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, 8 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.

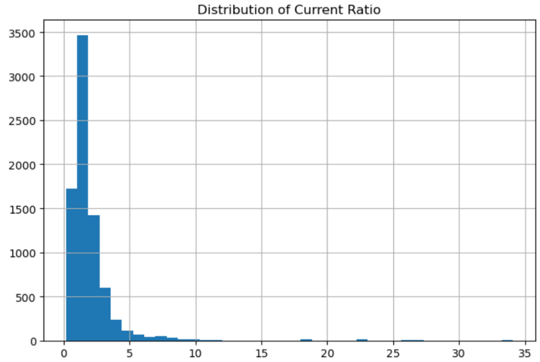


Figure 1: Distribution of Current Ratio

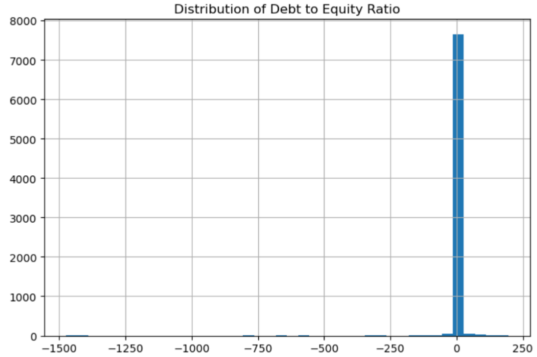


Figure 2: Distribution of Debt to Equity Ratio

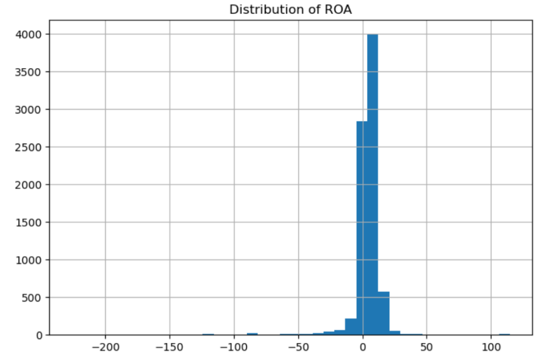


Figure 3: Distribution of ROA

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.

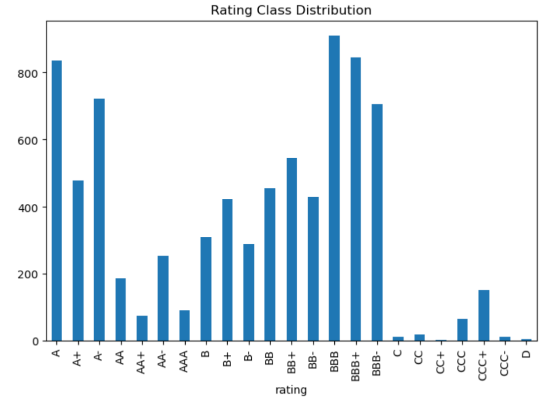


Figure 4: Rating Class Distribution

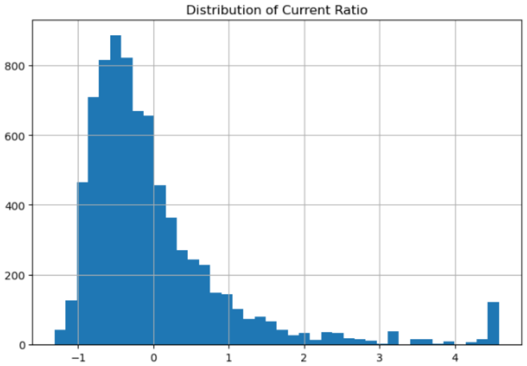


Figure 5: Distribution of Current Ratio after cleaning the data

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

This section describes the modeling phase of the CRISP-DM process. The goal is to train, tune, and select a multi-class classification model for corporate credit ratings based on financial ratios.

4.1 Model Choice

The model chosen was a classifier based on the random forest algorithm. Random Forests are an ensemble learning method based on decision trees and are good for analyzing tabled financial information. They handle non-linear relations, are robust against outliers, and do not need scaling. Moreover, they perform well in class imbalance situations when class weighting is involved.

It is modeled as a multi-class classification problem, with the target variable as an ordinal variable corresponding to different credit levels.

4.2 Hyperparameter Configuration

The following hyperparameters were fixed for all training runs:

- Number of trees ($n\_estimators$): 300
- Class weighting: balanced
- Random seed: 42
- Parallel execution: enabled

To control model complexity and reduce overfitting, the maximum tree depth was selected as the tuning parameter. The following values were evaluated:

- $max\_depth \in \{5, 10, 15, 20, None\}$

All other hyperparameters remained fixed to ensure comparability across runs.

4.3 Training and Validation Strategy

The prepared dataset was split into three disjoint subsets using stratified sampling:

- Training set: 70 percent
- Validation set: 15 percent
- Test set: 15 percent

Only the training and validation sets were used during hyperparameter tuning. Rare rating classes with insufficient samples were removed prior to splitting to ensure stable stratification. After filtering, 18 rating classes remained. The training set contains 5432 samples, and the validation set contains 1164 samples.

4.4 Training Runs

For each value of `max_depth`, one independent training run was executed. Each model was trained on the training set and evaluated on both the training and validation sets.

The following evaluation metrics were computed:

- Macro-averaged F1 score
- Classification accuracy

Macro F1 was chosen as the primary selection criterion because it weights all classes equally and reflects performance on minority rating classes.

The results show a clear improvement in validation Macro F1 with increasing tree depth. Shallow trees underfit the data, while deeper trees capture more complex decision boundaries. The highest validation Macro F1 score was achieved with `max_depth = None`. This configuration was therefore selected as the final model configuration.

A tuning curve visualizing validation Macro F1 as a function of `max_depth` was generated to support this decision. The visualization is shown in Figure 6.

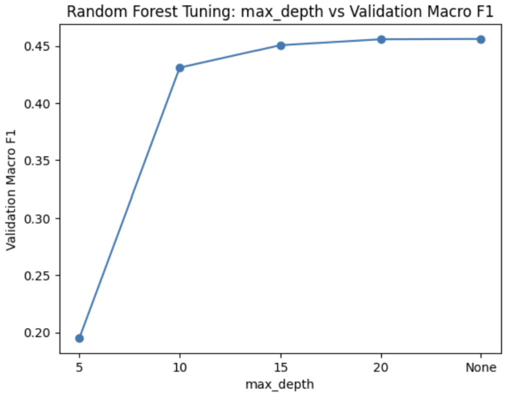


Figure 6: Random Forest Tuning: max_depth vs Validation Macro F1

4.5 Final Model Training

After selecting the optimal hyperparameter configuration, the final model was retrained using the combined training and validation datasets. This maximizes the amount of labeled data available for learning while keeping the test set untouched for final evaluation.

The final training dataset contains 6596 samples across 18 rating classes. The final Random Forest model uses the following configuration:

- *max_depth* = *None*
- *n_estimators* = 300
- class_weight = balanced

This final model is used exclusively for evaluation on the held-out test set in the next phase.

5 Evaluation

To evaluate the performance of our model we have to look at different metrics. The accuracy of our model is only 0.4553 this means only 45% of the classifications are correct. But this can be explained. We have 18 different classes and these classes are ordinal. A wrong classification is not so bad if it is not far away from the correct class. So we have to look at the distances. One way to do this is to look how many classifications have a distance of 1 or less, so the are in the right or in one of the neighboring classes. In this case our accuracy is 0.7191 which means nearly 72 percent of our classification do have a maximum distance of 1 from the right class.

Other measures to look at are the Mean Absolute Error (MAE) and Quadratic Weighted Kappa (QWK).

The Mean Absolute Error is the average of the distances. For our model the mea is 1.15, this means on average the classification is 1.15 classes away from the true class. For a model with 18 classes this is quite ok and a rather small error.

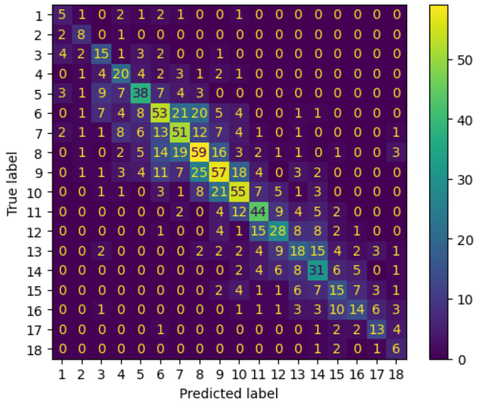
Quadratic Weighted Kappa measures the agreement between predictions and truth and penalizes big mistakes quadratically. This means if the prediction is very wrong it is weighted higher than if it is only slightly wrong. For our model qwk is 0.86 which is very good, 1 would be perfect.

Table 8: Model Performance Metrics

Metric	Value
Accuracy	0.4553
F1 Score (Macro)	0.4471
Precision (Macro)	0.4345
Recall (Macro)	0.4740
Mean Absolute Error	1.1529
Quadratic Weighted Kappa	0.8622
Accuracy (±1 notch)	0.7191

5.1 Comparison

Makwana, R., Bhatt, D., Delwadia, K. et al. Understanding and Attaining an Investment Grade Rating in the Age of Explainable AI. Comput Econ 66, 105–126 (2025) showed with the same dataset that the could reach a accuracy of 0.80. That is higher than our accuracy. The difference from their approach to ours is that they only used 9 features instead of the 18 features we used. This is most likely the reason for the difference in performance.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	5	1	0	2	1	2	1	0	0	1	0	0	0	0	0	0	0	0
2	2	8	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	4	2	15	1	3	2	0	0	1	0	0	0	0	0	0	0	0	0
4	0	1	4	20	4	2	3	1	2	1	0	0	0	0	0	0	0	0
5	3	1	9	7	38	7	4	3	0	0	0	0	0	0	0	0	0	0
6	0	1	7	4	8	53	21	20	5	4	0	0	1	1	0	0	0	0
7	2	1	1	8	6	13	51	12	7	4	1	0	1	0	0	0	0	1
8	0	1	0	2	5	14	19	59	16	3	2	1	1	0	1	0	0	3
9	0	1	1	3	4	11	7	25	57	18	4	0	3	2	0	0	0	0
10	0	0	1	1	0	3	1	8	21	55	7	5	1	3	0	0	0	0
11	0	0	0	0	0	0	2	0	4	12	44	9	4	5	2	0	0	0
12	0	0	0	0	0	1	0	0	4	1	15	28	8	8	2	1	0	0
13	0	0	2	0	0	0	0	2	2	2	4	9	18	15	4	2	3	1
14	0	0	0	0	0	0	0	0	0	2	4	6	8	31	6	5	0	1
15	0	0	0	0	0	0	0	0	2	4	1	1	6	7	15	7	3	1
16	0	0	1	0	0	0	0	0	0	1	1	1	3	3	10	14	6	3
17	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2	2	13	4
18	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	1	6

Figure 7: Confusion Matrix for the evaluation

The performance of a random classifier would be 1/18 because we have 18 classes. This results in 0,055 so 5,5% would be correctly classified. Our model is much better

5.2 Success criteria

The success criteria we defined in the beginning: The only criteria where we are successful is 'Keep the difference between validation and test F1 below 0.05.' F1 for validation set is 0.45 and for test 0.4471 so not even 1 percent difference. The other criteria are not fulfilled for reasons talked about before.

Table 9: Evaluation of Model Requirements

Requirement	Target	Observed	Met
Macro-averaged F1 score	≥ 0.70	0.447	No
Recall (lowest rating classes)	≥ 0.75	0.474	No
Validation–test F1 difference	≤ 0.05	0.01	Yes
Accuracy for classes above threshold	≥ 0.60	0.455	No

5.3 Bias

We have not protected attribute like gender, race, age in our dataset. So we will use the attribute 'sic_code' to test if there is a bias towards. The sic_code is a code which shows what main business are a company has.

The results show that model performance varies between industries. Some SIC groups achieve higher accuracy and lower mean absolute error, while others show larger prediction errors.

The directional bias is close to zero for most SIC groups, indicating that the model does not systematically over- or under-rate companies. One group shows a large directional bias, but this group contains only very few samples and is therefore not reliable. Overall, the observed differences indicate performance similar across industries rather than systematic discriminatory bias.

6 Deployment

6.1 Comparison and recommendations

The developed credit rating model shows mixed performance when compared to the predefined business success and data mining success criteria. While the overall classification accuracy (0.46) and

Table 10: Model Performance by SIC Group

SIC	Samples	Accuracy	MAE	±1 Notch	Directional Bias
7	98	0.5408	0.8571	0.7653	-0.1224
6	35	0.5714	0.9143	0.8000	-0.1143
2	235	0.4511	1.0979	0.7234	-0.0596
3	275	0.4727	1.1273	0.7491	-0.2836
5	102	0.4216	1.1471	0.6863	-0.3039
4	236	0.4492	1.2712	0.7288	0.0508
1	145	0.4207	1.2828	0.6345	0.0276
8	33	0.2727	1.3030	0.6667	-0.2121
9	5	0.4000	2.4000	0.4000	1.6000

macro-averaged F1 score (0.45) fall short of the strict data mining success targets, the ordinal performance metrics indicate that the model still provides substantial decision support value in the business context.

The requirement of achieving a macro-averaged F1 score of at least 0.7 on the test set is not met. Similarly, the recall for the lowest rating classes does not reach the target value of 0.75. These results indicate that the model is not yet suitable as a fully automated decision-making system, particularly for high-risk companies where misclassification could lead to financial losses. In addition, the gap between validation and test performance cannot be fully assessed due to missing validation metrics, limiting conclusions about generalization stability.

However, other metrics provide a more favorable picture with respect to the business objectives. The model achieves a Quadratic Weighted Kappa of 0.86 and a ±1 notch accuracy of approximately 72%. This indicates strong ordinal agreement between predicted and true ratings, meaning that most errors are small and close to the correct rating class. For a problem with 18 ordered rating classes, a mean absolute error of 1.15 rating notches is relatively low and suggests that the model captures the overall risk structure well.

From a business perspective, these results are sufficient to support analysts rather than replace them. The model can help reduce inconsistencies between analysts by providing a consistent baseline rating suggestion, especially for low- and medium-risk companies. This supports the objective of reducing manual review effort and speeding up rating decisions for new credit applications. However, the current performance is not sufficient to safely automate accept/reject decisions for high-risk companies, as the recall for the lowest rating classes remains below the required threshold.

6.2 Ethical aspects

The use of a machine learning model for corporate credit rating raises several ethical and risk-related concerns. As discussed in the AI risk assessment, historical credit ratings may already contain human bias, which the model can inherit and reinforce. In addition, incorrect predictions for low-rated companies pose a financial risk, while overly conservative predictions for high-quality companies may lead to lost business and reputational damage. The risk of concept drift is also relevant, as changing economic conditions can reduce model reliability over time if not properly monitored.

The bias analysis using SIC codes shows that model performance varies across industry groups, with some sectors exhibiting higher prediction errors than others. Although no strong systematic directional bias was observed for most groups, this uneven performance may still disadvantage certain industries if the model is used without safeguards. To mitigate these risks, the model should be deployed as a decision-support system with human oversight, particularly for high-risk cases, and its performance should be regularly monitored across industries and over time.

6.3 Monitoring plan

During deployment, several aspects of the credit rating model must be continuously monitored to ensure reliable and responsible operation. First, model performance metrics such as accuracy, mean absolute error (MAE), and ±1-notch accuracy should be tracked over time. A significant deterioration in these metrics, for example an increase in MAE above 1.5 or a drop in ±1-notch accuracy below 60

Second, data and distributional changes should be monitored to detect concept drift. Shifts in the distribution of financial ratios or rating predictions compared to the training data indicate changing economic conditions and may invalidate the model. Additionally, fairness across industries, monitored using SIC-based performance metrics, should be reviewed regularly. A sustained increase in error or directional bias for specific SIC groups should trigger further investigation and potential model adjustment. Finally, human override rates and analyst feedback should be monitored; a sharp increase in overrides is a strong signal that the model's recommendations are no longer aligned with expert judgment and require intervention.

6.4 Reproducibility reflection

The project documents the data source, use case, modeling approach, and evaluation metrics clearly, which supports reproducibility at a conceptual level. The use of standard libraries and well-defined performance measures further helps ensure that the overall workflow can be understood and repeated.

However, full reproducibility is limited by missing details such as exact preprocessing steps, hyperparameter settings, random seeds, and fixed dataset versions. Without this information, rerunning the analysis may lead to slightly different results. Providing more detailed configuration and experiment tracking would further improve reproducibility.

7 Conclusion

7.1 Comments beyond the provenance graph

All required information should be in the provenance graph.

7.2 Overall findings and lessons learned

This project demonstrates that machine learning models can provide meaningful support for corporate credit rating tasks by capturing the ordinal structure of ratings and producing consistent predictions. While traditional classification metrics such as accuracy and macro-averaged F1 score are relatively low due to the large number of ordered classes, ordinal metrics such as mean absolute error and quadratic weighted kappa show strong performance, indicating that most prediction errors are small and close to the true rating.

A key lesson learned is that business suitability cannot be judged by a single metric. Although the model does not fully meet all data mining success criteria, it still delivers value as a decision-support tool when combined with human oversight. Additionally, fairness and industry-specific performance analysis are essential, as model behavior can vary across sectors. Overall, careful metric selection, bias monitoring, and controlled deployment are crucial for responsible and effective use of machine learning in credit risk assessment.

7.3 Feedback

What I liked about this exercise is working with a real dataset and just seeing the whole machine learning process, but I think it shouldn't be mixed with Provenance. For someone who have never worked with it,it just makes the exercise more complicated. Split it in two exercise, one for Provenance and one for machine learning.