BI2025 Experiment Report - Group 54

Alexander Resch*
TU Wien
Austria

Jakob Kimeswenger†
TU Wien
Austria

Abstract

This report documents the machine learning experiment for Group 54, following the CRISP-DM process model.

CCS Concepts

• **Computing methodologies** → **Machine learning**.

Keywords

CRISP-DM, Provenance, Knowledge Graph, Machine Learning

ACM Reference Format:
Alexander Resch and Jakob Kimeswenger. 2025. BI2025 Experiment Report - Group 54. In *Proceedings of Business Intelligence (BI 2025).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Business Understanding

1.1 Data Source and Scenario

We decided to use the "Corporate Credit Rating with Financial Ratios" dataset available on Kaggle. The dataset contains a set of financial ratios measuring liquidity, leverage and profitability for multiple companies, together with a corporate credit rating label. The scenario is a financial institution that wants to assess corporate credit risk based on financial statement information. The model should assist analysts in assigning credit ratings in decision on creditworthiness.

1.2 Business Objectives

- Support analysts with consistent credit rating decisions for coporate clients.
- Identify companies with high credit risk early, to reduce credit losses.
- Decrease the effort for manual review for low risk companies.
- At new credit applications, speed up the rating decisions.

1.3 Business Success Criteria

- Reduce inconsistencies in ratings between analysts by 15 percent.
- Cut down average review time by 20 percent per company.
- Keep the share of wrongly accepted high-risk firms below 10 percent.
- Avoid blocking of more than 30 percent of low-risk firms.

1.4 Data Mining Goals

- Train a multi-class classifier that predicts corporate credit ratings from financial ratios.
- Provide calibrated probability scores to rank firms by risk.
- Identify the most important financial ratios driving the rating decisions.
- Support later analysis of class-wise performance for different rating levels.

1.5 Data Mining Success Criteria

- Achieve a macro-averaged F1 score of at least 0.7 on the test set.
- Achieve at least a recall of 0.75 for the lowest rating classes.
- Keep the difference between validation and test F1 below 0.05.
- Maintain accuracy for all classes rated above 0.60 where sample size is sufficient.

1.6 AI risk aspects

- Historical ratings can contain human bias, which the model will then learn.
- Financial ratios behave differently across industries, which can disadvantage some sectors.
- Wrong predictions for low-rated companies may result in financial losses.
- Wrong predictions about high-quality companies lead to loss of business and reputational damage.
- Concept drift can occur because economic conditions change over time.

*Student A, Matr.Nr.: 12017130
†Student B, Matr.Nr.: 12122531

BI 2025, -

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

2 Data Understanding

2.1 Load data

Credit ratings and financial ratios for corporations. Load the corporate credit rating dataset from CSV, clean column names, parse the `rating_date` field and created a hierarchical time index (year, month, day) with an additional `day_of_week attribute`.

Table 1: Most important Raw Data Features

Feature Name	Data Type	Description
binary_rating	integer>	Good vs bad rating indicator
current_ratio	double>	Current assets divided by current liabilities
debt_equity_ratio	double>	Leverage ratio total debt relative to equity
rating	string>	Long term credit rating symbol (for example AAA, BBB-)
rating_date	dateTime>	Date at which the credit rating was assigned

2.2 Attribute types, units, semantics

Document attribute groups, units and semantics for the corporate credit dataset. Financial ratios (liquidity, leverage, profitability, efficiency, growth, cash flow) are dimensionless and express relative quantities. Absolute quantities such as total assets, total liabilities, revenues, and cash are measured in USD. `rating` is an ordered categorical label with classes 0 to 8, where higher values indicate higher credit quality. `binary_rating` groups low quality (0,1,2) versus all other classes. `rating_date` records the decision date and defines the time axis for potential concept drift. Derived calendar attributes (year, month, day, day_of_week) support temporal analysis and checks for seasonality.

2.3 Structure of dataset

Table 2 summarizes the basic structure of the corporate credit dataset, including number of rows, columns, data types, missing values, and unique values per column.

2.4 Distribution analysis and skewness

Analyze distributions and skewness of key financial ratios and inspect the frequency distribution of `ratings`, `binary_rating`, `rating_agency` and `sector`. Table 3 and 4 summarizes numeric distributions, skewness, and categorical frequency counts for the corporate credit dataset:

2.5 Correlation analysis

Compute pairwise correlations for numerical attributes. Strong absolute correlations above 0.7 or below -0.7 indicate multicollinearity and redundant information. These findings influence feature selection and model design. Correlation between financial ratios and the binary credit rating is shown in Table 5.

Column name	Data type	Missing	Unique
rating_agency	object	0	7
corporation	object	0	1377
rating	object	0	23
rating_date	datetime64[ns]	0	1414
cik	int64	0	686
binary_rating	int64	0	2
sic_code	float64	0	240
sector	object	0	12
ticker	object	0	678
current_ratio	float64	0	2521
long_term_debt___capital	float64	0	2241
debt_equity_ratio	float64	0	2484
gross_margin	float64	0	2601
operating_margin	float64	0	2648
ebit_margin	float64	0	2648
ebitda_margin	float64	0	2649
pre_tax_profit_margin	float64	0	2649
net_profit_margin	float64	0	2642
asset_turnover	float64	0	2424
roe___return_on_equity	float64	0	2651
return_on_tangible_equity	float64	0	2648
roa___return_on_assets	float64	0	2632
roi___return_on_investment	float64	0	2641
operating_cash_flow_per_share	float64	0	2590
free_cash_flow_per_share	float64	0	2585
day_of_week	object	0	7

Table 2: Dataset structure overview with data types and cardinalities

Variable	Count	Min	Max	Mean	Median	Skew
current_ratio	7805	0.17	34.08	1.93	1.50	7.26
debt_equity_ratio	7805	-1473.10	194.38	0.18	0.75	-27.88
gross_margin	7805	-87.68	100.00	42.43	39.64	0.35
operating_margin	7805	-461.79	93.99	11.53	12.12	-9.75
asset_turnover	7805	0.07	8.50	0.85	0.66	2.77
roa___return_on_assets	7805	-226.44	114.72	4.58	4.77	-5.27
roe___return_on_equity	7805	-11258.21	7038.46	15.95	12.50	-9.28

Table 3: Descriptive statistics of numerical financial ratios

Variable	Unique	Top value	Frequency
rating	23	BBB	910
binary_rating	2	1	5099
rating_agency	7	Egan-Jones Ratings Company	2826
sector	12	Other	1251

Table 4: Overview of categorical variables and most frequent values

2.6 Outlier detection

Identify potential outliers in key financial ratios using a z-score based approach (Table 6). Ratios with an absolute z-score larger than 3.0 are flagged as outliers. After inspecting the report, the decision is to cap extreme leverage and profitability ratios in the preparation phase instead of dropping rows, to keep as much data as possible.

Variable	Correlation with binary_rating
current_ratio	-0.162
gross_margin	0.151
operating_margin	0.201
ebit_margin	0.202
ebitda_margin	0.158
pre_tax_profit_margin	0.224
net_profit_margin	0.207
roa___return_on_assets	0.253
roe___return_on_equity	0.033
asset_turnover	0.004

Table 5: Correlation between financial ratios and binary credit rating

Feature	# Outliers	Share (% of 7805)
current_ratio	100	1.28
debt_equity_ratio	31	0.40
gross_margin	5	0.06
operating_margin	81	1.04
roa___return_on_assets	92	1.18
roe___return_on_equity	37	0.47

Table 6: Outlier counts per numerical feature using z-score thresholding

2.7 Plausibility check

Check the plausibility of key financial ratios by counting values with clearly unrealistic ranges, for example extreme negative or positive liquidity and profitability ratios. Summary of plausibility checks for financial ratios in the corporate credit dataset:

Plausibility check rule	Violations (count)
current_ratio < 0	0
gross_margin outside $[-100, 100]$	0
operating_margin outside $[-100, 100]$	56
roa < -100	4
roa > 100	4
roe < -1000	11
roe > 1000	20

Table 7: Summary of plausibility checks for selected financial ratios

2.8 Visual exploration of key variables

Produce histograms for selected financial ratios and a bar chart for rating classes. The plots Figure 1-4 show skewed distributions, long tails, and strong class imbalance. These observations support later decisions on outlier handling, transformation, and class weighting strategies.

2.9 Sensitive attributes and minority groups

The dataset does not contain explicit sensitive attributes such as gender or ethnicity. Minority groups appear in rare target classes, especially rating classes 0 and 1. These classes have few observations and need attention in evaluation and model design, for example through class weights or sampling strategies.

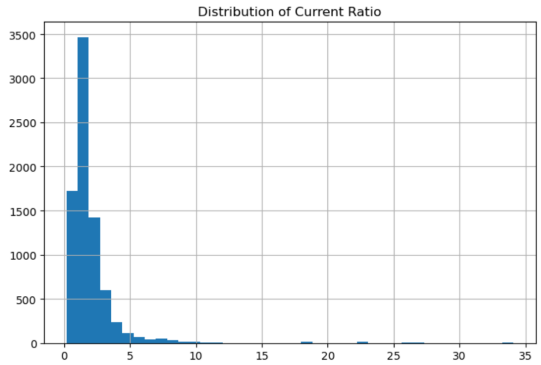

Figure 1: Distribution of Current Ratio

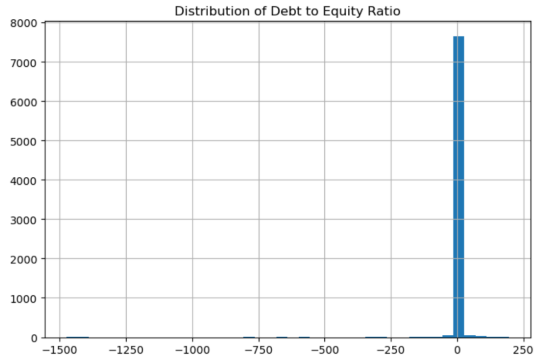

Figure 2: Distribution of Debt to Equity Ratio

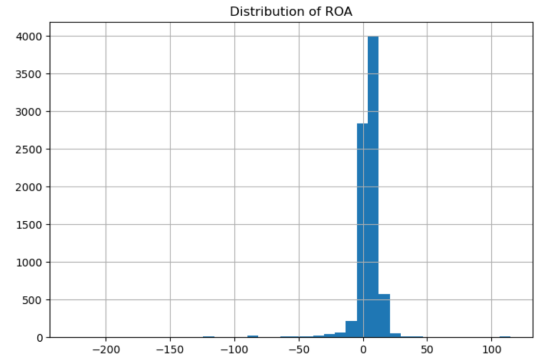

Figure 3: Distribution of ROA

2.10 Risks and bias in the data

Historical ratings reflect past human judgment and internal policies. This creates systematic bias in the target labels. Missing sector information hides structural differences between industries and firm sizes. `rating_date` spans multiple years, so changes in the economic cycle and internal rating guidelines influence label stability and calibration.

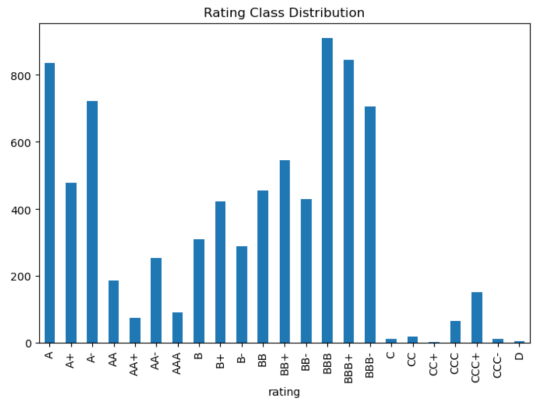

Figure 4: Rating Class Distribution

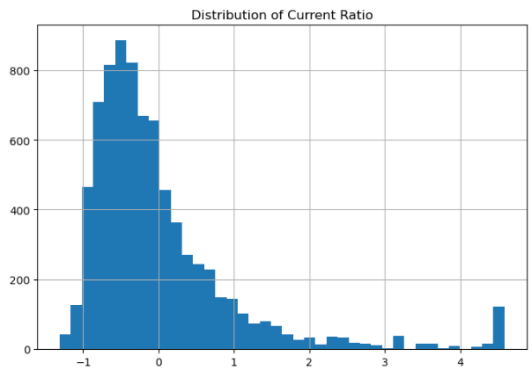

Figure 5: Distribution of Current Ratio after cleaning the data

Open questions for a domain expert:

- Did rating guidelines change during the covered period.
- Are some industries overrepresented in the dataset.
- Are financial ratios comparable across company sizes for all rating classes.

2.11 Planned actions for data preparation

Planned preparation actions based on the data understanding findings:

- Cap extreme outliers in leverage, liquidity, and profitability ratios instead of dropping rows, to keep more observations.
- Apply suitable strategies for missing values, for example dropping rows with many missing entries and imputing single missing ratios.
- Encode rating as ordinal target and keep binary_rating as helper target for alternative evaluation.
- Address class imbalance with stratified splits and class weights during model training.
- Normalize or transform strongly skewed ratios, for example with log transforms for strictly positive variables.

3 Data Preparation

3.1 Handle outliers

During the outlier handling phase, extreme values that were identified in the Data Understanding phase were addressed in order to reduce their influence on the analysis. Instead of removing entire observations, these outliers were treated so that more data points could be retained while still improving the overall quality and stability of the dataset. The Figure 5 shows how the Distribution of Current Ratio after cleaning the data looks like. We can see that the outliers where removed.

3.2 Handle missing values

In the missing value handling step, observations with a high proportion of missing data were removed to ensure data reliability, while isolated missing values in financial ratio variables were imputed using robust column medians. Non-ratio variables were excluded from imputation.

3.3 Encode rating

In this step, the credit rating variable was transformed into an ordinal numeric scale to reflect the natural ordering of credit quality from higher to lower ratings. This makes the rating variable easier to use in the modeling process, while still preserving the relative differences between rating categories. In addition, the binary rating variable was kept unchanged and used as a supporting target for alternative evaluation and comparison of model results.

3.4 Not used

3.4.1 Binning. Binning was not applied to the financial ratio variables. Although binning can simplify continuous variables by grouping them into categories, it also leads to a loss of information. In this dataset, the ratios contain meaningful continuous variation that is important for modeling credit risk. Since outliers were already handled and the models used can work well with continuous variables, keeping the original ratio values was considered more appropriate than applying binning.

3.4.2 Transformations. Additional transformations of numerical variables, such as logarithmic or power transformations, were considered to address skewed distributions. These transformations were not applied because several financial ratios contain zero or negative values, and applying a uniform transformation could distort the original economic meaning of the variables. After handling outliers, the remaining distributions were considered acceptable for further analysis.

3.4.3 Scaling. Feature scaling was considered to bring numerical variables onto a comparable scale. This step was not applied at this stage, as the decision depends on the final choice of modeling approach. Scaling can be applied later if required by the selected model, but was not enforced during the data preparation phase.

3.4.4 Attribute removal. The removal of attributes was considered to reduce dimensionality and remove potentially redundant information. However, all remaining variables were deemed relevant either as financial indicators, identifiers, or target variables. Therefore, no attributes were removed during data preparation.

3.5 Derived attributes

The possibility of creating additional derived attributes was considered, such as combining financial ratios or creating summary indicators for liquidity, leverage, or profitability. While these derived features could potentially capture more complex relationships in the data, they were not created at this stage. The original financial ratios already provide sufficient information, and adding derived attributes would increase complexity without clear benefits for the current analysis.

3.6 External data sources

The use of additional external data sources was considered to improve the prediction of corporate credit risk. Possible external attributes include macroeconomic indicators such as interest rates or economic growth, as well as industry-level information to capture differences between sectors. Market-based data like stock price volatility could also provide useful information. These data sources were not included in this analysis and could be explored in future work.

4 Modeling

4.1 Hyperparameter Configuration

The model was trained using the following hyperparameter settings:

Table 8: Hyperparameter Settings

Parameter	Description	Value
Learning Rate	...	1.23

4.2 Training Run

A training run was executed with the following characteristics:

- **Algorithm:** Random Forest Algorithm
- **Start Time:** 2025-12-15 14:20:28
- **End Time:** 2025-12-15 14:20:28
- **Result:** R-squared Score = 1.2300

5 Evaluation

6 Deployment

7 Conclusion