58,Inf)
PROFITABILITY	[-Inf,50), [50,52), [52,58), [58,Inf)
SHORT_TERM_LIQUIDITY	[-Inf,50), [50,52), [52,60), [60,Inf)
MEDIUM_TERM_LIQUIDITY	[-Inf,42), [42,50), [50,51), [51,56), [56,Inf)
TURNOVER	<7M, 7–8M, 8–19M, 19–24M, 24–33M, 33–41M, 41–51M, >51M
INDUSTRY	(A,L&F, EG&W, EI), (H&L, M), (OM&CI, O), (P&CS), (T, TS&CI)

Table 3: Variable binning results after `woebin()`

7 Model building and performance information

7.1 Agenda

Before building the models, we split the data by randomly selecting 80% of them for the training set; the rest will form the test set. For each type of model, we will try several variants to select the best-performing one. The models to be compared include:

- Logistic Regression,
- Probit Regression,
- Linear Regression.
- Expert Model,
- Segmentation by `GROUP_FLAG`

After selecting the best variant within each model type, we will compare these models against each other.

7.2 Methods

In our analysis, we will use the **forward** and **backward** selection methods for selecting variables in logistic regression modelling.

- **Forward selection** starts with an empty model and gradually adds variables. At each step, the variable that most improves the model's fit is added.
- **Backward selection** begins with a full model containing all candidate variables. Variables that contribute the least to the model fit are removed step by step.

To select an appropriate logistic regression model in R, we will use the `step` function, which implements both forward and backward selection while minimizing the AIC criterion. However, we should note that the resulting model is not always guaranteed to be the best possible one.

The **Akaike Information Criterion (AIC)** is a commonly used measure for model selection. It balances the trade-off between model fit and model complexity. The AIC is defined as:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

where:

- k is the number of estimated parameters in the model,
- $\hat{L}$ is the maximum value of the likelihood function for the model.

A lower AIC value indicates a better model in terms of the trade-off between goodness of fit and simplicity.

7.3 Metrics

In our analysis, we incorporate several evaluation metrics, presented below, to assess model performance. This choice is particularly motivated by the imbalance in our outcome variable, where the number of repaid loans significantly exceeds the number of defaulted ones.

- **True Positive Rate (TPR)** - also known as Recall or Sensitivity:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **False Positive Rate (FPR)**:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

- **F1 Score** - the harmonic mean of Precision and Recall:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}$$

- **AUC (Area Under the ROC Curve)** - the area under the Receiver Operating Characteristic (ROC) curve, which plots TPR against FPR.

- **Kolmogorov–Smirnov statistic** measures the maximum separation between the distributions of defaulted and non-defaulted clients

$$KS = \max_x |F_{\text{default}}(x) - F_{\text{non-default}}(x)|$$

  - $F_{\text{default}}(x)$ – cumulative distribution function (CDF) of scores for defaulted clients (label = 1),
  - $F_{\text{non-default}}(x)$ – CDF of scores for non-defaulted clients (label = 0).

  Higher values indicate better model discrimination.

7.4 Optimal cut-off

We used the **Youden method** to determine the optimal classification threshold. The Youden method selects the threshold that maximizes Youden's J statistic, defined as

$$J = \text{Sensitivity} + \text{Specificity} - 1,$$

which identifies the cutoff point that optimally balances sensitivity and specificity in binary classification tasks.

7.5 Best model choice

- All models use the same evaluation metrics and the same method for determining the optimal cutoff threshold to ensure consistency and comparability of results.
- For each model category (logistic regression, probit model, linear regression, expert model), several candidate models are created and evaluated using 10-fold cross-validation.
- The best-performing model from each category is selected based on cross-validation results.
- The selected models are then compared against each other on a separate test dataset to determine the final model choice.

7.6 Modeling starting point

In all modeling cases, we begin with a **full model** specification, which includes all explanatory variables that passed initial filtering based on Information Value (IV). The variable `GROUP_FLAG` is excluded from this model, as it showed the lowest IV in the dataset and therefore provides negligible discriminatory power.

The full model serves as the reference point for model simplification procedures. Specifically, **model_2** excludes the variable `TURNOVER`, which had the second lowest IV, allowing us to evaluate the impact of removing low-informative variables.

In addition, we implement both *forward* and *backward* stepwise selection procedures. The forward approach starts from the null model (intercept only) and incrementally adds variables from the full model, while the backward strategy begins with the full model and sequentially eliminates non-contributive predictors.

This structure ensures a transparent and consistent comparison across models and selection strategies, while aligning variable choice with predictive relevance as measured by IV.

8 Binary response models: logit and probit

8.1 Logistic regression

First, we focus on logistic regression, a widely used statistical method for modelling binary outcome variables. In our case, the binary outcome variable Y corresponds

to `DEFAULT_FLAG`, which indicates whether a default has occurred. The predictor variables $X_1, X_2, \ldots, X_n$ represent the explanatory variables used in the model.

The logistic regression model can be written as:

$$p := \mathbb{P}(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

The model uses the **logit link function**, which is defined as:

$$g(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

Models

Below are shown the structures of all models (see Table 4), including Model 3, which was obtained through forward and backward selection methods. Both methods identified the same set of variables to be included in the final model.

	Without WOE Binning	**WOE Binning**
Model	**Description**	**Description**
Model 1	Full model	Full model
Model 2	Full model without – `TURNOVER`	Full model without – `TURNOVER`
Model 3 (forward)	Full model without – `MEDIUM TERM LIQUIDITY` – `INDUSTRY`	Full model without – `MEDIUM TERM LIQUIDITY`

Table 4: Descriptions of logistic models used in the analysis

Performance

Now, we want to evaluate our models using **10-fold cross-validation**. The results of this evaluation are presented below.

	Without WOE Binning			**WOE Binning**		
Metric	**Model 1**	**Model 2**	**Model 3**	**Model 1**	**Model 2**	**Model 3**
AUC	0.9644	0.9647	0.9656	0.9706	0.9699	0.9707
KS	0.8622	0.8628	0.8627	0.8779	0.8762	0.8784
F1	0.9437	0.9440	0.9412	0.9455	0.9519	0.9468

Table 5: Comparison of model performance metrics with and without WOE binning

Despite the comparability of the models, we decided to select **Model 2 without WOE** and **Model 2 with WOE binning** categories, as our decision was mainly based on the **F1 score** metric which is optimal for unbalanced data sets.

8.2 Probit specification

In the next step we consider **probit link function**, the second most popular one model in regression problems with a dichotomous dependent variable. It is defined as:

$$g(p) = \mathbf{\Phi}^{-1}(p).$$

This approach as we can see uses the normal distribution function.

Models

Below are shown the structures of all models (see Table 6), including Model 3, which was obtained through forward selection method and model 4, which was obtained through backward selection method.

	Without WOE Binning	WOE Binning
Model	**Description**	**Description**
Model 1	Full model	Full model
Model 2	Full model without – `TURNOVER`	Full model without – `TURNOVER`
Model 3 (forward)	Full model without – `PROFITABILITY` – `MEDIUM TERM LIQUIDITY` – `INDUSTRY` – `TURNOVER`	Full model without – `MEDIUM TERM LIQUIDITY`
Model 4 (backward)	Full model without – `MEDIUM TERM LIQUIDITY` – `TURNOVER`	

Table 6: Descriptions of probit models used in the analysis

Performance

The models are now evaluated through **10-fold cross-validation**, and the outcomes of this procedure are reported below.

	Without WOE Binning				WOE Binning		
Metric	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**
AUC	0.9394	0.9370	0.9395	0.9205	0.9707	0.9699	0.9707
KS	0.8214	0.8131	0.8213	0.7813	0.8784	0.8768	0.8796
F1	0.9474	0.9491	0.9477	0.9486	0.9458	0.9502	0.9464

Table 7: Comparison of model performance metrics with and without WOE binning

Despite the comparability of the models, we decided to select **Model 2 without WOE** and **Model 2 with WOE binning** categories, as our decision was mainly based on the **F1 score** metric.

8.3 Logit and probit comparison

We now evaluate the best-performing models from the previous 10-fold cross-validation procedure on the independent test set. This comparison focuses on models using two different link functions: *logit* and *probit*.

Logit models on the test set

The following confusion matrices present the performance of the selected logistic regression models on the test dataset. Both models use the logit link function: the first one is built on the original data, while the second is based on WOE-transformed variables.

Model	Prediction	Actual 0	Actual 1
Logit (Original Variables)	0	965 (TN)	5 (FN)
	1	95 (FP)	86 (TP)
Logit (WOE)	0	982 (TN)	5 (FN)
	1	78 (FP)	86 (TP)

Table 8: Confusion matrices for logit models (Test Set)

Probit models on the test set

Similarly, we now evaluate the performance of the probit regression models on the test dataset. The first model is based on the original variables, while the second uses WOE-transformed predictors.

Model	Prediction	Actual 0	Actual 1
Probit (Original Variables)	0	973 (TN)	7 (FN)
	1	87 (FP)	84 (TP)
Probit (WOE)	0	992 (TN)	6 (FN)
	1	68 (FP)	85 (TP)

Table 9: Confusion matrices for probit models (test set)

Performance metrics summary

The table below compares the classification performance of the four evaluated models: logit and probit, each estimated on both raw and WOE-transformed variables. The comparison focuses on AUC, KS, and especially F1-score, which is the main evaluation criterion in this analysis.

Model	AUC	KS	F1 Score
Logit (Original Variables)	0.9622	0.8554	0.9507
Logit (WOE)	0.9648	0.8715	0.9595
Probit (Original Variables)	0.9585	0.8410	0.9539
Probit (WOE)	0.9644	0.8699	0.9640

Table 10: Performance metrics for logit and probit models (test set)

The performance differences between the logit and probit models are minimal, indicating that the choice of link function has little practical impact in this context. However, models using WOE-transformed variables show consistently higher F1 scores than their raw counterparts, suggesting that variable binning improves classification accuracy — particularly in balancing precision and recall.

9 Linear regression

As requested by the IT department now we are building linear model. To do such thing we have to convert our target variable `DEFAULT_FLAG` to numeric type because this statistical method is used for modelling relationships between a continuous outcome variable and one or more explanatory variables. The linear regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon,$$

where ϵ is the error term, assumed to follow a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$, capturing unobserved variability. The model is estimated by minimizing the sum of squared residuals, i.e., the differences between observed and predicted values:

$$\min_{\beta} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2,$$

where $\hat{Y}_i$ denotes the predicted value for the i-th observation.

9.1 Models

Below are shown the structures of all models (see Table 11), including Model 3, which was obtained through forward and backward selection methods. Both methods identified the same set of variables to be included in the final model.

Model	Description
Model 1	Full model
Model 2	Full model without – `TURNOVER`
Model 3 (forward)	Full model without – `PROFITABILITY` – `MEDIUM TERM LIQUIDITY` – `TURNOVER`

Table 11: Descriptions of linear models used in the analysis

9.2 Performance

Now, we want to evaluate our models using **10-fold cross-validation**. The results of this evaluation are presented below.

Metric	Model 1	Model 2	Model 3
AUC	0.9602	0.9603	0.9603
KS	0.8367	0.8356	0.8387
F1	0.9309	0.9313	0.9378

Table 12: Comparison of model performance metrics

Despite the comparability of the models, we decided to select **Model 3**, as our decision was mainly based on the **F1 score** metric.

9.3 Test set evaluation

Following cross-validation, the final linear regression model is evaluated on the independent test set. Although linear regression is not typically used for classification tasks, we apply a threshold-based classification to assess its predictive performance.

The confusion matrix below summarizes the classification results, where the prediction is based on the optimal threshold obtained using the Youden index.

Prediction	Actual 0	Actual 1
0	965 (TN)	6 (FN)
1	95 (FP)	85 (TP)

Table 13: Confusion matrix for linear regression model (Test Set)

9.4 Comparison with classification models

Compared to previous classification approaches, the linear regression model yields slightly inferior performance. While its results remain competitive, the F1 score and KS statistic are both lower than those of the best-performing models. The

table below summarizes the test set metrics for the linear model versus the top WOE-based probit classifier:

Model	AUC	KS	F1 Score
Linear Regression	0.9615	0.8444	0.9503
Probit (WOE-based)	0.9644	0.8699	0.9640

Table 14: Performance comparison: linear regression vs. WOE-based probit

In addition to performance, a key limitation lies in the nature of the predictions. Unlike logistic or probit regression, which naturally output values interpretable as probabilities in the $[0, 1]$ range, linear regression can yield predictions outside this interval. For example, some predicted values on the test set include:

```
-0.024, 0.269, -0.091, -0.039, -0.080, 0.050
```

This makes the interpretation of outputs less intuitive and complicates threshold selection. Although classification is still possible using a threshold (e.g., based on the Youden index), the lack of probabilistic meaning weakens its practical applicability in credit risk modeling.

Given these factors, linear regression is not recommended as a primary model for default prediction in this context.

10 Experts model

The next model we built is based on expert analysis who proposed a model with the following specification:

$$p = \frac{1}{1 + e^{-0.1 \times Score}},$$

where the $Score$ is the weighted average of the following variables:

Variable	WEIGHT
PRODUCT DEMAND	20%
OWNERS MANAGEMENT	10%
ACCESS CREDIT	10%
PROFITABILITY	15%
SHORT TERM LIQUIDITY	25%
MEDIUM TERM LIQUIDITY	20%

Table 15: Criteria weights for expert model

10.1 Model definition

To apply the expert-defined model within the `glm` framework, a custom link function was introduced. The model is constructed using the following logistic regression formula:

$$\texttt{glm(DEFAULT\_FLAG} \sim \texttt{Score, family = binomial(link = custom\_logit\_01()))}$$

Here, `Score` represents the weighted average of selected continuous variables, as defined in Table 15. The use of a custom link allows us to match the expert-specified shape of the probability function within a standard regression framework, without altering the data or structure of the model.

The estimated model coefficients are both statistically significant ($p < 0.001$), confirming that the expert-defined `Score` is a strong predictor of default risk, with higher scores associated with lower default probability.

10.2 Evaluation on test set

The performance of the expert model was evaluated on the test set. The confusion matrix is presented below:

Prediction	Actual 0	Actual 1
0	906 (TN)	10 (FN)
1	154 (FP)	81 (TP)

Table 16: Confusion matrix for expert model (test set)

Key performance metrics calculated on the test set are summarized below:

Model	AUC	KS	F1 Score
Expert Model	0.9385	0.7448	0.9170

Table 17: Performance metrics for expert model (test set)

10.3 Model comparison and justification

Although the expert model is intuitive and straightforward to implement, it is clearly outperformed by statistically derived models. On the test set, its F1 score (0.9170) and KS statistic (0.7448) are lower than those of the best WOE-based models, such as the probit model (F1 = 0.9640, KS = 0.8699).

This performance gap highlights the benefits of data-driven modeling, which adapts to real patterns in the dataset rather than relying on fixed weights. From a business perspective, more accurate classification reduces credit risk by improving both acceptance and rejection decisions.

For these reasons, we recommend implementing the statistically optimized model, which offers both stronger predictive power and better flexibility over time.

11 Segmentation analysis by financial holding type

Our final task involved building two distinct models to differentiate between independent entities and subsidiaries within financial holdings. The dataset was partitioned based on the `GROUP_FLAG` variable, creating separate training and test sets for each entity type. Both models utilized the same predictor variables as in Model 2 with logistic link function (excluding `TURNOVER`).

11.1 Data segmentation

The dataset was divided according to financial holding status:

- **Independent Entities** (`GROUP_FLAG = 0`): Companies not part of any financial holding
- **Subsidiaries** (`GROUP_FLAG = 1`): Entities belonging to financial holdings

This segmentation allowed us to create specialized models capturing potential differences in default risk factors between these two types of organizations.

Model	Description
Model 4	Full model for Independent Entity
Model 5	Full model for Subsidiary of Financial Holding

Table 18: Descriptions of models with respect of Financial Holding used in the analysis

11.2 Performance evaluation

The performance of those both models was evaluated on the test set. The models demonstrated substantially different classification patterns. The confusion matrix for Model 4 is presented below:

Prediction	Actual 0	Actual 1
0	345 (TN)	1 (FN)
1	22 (FP)	28 (TP)

Table 19: Confusion matrix for Model 4 (Independent Entities)

The confusion matrix for Model 5 is presented below:

Prediction	Actual 0	Actual 1
0	607 (TN)	4 (FN)
1	86 (FP)	58 (TP)

Table 20: Confusion matrix for Model 5 (Subsidiaries)

Comparative metrics

The models showed distinct performance characteristics. Both models were evaluated using standard metrics with optimal thresholds determined by the Youden index.

Model	AUC	KS	F1 Score
Independent Entity	0.9683	0.9056	0.9677
Subsidiary	0.9404	0.8114	0.9310

Table 21: Performance metrics for segmented models (test set)

12 Summary

Table 22 demonstrates that segmentation by entity type (independent vs. subsidiary) provides valuable insights. WOE transformation appears to be beneficial across different model types, while the experts model's performance suggests potential limitations. Linear regression model also has some bounders in binary classification problem and should be avoided for this task.

Model	AUC	KS	Accuracy	F1
Logit (Original Variables)	0.9622	0.8554	0.9131	0.9507
Logit (WOE)	0.9648	0.8715	0.9279	0.9595
Probit (Original Variables)	0.9585	0.8419	0.9192	0.9544
Probit (WOE)	0.9644	0.8699	0.9357	0.9640
Linear	0.9615	0.8444	0.9123	0.9503
Experts model	0.9385	0.7448	0.8575	0.9170
Independent Entity	0.9684	0.9056	0.9419	0.9677
Subsidiary	0.9404	0.8114	0.8808	0.9310

Table 22: Comparison of models performance metrics

After the whole analysis, we can conclude that model for Independent Entity demonstrated outstanding performance which suggests that differentiated risk management strategies should be developed for independent versus subsidiary entities. If segmentation by entity type is not feasible (e.g., due to data constraints or operational limitations), the Probit model with WOE-transformed variables emerges as the most effective general-purpose model. However, the Logit model with WOE also performs nearly as well and could be a valid alternative. Across all cases, WOE transformation improves predictive performance, reinforcing its value in credit risk modelling.