Credit Risk Modeling - ILAB OeNB

outstanding style and structure

By Anastasia Pryshchepa, Leonid Kuzmishin,
Tristan Leiter, Martin Yago Tortajada

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

ÖNB OESTERREICHISCHE NATIONALBANK EUROSYSTEM

Motivation and Research Tasks

1. Develop and evaluate supervised credit scoring models for Austrian SMEs using raw balance-sheet data.

2. Does the use of financial ratios improve model performance and generalization compared to the raw-data approach?

3. How do time-series dynamics and lagged variables influence the model's predictive power and feature importance over time?

Balance Sheet Data

Before we look further into data exploration, we explore how the variables are constructed.

mistake in the translation?
current assets =
umlagevermoegen

Balance-sheet formulas:

Sum to	Item 1	Item 2	Item 3	Item 4	Item 5
Total assets (f1)	Invested Capital (f2)	Current assets (f3)			
Current assets (f3)	Inventories (f4)	Cash (f5)	X		
Equity (f6)	Retained earnings (f7)	Net profit (f8)	Retained earnings (f9)	X	
Total assets (f1)	Equity (f6)	Provisions (f10)	Liabilities (f11)	X	

Table 1: Overview of basic accounting.

Sum to	Deviation greater than 5%	Number of Observations
Row 1	3.18%	7,760
Row 4	7.06%	17,250

Table 2: Deviations in the dataset concerning basic accounting.

Balance Sheet Data

Before we look further into data exploration, we explore how the variables are constructed.

Balance-sheet formulas:

Sum to	Item 1	Item 2	Item 3	Item 4	Item 5
Total assets (f1)	Invested Capital (f2)	Current assets (f3)			
Current assets (f3)	Inventories (f4)	Cash (f5)	X		
Equity (f6)	Retained earnings (f7)	Net profit (f8)	Retained earnings (f9)	X	
Total assets (f1)	Equity (f6)	Provisions (f10)	Liabilities (f11)	X	

Table 1: Overview of basic accounting.

Sum to	Deviation greater than 5%	Number of Observations
Row 1	3.18%	7,760
Row 4	7.06%	17,250

Table 2: Deviations in the dataset concerning basic accounting.

missing positions?
row 1:
Leitner: "much more deviation than i expected" .
ADD also accrued expense + differed assets
Around 50 to 100 dont fulfill this equation.
Depending on the size of the firm, different reporting requirements, some items missing.

SIMPLY MAKE ALL THE ITEMS POSITIVE.
ROw 3 has also problem with EQUITY; negative equities. Earnings is included in the net profit. Gewinnrücklage is also included.

IN A NUTSHELL: replace with 4 formulas we will get from Leitner

Balance Sheet Data

Examples of inconsistent observations.

Sum to	Deviation greater than 5%	Number of Observations
Row 1	3.18%	7760
Row 4	7.06%	17250

Table 2: Deviations in the dataset concerning basic accounting.

ID	Size	sector	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	y
256642	Tiny	service	**52**	**59**	**11**	0	10	-27	0	-32	0	1	71	0
85999	Tiny	construction	**1**	0	1	0	0	**105**	0	70	75	**1**	**18**	0

Table 3: Inconsistent observations.

Dependent variable

The dataset includes an extreme class imbalance (0.8% of observations are defaults).

Imbalanced Dataset

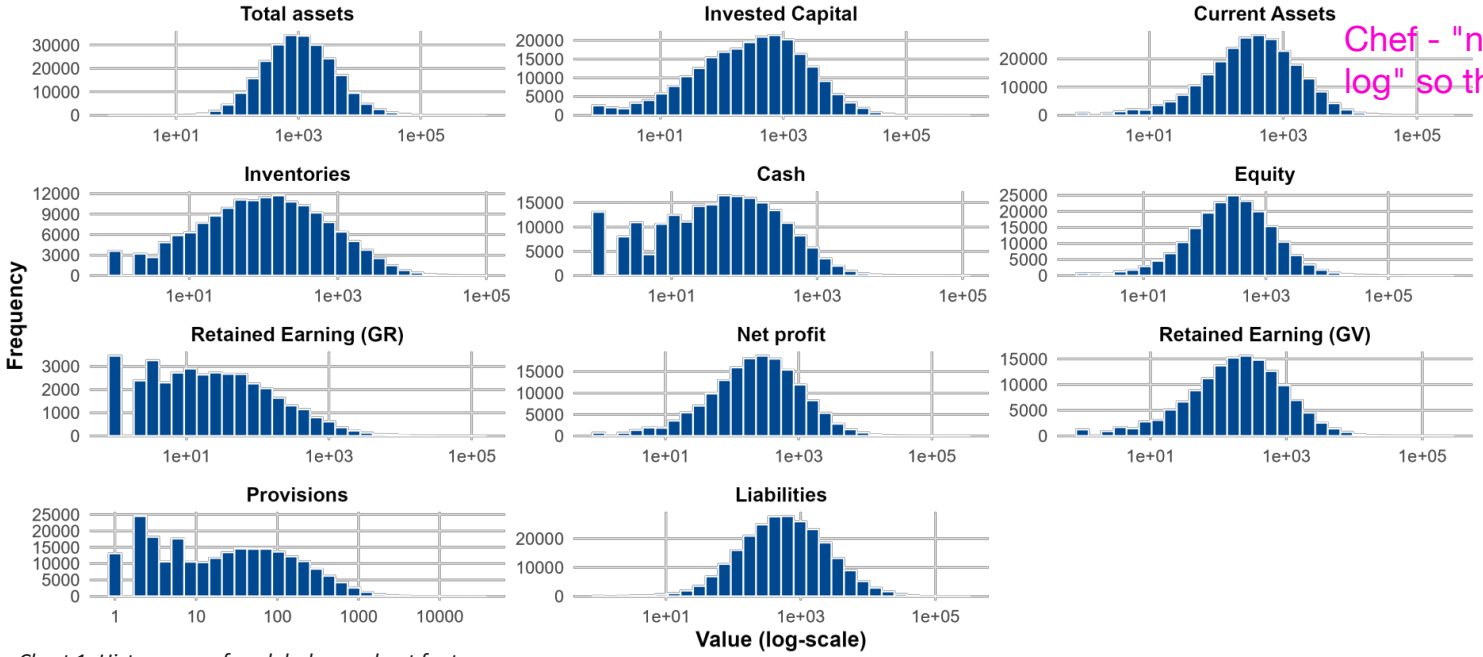
Default Status	Number of Firms	Proportion
1 (Default)	2,096	0.008576
0 (No Default)	242,305	0.991424
Total	**244,401**	**1.000000**

Table 4: Absolute and relative distribution of the dependent variable.

The **imbalanced dependent variable** can be a **challenge** for **model development**.

Feature distribution

Raw financial data exhibits a heavy right tail. We log-transform the raw data to visualize their distributions.



Chef - "normally you transform with the log" so this RESULT IS OK.

Chart 1: Histograms of each balance-sheet feature.

- Most balance-sheet features show a log-normal distribution after transformation. Without the log-transformation, the values would be too uneven to visualize properly. Additionally, it will be crucial for the logistic regression model.
- Only retained earnings (GR) and provisions deviate considerably.

Extreme values

Since the data set has negative values and zeros, we used signed log transformation to visualize the outliers.

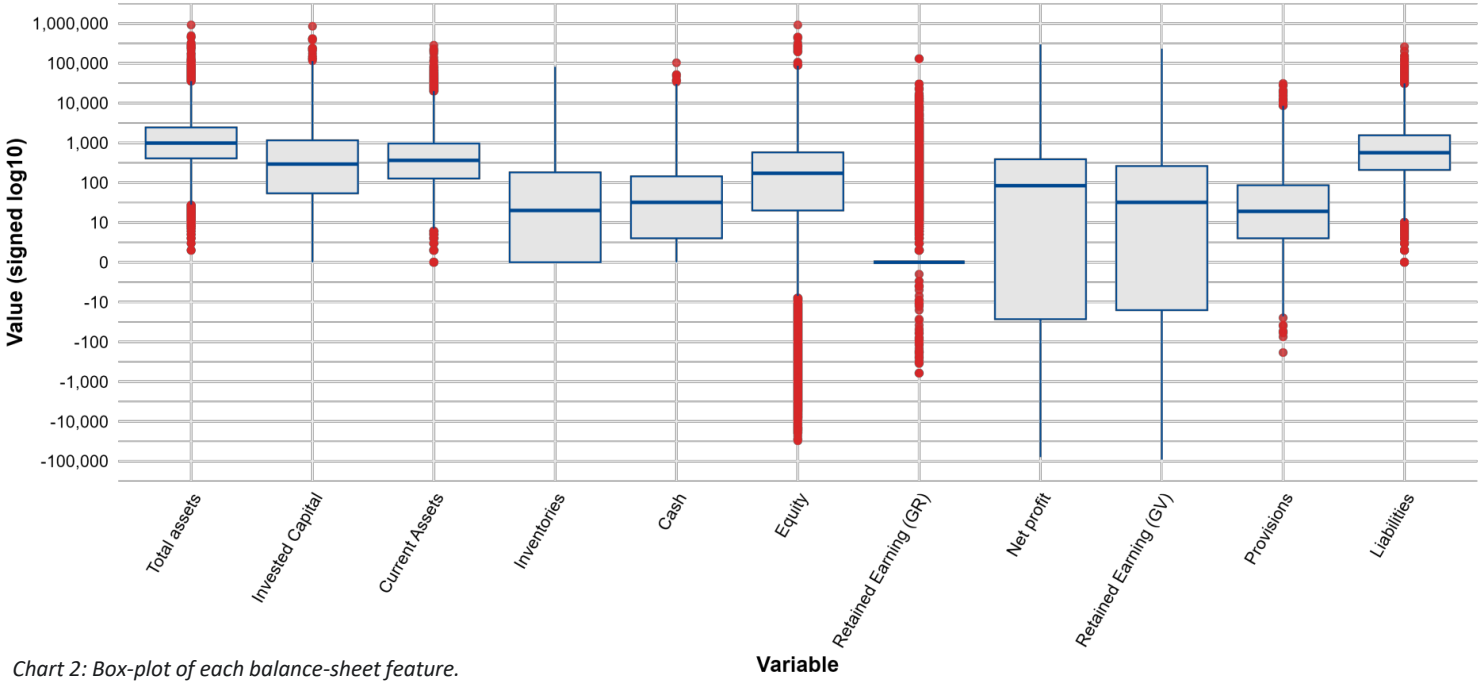


Chart 2: Box-plot of each balance-sheet feature.

- All predictors from the Balance Sheet are highly skewed and heavy-tailed distributed. Equity (f6), Retained Earnings GR (f7) have dense clusters of extreme values. Presence of numerous outliers suggests considerable variability in the underlying Balance Sheet Data.
- To deal with outliers we can apply robust scaling or weight of evidence.

Feature variation

All balance sheet features show a large discrepancy of the mean and standard deviation.

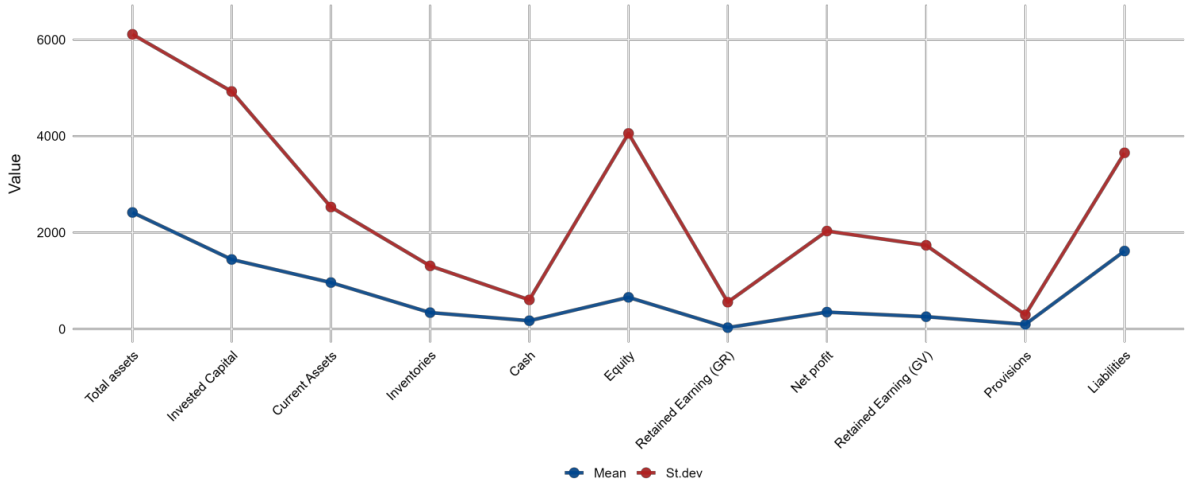


Chart 3: Mean and Standard-deviation of the balance-sheet features.

Variable	Mean	St.dev	Skewness
Total assets	2416.3	6109.6	33.5
Invested Capital	1441.6	4925.5	40.9
Current Assets	963.0	2529.3	28.3
Inventories	339.7	1308.6	13.9
Cash	171.7	604.4	40.8
Equity	656.5	4055.1	88.0
Retained Earning (GR)	28.4	556.4	164.4
Net profit	350.2	2031.0	44.2
Retained Earning (GV)	254.7	1735.8	36.2
Provisions	96.6	294.5	26.5
Liabilities	1616.4	3652.8	10.1

Table 5: Summary statistics for the balance-sheet features.

- There is high variability in each feature as the standard deviation is greater than the mean for every feature.
- Total assets (f1), invested capital (f2), net profit (f8), retained earnings (f9) and liabilities (f11) all showcase large spreads, which implies extreme values in the upper tail.
- Generally, the data is not centered around one value and includes many extreme observations.

- All distributions show a positive skewness.
- Equity (f6) and retained earnings (GV; f7) both have a particularly high skewness, implying that a few large observations dominate the shape of the distribution.

We **standardize** the features before modelling.

Dependency structure

We identify several key features, which separate the dependent variable well.

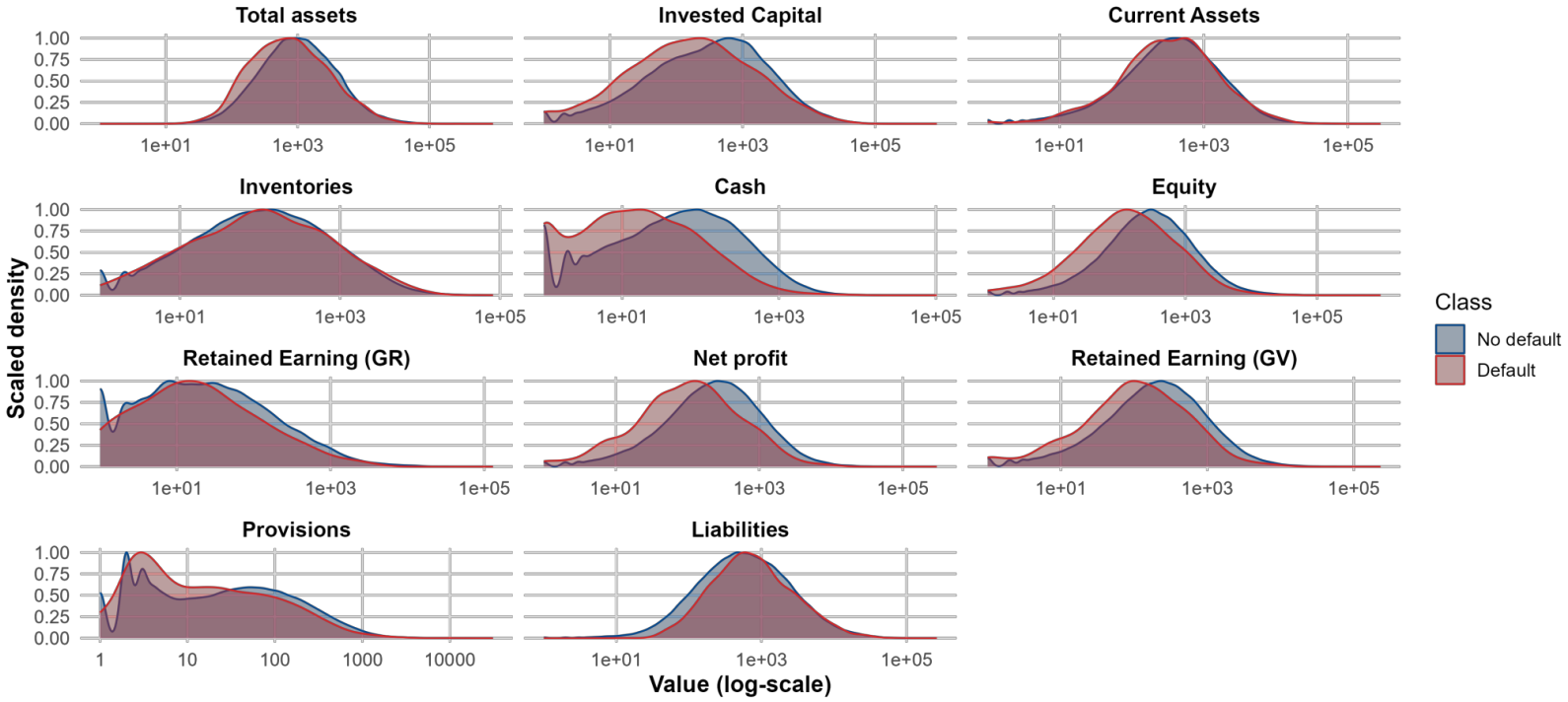
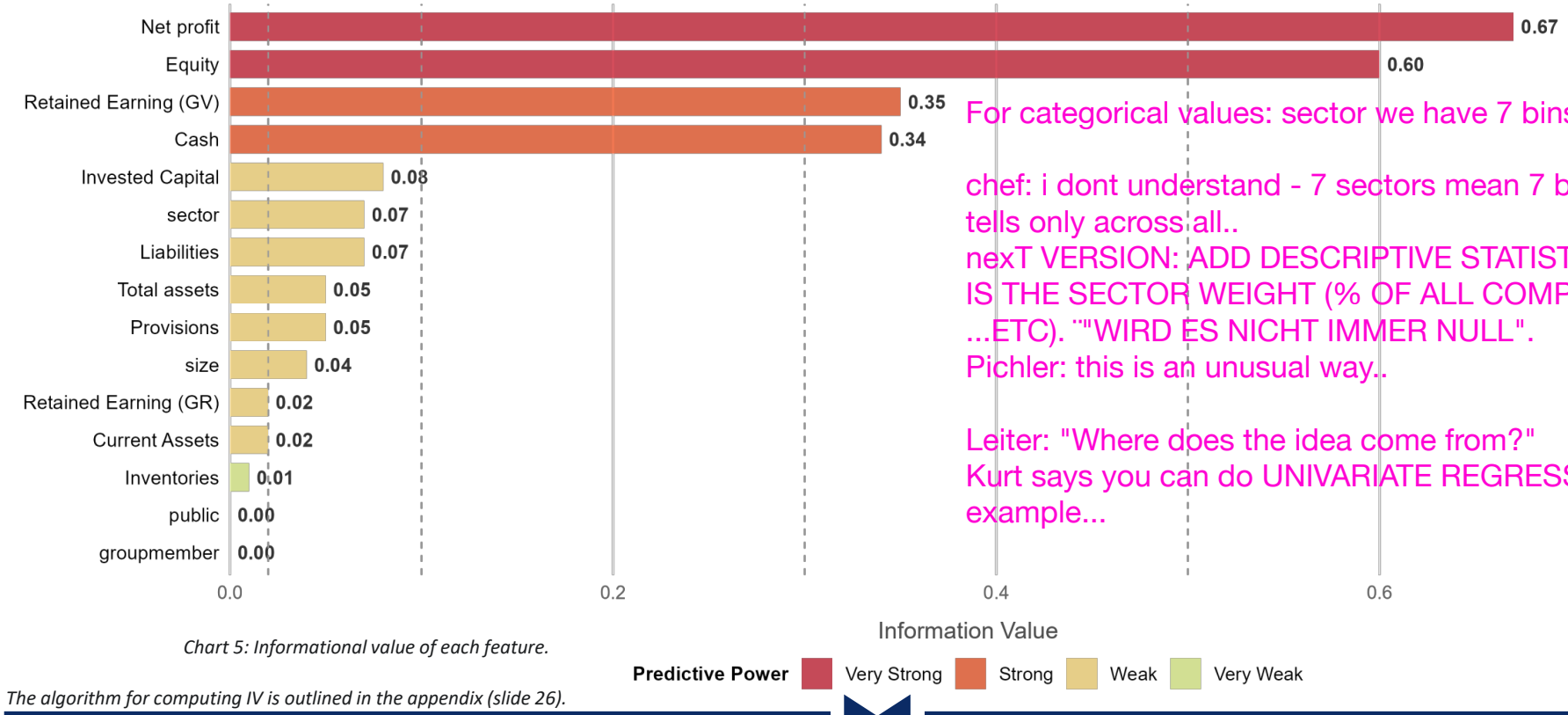


Chart 4: Distribution of each feature dependent on default status.

Defaulting companies tend to have **lower cash** and **equity** levels, as well as **less invested capital, net profit** and **retained earnings**!

Informational value

Information Value (IV) analysis identifies Profitability and Equity as the dominant drivers of creditworthiness.



For categorical values: sector we have 7 bins --

chef: i dont understand - 7 sectors mean 7 bins. This tells only across all..
nexT VERSION: ADD DESCRIPTIVE STATISTICS, WHAT IS THE SECTOR WEIGHT (% OF ALL COMPANIES E.G. ...ETC). ""WIRD ES NICHT IMMER NULL".
Pichler: this is an unusual way..

Leiter: "Where does the idea come from?"
Kurt says you can do UNIVARIATE REGRESSION, for example...

Chart 5: Informational value of each feature.

The algorithm for computing IV is outlined in the appendix (slide 26).

Similarly, **net profit**, **equity**, **retained earnings (GV)**, and **cash** are highly predictive.

Significance tests

Welch's t-tests confirm that differences in financial means between groups are statistically significant, not random.

CHEF: when you have a categorical TRUE-false (actually binary).
ARE THE logs applied to ratios really priority?

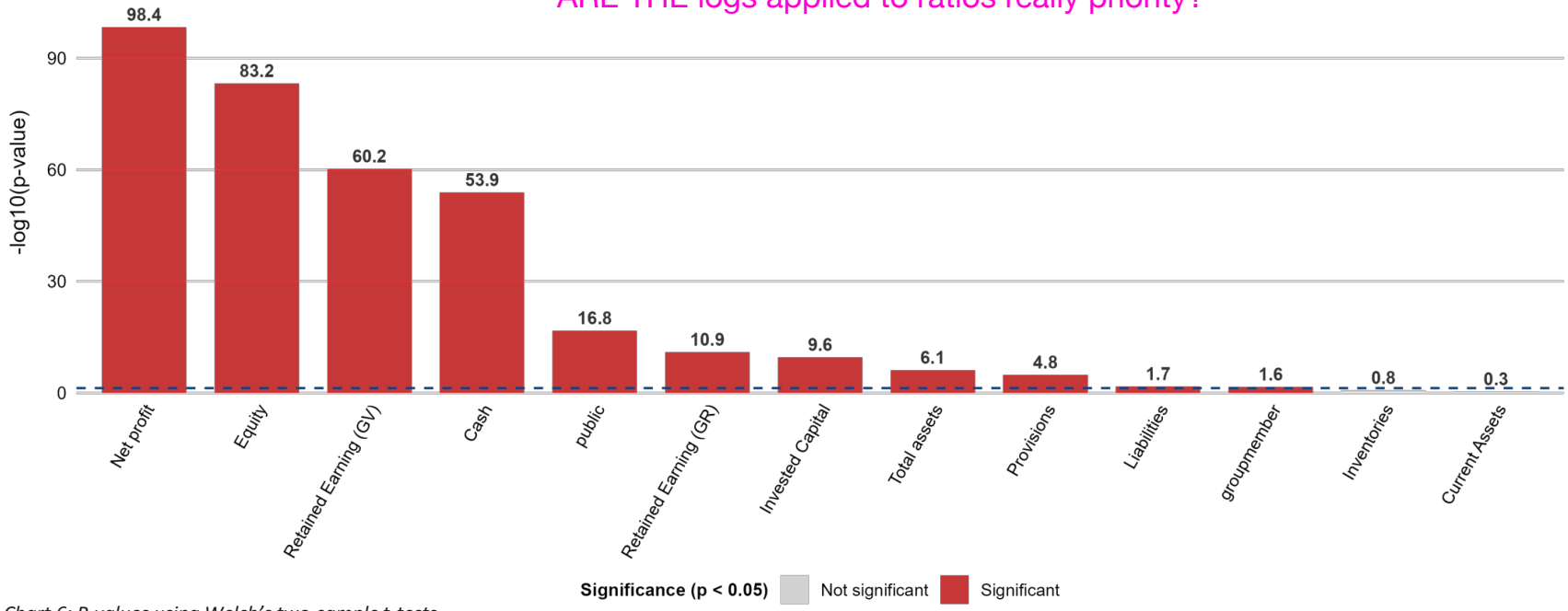


Chart 6: P-values using Welch's two-sample t-tests.

Only **inventories** and **current assets** show weaker statistical separation.

Multicollinearity

Strong correlations within 'Scale' and 'Profitability' clusters might be an issue for modelling.

Summary:

Company Scale Cluster:

- Invested Capital (f2) and total assets (f1) show a correlation of 0.92, indicating that they provide almost identical information.
- Equity (f6) and liabilities (f11) each also have a high correlation with total assets (0.79 and 0.74 respectively).

Profitability Cluster:

- Net profit (f8) and retained earnings (GV; f9) exhibit a correlation of 0.91.

pichler: ""WHAT IS THE MOTIVATION?"
tristan: " avoid noise"
pichler: "What noise?" Techniques in ml are constructed to deal with multicolinearity
The correlation of levels is expected to be high .
tristan:
Chef: It would make sense TO TAKE THIS PATH if you followed OLS, classical approach..
PARSIMONIUOS MODEL.
But in ML i dont care whether the model is parsimonious at all.
LEonid: we only have 200k observations
Nastia: for us to understand
Pichler: CORRELATIONS- REPEAT WITH THE STANDARDISED !

Multicollinearity:

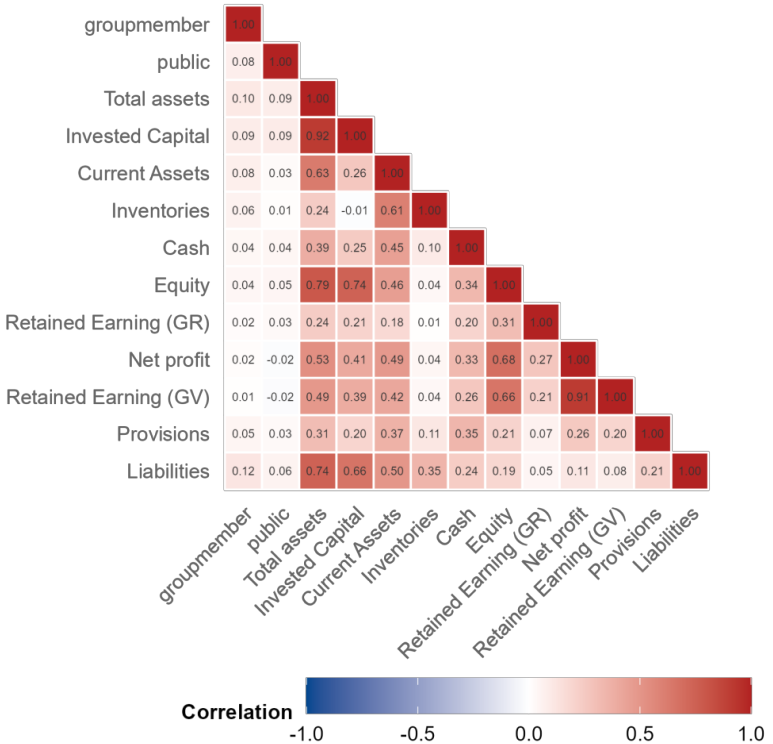


Chart 7: Correlation of all features with each other.

NUT-SHELL: dont remove variables

Feature Selection

We prioritize features with high IV while removing redundant variables to maximize model parsimony.

Summary:

	Feature	IV	Welch t-test*	Importance
f1	Total Asset	0.05	8	Medium
f2	Invested Capital	0.08	7	Medium
f3	Current Assets	0.02	13	Low
f4	Inventories	0.01	12	Low
f5	Cash	**0.34**	**4**	**High**
f6	Equity	**0.60**	**2**	**High**
f7	Retained Earnings (GR)	0.02	6	Medium
f8	Net profit	**0.67**	**1**	**High**
f9	Retained Earnings (GV)	**0.35**	**3**	**High**
f10	Provisions	0.05	9	Medium
f11	Liabilities	0.07	10	Medium
	Public	0.00	5	Low
	Sector	0.07		Low
	Size	0.04		Low
	Groupmember	0.00	11	Low

Table 6: Summary of all features and their importance.

**The column contains a ranking of the most significant p-value (=1) to the least significant one (=13).*

Conclusion:

- We propose dropping **total assets** since it is of medium importance, but highly correlated with the important feature invested capital.
- Additionally, we also drop **retained earnings (GV)** since it is highly correlated with net profit.

Dropping the **features** refers to the **first use case**.
For now, we leave it open whether to include them or not for the models with the ratios.

CHEF: DO EDA
- HOW MANY YEARS
- HOW MANY FIRMS PER sector
- DEFAULTS and NOn-def

and all the things that are useful for the stratification.. think ahead.

Data Splitting and Sampling

We split the data 70/30 and use stratified sampling to preserve the ratio of defaults to non-defaults.

0.8% Defaults

leitner: WHAt are conditions for the stratified sampling?
CHEF: this is an art. What are the STRATIFIERS?
You want to preserve the DR - sure ok. byt how exactly? Stratified per sector? Per .. ?
THIS IS THE CORRECT PATH but we want to add: THINKABOUT ALTERNATIVE STRATITIFYING STRATEGIES. Will you present only DEfault and Sector? TAKE into account size?
Do you do the sector cross .. in parallel?

Stratified Sampling

Training Set (70%)
N = 171,081
Preserves 0.8% Ratio.

Test Set (30%)
N = 73,320
Preserves 0.8% Ratio.

leitner
split ACCORDING TO sector -if you have enoiugh observations. most of the firms here have the classical time-series. IF you do some modelling inthe decision tree domain: Think of THE evolution over time. Dont use the same firm for training and evaluation. TAKE THE FIRM WITH ITS FULL TIME DIMENSION.
OUT OF TIME is not the same problem as out of sample.

Total Dataset
N = 244,401

Popular Metrics

There are a variety of different loss measures, which could be used.

Precision / Recall.

Simple metrics can be misleading. Basically a „naive“ classifier that predicts all 0s or 1s, which can result in high scores on one dimension while failing completely in the other one.

CHEF:
JUST GIVE ME THE ROC AND AUC
CHEF: AUCs per SECTOR. Multivariate --> if i think about ML models, forward NN.. my interest is just robust models. Don´t leave predictors out.

F1 Score.

Values precision and recall equally. In our task, recall is critical, while precision is naturally low (1-2%). Using F1 sacrifices recall to boost precision.

F-beta.

Requires defining the relative cost of False Positives vs. False Negatives. Inferring these costs directly from the data is difficult and often arbitray.

RoC-AuC.

Often overoptimistic for imbalanced data. It can score two vastly different models with nearly identical high scores, masking performance issues.

ROC-AUC

The "easy" negative problem.

Scenario: 1000:1 Ratio.

Assume we have a dataset with 500,000 non-defaults and 500 defaults.

Model A (Better) **AUC: 0.999**

Mixes 1,000 non-defaults with 500 defaults, then ranks the remaining 499,000 “easy” non-defaults correctly.

Model B (Worse) **AUC: 0.990**

Mixes 10,000 non-defaults with 500 defaults, then ranks the remaining 490,000 “easy” non-defaults correctly.

Result: Absolute difference in AUC is only 0.009, but it fails to reflect that Model B is 10x worse than Model A.

PR-AUC

PR-AUC exposes the true performance gap masked by ROC-AUC in imbalanced data.

Better differentiation

- **Sensitivity to False Positives:** Unlike ROC-AUC, the Precision-Recall curve drops drastically when false positives are introduced among the top-ranked items.
- **Model Comparison/Selection:** PR-AUC clearly separates Model A from Model B (the previous example), providing a reliable signal for optimization.

Strategy:

PR-AUC will be used as the **primary metric** for **hyperparameter tuning**. ROC-AUC will be reported for standard compliance.

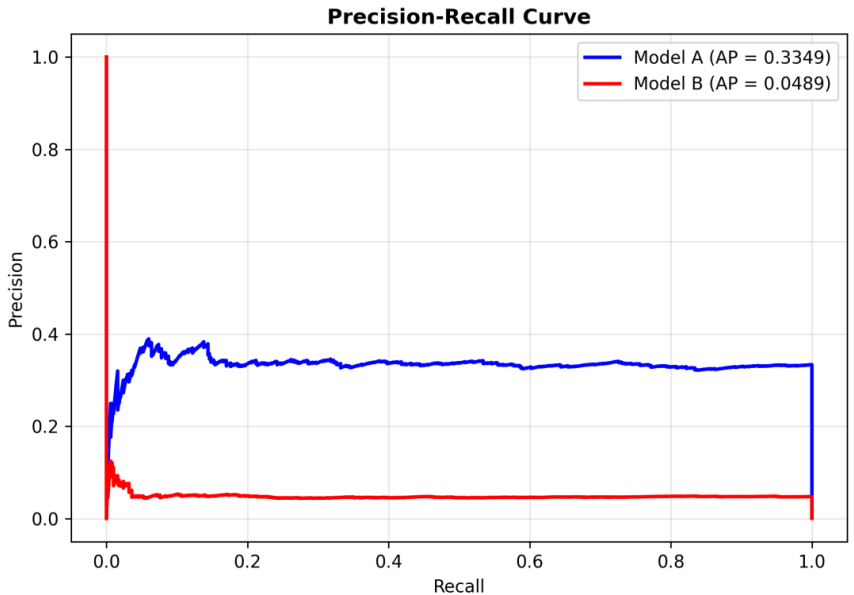


Chart 9: Precision-recall curve of Model A and B based on the previous slide.

Leonid: about the difference in the curves.
pichler: WHAT IS the major source of info input to decide?
Leonid: internet.
pichler: there is a huge literature about scoring methods. EXTREMELY INTERESTING AND DELICATE. comon knowledge - there is no CONSENSUS.
TO START; AUC. INSTItutionally wanted. At least for the first round of analysis. Further discussion in 2nd round, maybe extra analysis. Maybe not PR. Kurt is not expert either and we cannot really tell. Proceed with traditional AUC.
CHEF: portfolio dependent. THER IS A PAPER YOU MOIGHT WANT TO READ, ABOUT ROC PRESENTE BY OENB. Look up "PROPER SCORING RULES".
Certain statsictical feature, "BRIER score" . . BUT YOU ARE IN THE RIGHT TRACK. START WITH ROC.

Threshold considerations

For a confusion matrix we need a suitable threshold.

Analyzing Mistakes

Visualizing mistakes is valuable, but it requires selection of a specific decision threshold to classify predictions as 0 or 1.

Proposed Approach

Instead of guessing a beta for the F-beta score, we define thresholds on train based on Recall requirements:

- **Scenario A:** Fix Recall at 0.99 (to catch almost all defaults).
- **Scenario B:** Recall at 0.90 (to allow for a 10% miss-rate).

We then inspect the resulting **Precision, Recall** on test at these fixed operating points.

Open Questions

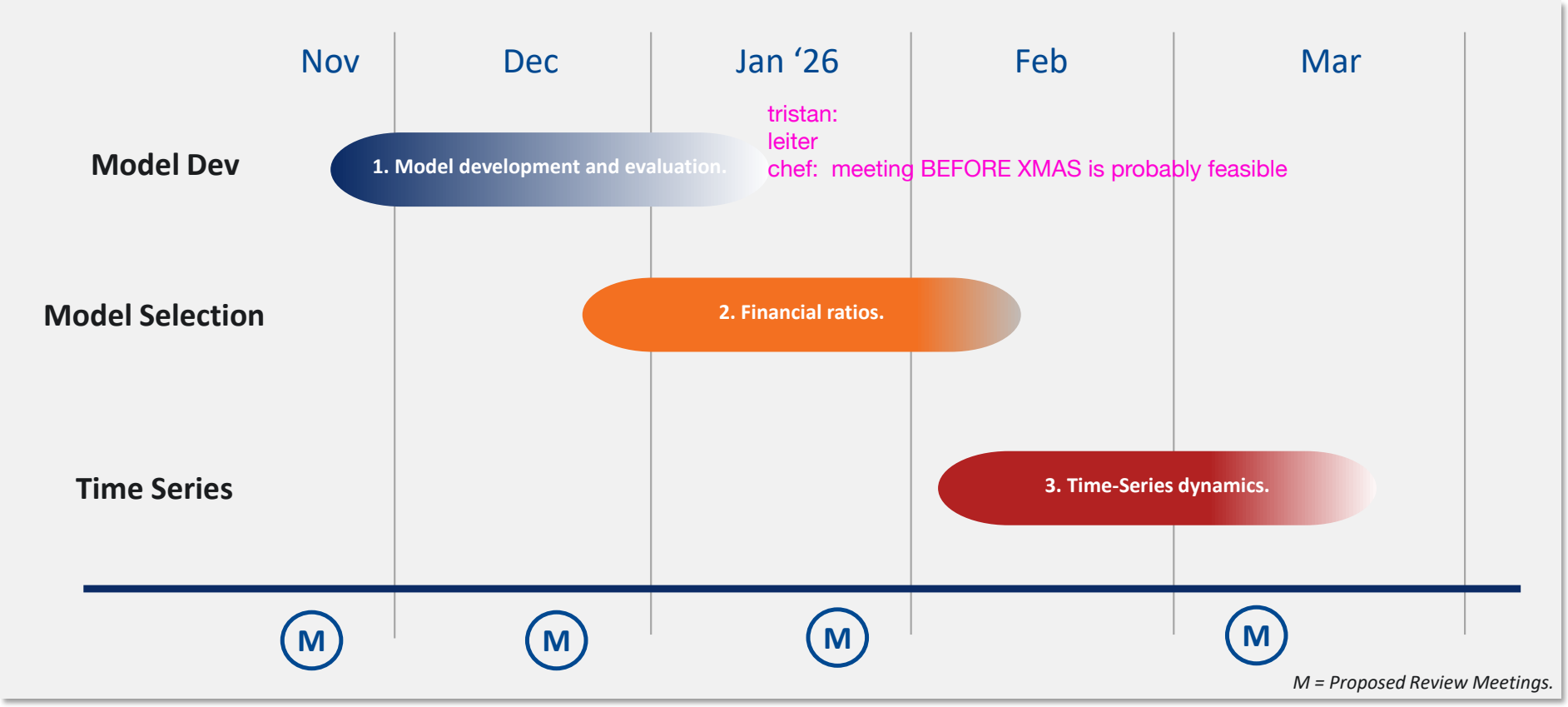
What is the optimal trade-off between FP and FN? Without explicit costs for False Negatives, anchoring to high Recall is the safest starting strategy.

For 73,000 observations of which 700 are positive, would we rather have:

- 25,000 predicted as negative, of which 25 are False Negative (focus on catching all positives)
- 43,000 predicted as negative, of which 85 are False Negative
- 58,000 predicted as negative, of which 200 are False Negative (focus on catching "easy" positives)

Project Roadmap

Timeline overview for implementation and delivery (Nov 2025 – March 2026).



Outlook: Data processing

Summary of data processing and open questions.

Categorical data

For now, we put them in the model, but we are considering
using, for example, target encoding. We will focus more on this aspect in the next step.

about "machines figuring out on their own".
This is not supervised learning.
CHEF: for now do just POSITIONS, TRY WITH SOME standardised RATIOS, .. only if you have so much time, then do some standard in between "PCA, auto-encoder, ..
Try things like "divide by
Replace with empirical percentile in the cum. distr function. for example, "this firm is the 5th in rank". It gives you a value between 0 and 1. Think about ways of normalising data.
Creative ideas. Robust scaling also good ! BRAVO NASTIA

Handling Imbalance

- Assign weights to observations
- Other techniques (for example balanced bagging)

Feature generation

Comparing models fit with and without ratios doesn't seem entirely "fair".

- Ratios add 15 more features (doubling *#.* of features)
- We can use deep feature generation to create 15 new features for the models fit without ratios
- This is also related to machines figuring stuff out on their own

DONT DO DEEP FEATURE GENERATION.

chef & leiter: dont do weighting techiques FOR THE IMBALANCE.
Upsampling problematic. Just start with the data as is.
IF you use whatever packages, you have to know what you doing-

Open Questions

Are we allowed to apply deep feature generation? Strictly speaking it is not entirely covered under supervised learning.

Outlook: Modelling

Summary of the next steps for modelling and open questions.

GLMs

- Logistic regression
- Regularization

CHEF : these models look good.

Greedy search? Should be fine, Kurt will also give his opinion.

Tree-based models

- Random forest
- Boosting

NNs

- Existing architecture(s)

Open Questions

Usually we are taught about applying a grid search. However it is computationally expensive and sort of "shooting in the dark".

- Can we use Bayesian optimization algorithms and maybe zero-shot algorithms?
- There are a lot of concepts in NNs: Implementing the NN manually seems redundant given available tools. Is it okay to use a library instead?

Outlook: Modelling via AutoML

A possible high-performance benchmark with an automated end-to-end pipeline.

Summary of AutoML

- Variety of different models (variations of Boosting, NN, and many more)
- Bagging and cross-validation
- Hyperparameter tuning
- Models based on output of other models
- Ensemble of all these models

AUTOML - CHEF:
TRY IT OUT!

What it needs

- Data
- Some implementations can even work with raw data, handling NAs and other data issues (via value imputation, feature encoding, and many other applications)

Open Questions

Can we employ the AutoML toolbox?

Appendix 1 – Scaling methods

Robust Scaling

$$x_{scaled} = \frac{x - median\ (x)}{IQR(x)}$$

Robust Scaler transforms centers the variable and interquartile range scales the spread. This way, it reduces the influence of extreme values without removing them, preserves the underlying structure. The method is suitable for using in tree-based models or boosting.

Weight of evidence

$$\text{WOE} = \ln\left(\frac{\%\text{Good}}{\%\text{Bad}}\right)$$

A large positive WOE means that the bin is strongly associated with non-defaults, while a large negative value demonstrates the opposite. After calculating WOE, the original value is replaced by WOE of the bin it belongs to. The method is suitable for using in glm models and iff the outcome is binary.

Appendix 2 – Stratified sampling algorithm

1. Group Data by Stratum: The function first identifies all unique values in the specified strata column (y). In this case, it creates two distinct groups: one for all the defaulting firms (y = 1) and another for all the non-defaulting firms (y = 0).

2. Sample Proportionally within Each Group: The function then takes an 80% sample from the defaulting firms group and a separate 80% sample from the non-defaulting firms group.

3. Create the Training Set: These two separately sampled subsets (80% of defaults and 80% of nondefaults) are combined to form the final training set (Train).

4. Create the Test Set: The remaining 20% of the defaulting firms and the remaining 20% of the non-defaulting firms are combined to form the final testing set (Test).

Appendix 3 – Information value algorithm

1. Binning (Grouping): The function first discretizes each continuous independent variable into a set of bins or groups (e.g., by quantiles). For categorical variables, each category is treated as a separate group.

2. Counting Goods and Bads: For each bin of a variable, the algorithm counts the number of defaults and non-defaults.

3. Calculating Proportions: It then calculates the proportion of the total goods and total bads that fall into that specific bin.

$$\%\text{Goods} = \frac{\text{Number of Goods in the bin}}{\text{Total number of Goods in the entire data set}};\ \%\text{Bads} = \frac{\text{Number of Badsin the bin}}{\text{Total number of Bads in the in the entire data set}}:$$

4. Calculating Weight of Evidence (WoE): The WoE for each bin measures how much the evidence in that bin supports one outcome over the other. A large positive WoE means the bin is strongly associated with non-defaults, while a large negative WoE means it is strongly associated with defaults.

5. Calculating Information Value: Finally, the total IV for the entire variable is calculated by summing a weighted value across all its bins. This single number represents the total predictive power of the variable.

$$\text{IV} = \sum(\%\text{Goods} - \%\text{Bads}) \cdot \text{WoE}$$

Appendix 4 – Outliers

Since the data set has negative values and zeros, we used signed log transformation to visualize the outliers. The transformation doesn't change the rank or ordering of the data, it only rescales it. Signed log keeps negative values negative, but compresses their magnitude in the same way as positive values.

$$\text{signedlog}(x) = \text{sign}(x) \cdot \log_{10}(1 + |x|)$$

THANK YOU FOR YOUR ATTENTION!