Developing an Internal Rating System

SE Applied Risk Management

Winter Term 2024/2025

Group 4 - Sonja Katzensteiner, Mikhail Kazakov, Sebastian Rous

Approach to creating a Rating Model

Baseline Approach

- Logistic Regression
- Standard in credit risk models
- Clear explainability

Machine Learning Approach

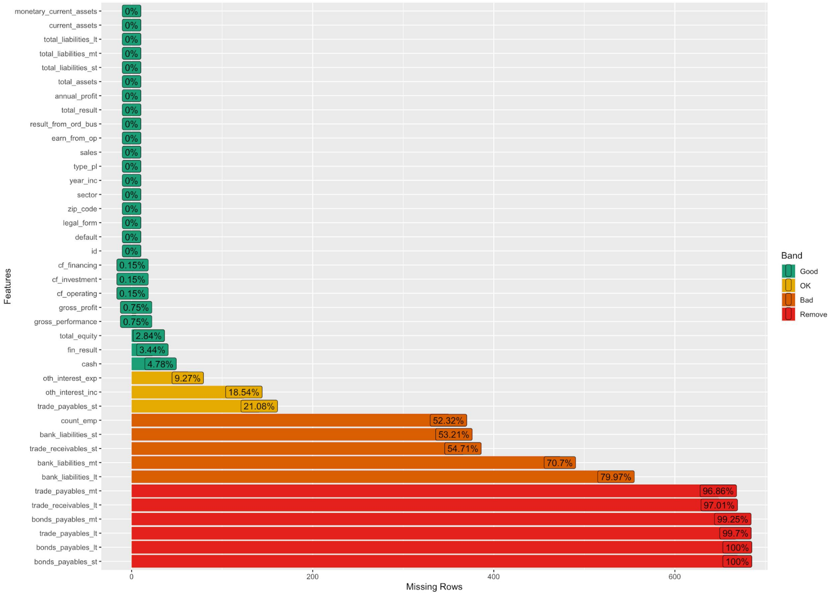
- Random Forest
- Gradient Boosting
- Useful for non-linear relationships

Data Management

Imputation of Missing Values

- Sales ← Gross Performance
- Gross Performance ← Sales
- Total Equity ← Total Assets - Total Liabilities
- Interest Expenses ← Total Liabilities x avg. Debt Interest
- Financial Result ← Total Result - Earnings from Op.
- Total Assets ← Total Liabilities + Total Equity

- Remaining NAs = 0



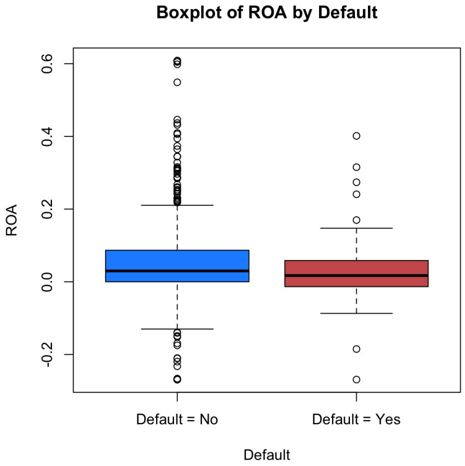
Feature Engineering

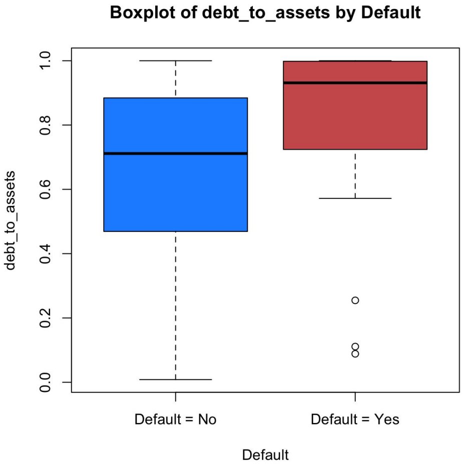
Profitability Ratios

- Ebit Margin
- Net Profit Margin
- ROA
- ROE
- Return on operating Profit

Solvency Ratios

- Debt/Assets
- Debt/Equity
- Interest Coverage





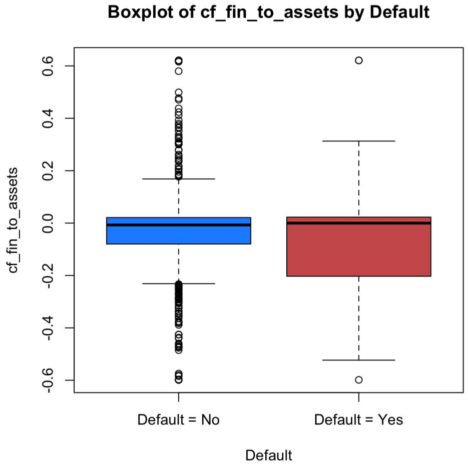
Feature Engineering

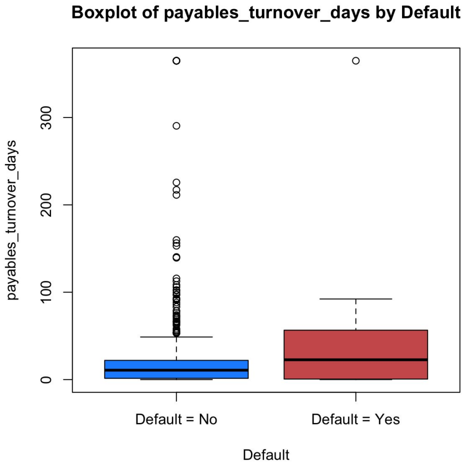
Cash Flow Ratios

- Operating CF to Debt
- Operating CF to Sales
- Investment CF to Assets
- Financing CF to Assets
- Cash Flow Coverage

Turnovers

- Payables Turnover Days
- Receivables Turnover Days





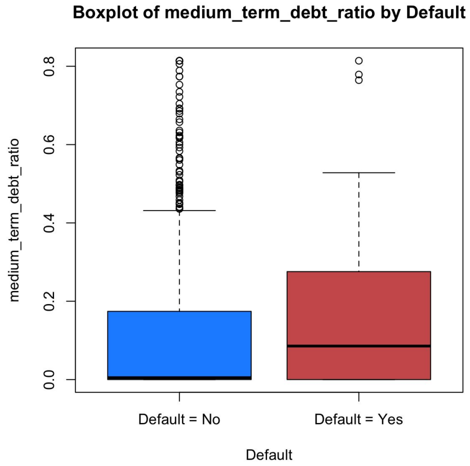
Feature Engineering

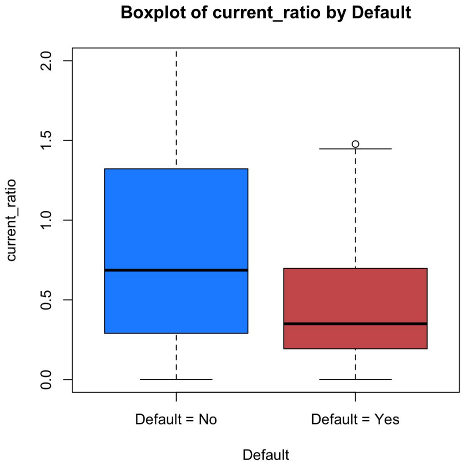
Leverage Ratios

- Short term Debt Ratio
- Medium term Debt Ratio
- Long term Debt Ratio
- Bank Debt Ratio
- Trade Payables Ratio
- Bond Debt Ratio

Liquidity Ratio

- Current Ratio
- Cash Ratio





Outlier Management

Winsorizing

Winsorizing: Cap extreme values
Minimum: 1% quantile
Maximum: 99% quantile

Max value for Payables/Receivables Turnover Days → 365

Winsorizing + Log Transformation

Winsorizing: Cap extreme values
Minimum: 1% quantile
Maximum: 99% quantile

Logarithmic Transformation:
transform skewed distributions
Operating CF to Sales, Interest Coverage, Return on operating CF, CF coverage ratio

Max value for Payables/Receivables Turnover Days → 365

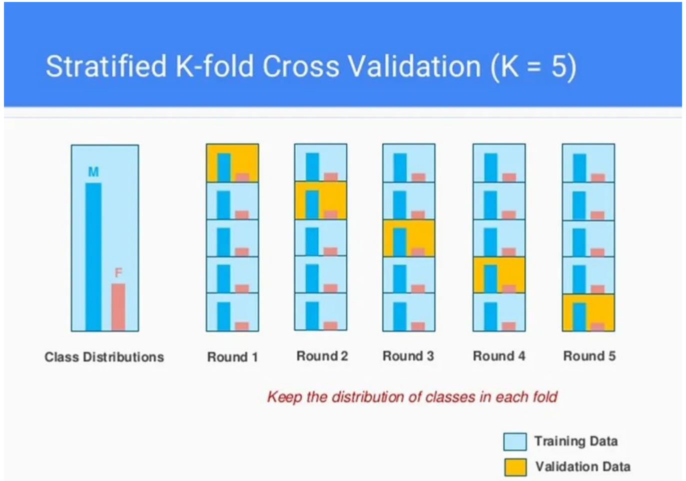
Model Training

Models Trained

- Logistic Regression
- Random Forest
- Gradient Boosting

Training Technique

- Repeated Stratified k fold Cross-Validation
- Optimizes model stability and prevents overfitting through diverse train-test splits.

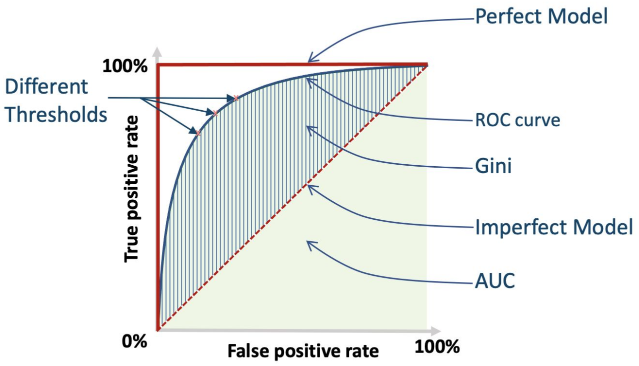


https://medium.com/@ompramod9921/cross-validation-623620ff84c2

Comparing Model Performance

Evaluation Metrics

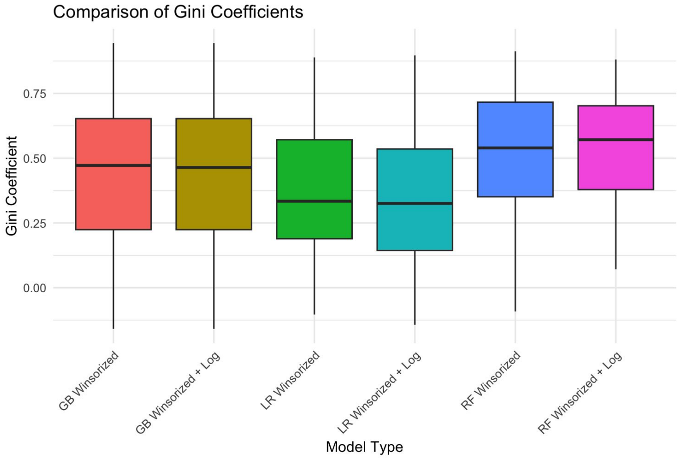
- ROC Curve and AUC (Area Under Curve)
- Mean Gini Coefficient as a proxy for predictive power.



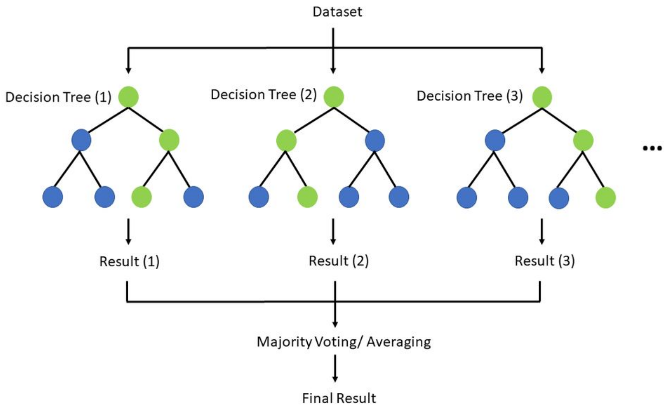
https://yassineelkhal.medium.com/confusion-matrix-auc-and-roc-curve-and-gini-clearly-explained-221788618eb2

Best Model

Random Forest (Winsorized + Log) best mean Gini coefficient with 0.53



Random Forest

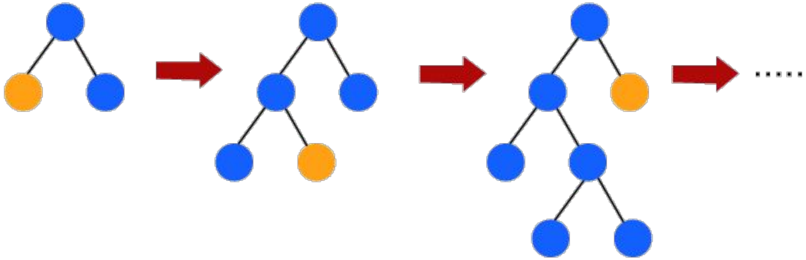


Mechanics

- **Many decision trees** train on shuffled data with **random features.**
- Each tree predicts **default** or **non-default**, majority vote decides.
- **Default risk** is based on the percentage of trees predicting “**Yes**”.

https://de.wikipedia.org/wiki/Random_Forest

Gradient Boosting



Mechanics

- The **first** tree makes a **prediction**, but there are **mistakes**.
- The **next** trees correct these **mistakes** by learning from how **errors** change (gradient)
- Step by step, **all** trees improve and combine for a strong final **prediction**

https://www.linkedin.com/pulse/introduction-gradient-boosting-machines-gbm-powerful-ensemble-sachin-fhdlc/

Why Random Forest Outperformed

- **Captures Non-Linearity:** Unlike Logistic Regression, which assumes linear patterns.
- **More Robust to Noise:** Avoids overfitting by averaging trees, unlike GBM, which struggles with noisy data.
- **Balanced Bias-Variance**: Avoided GBM's overfitting and LR's oversimplification
- **Reliable & Stable Predictions:** Majority voting ensures that no single tree dominates, leading to more consistent and interpretable risk assessments.

Making Predictions

Final Training

Train Final Random Forest model on the entire training data set (Winsorized + Log)

Test Data Preparation

Align test dataset types to the same as training dataset
Application of the same Data Cleaning, Feature Engineering, and Outliers handling to the test data set as to the training data we did earlier

Run model

Predict default probabilities on the cleaned test data set

Results

Top 10 predicted Company Defaults

id	C_289	C_211	C_550	C_448	C_539	C_847	C_471	C_517	C_857	C_086
DP	0.496	0.444	0.44	0.434	0.424	0.424	0.394	0.388	0.388	0.386

Gini Results

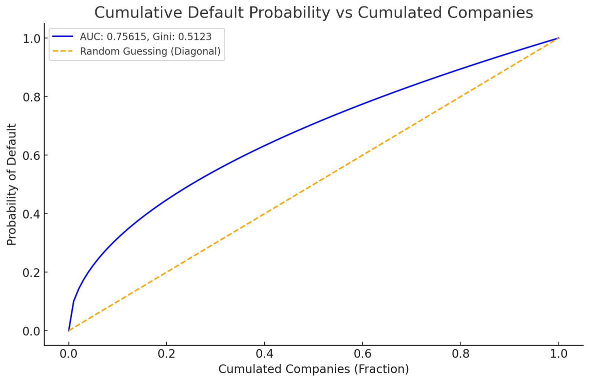
Out-of-Sample Gini is 0.5123

Gini = AUC x 2 -1

Area Under Curve 0.7561

Gini coefficient	<= 40 %	→ poor result
Gini coefficient	41% - 60 %	→ good result
Gini coefficient	> 60 %	→ excellent result

Source: VO Slides Part 1 - Applied Risk Management, P. 68



Limitations of the model

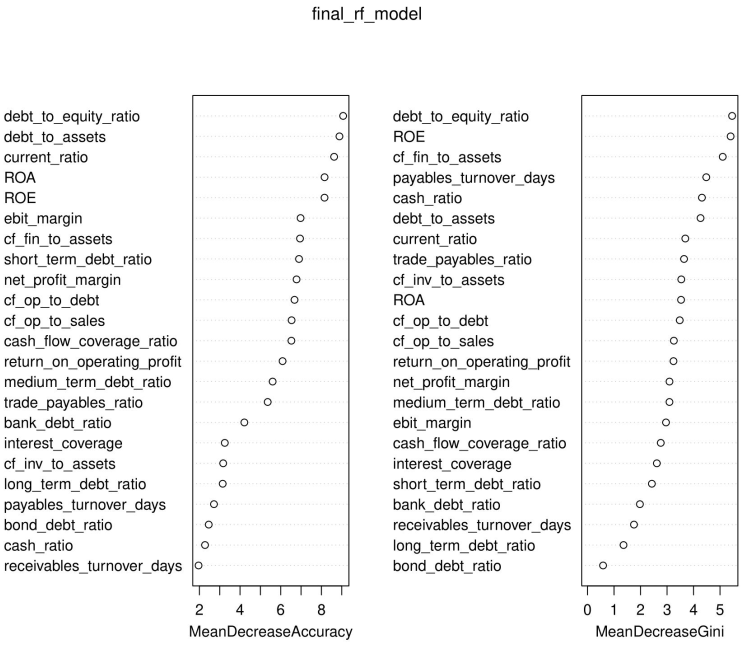
- Limited interpretability
- Extreme values even after data transformation
- Economically weird relations in random forests model
- Reliance on sales
- Overfitting in some cases

Importance of the variables

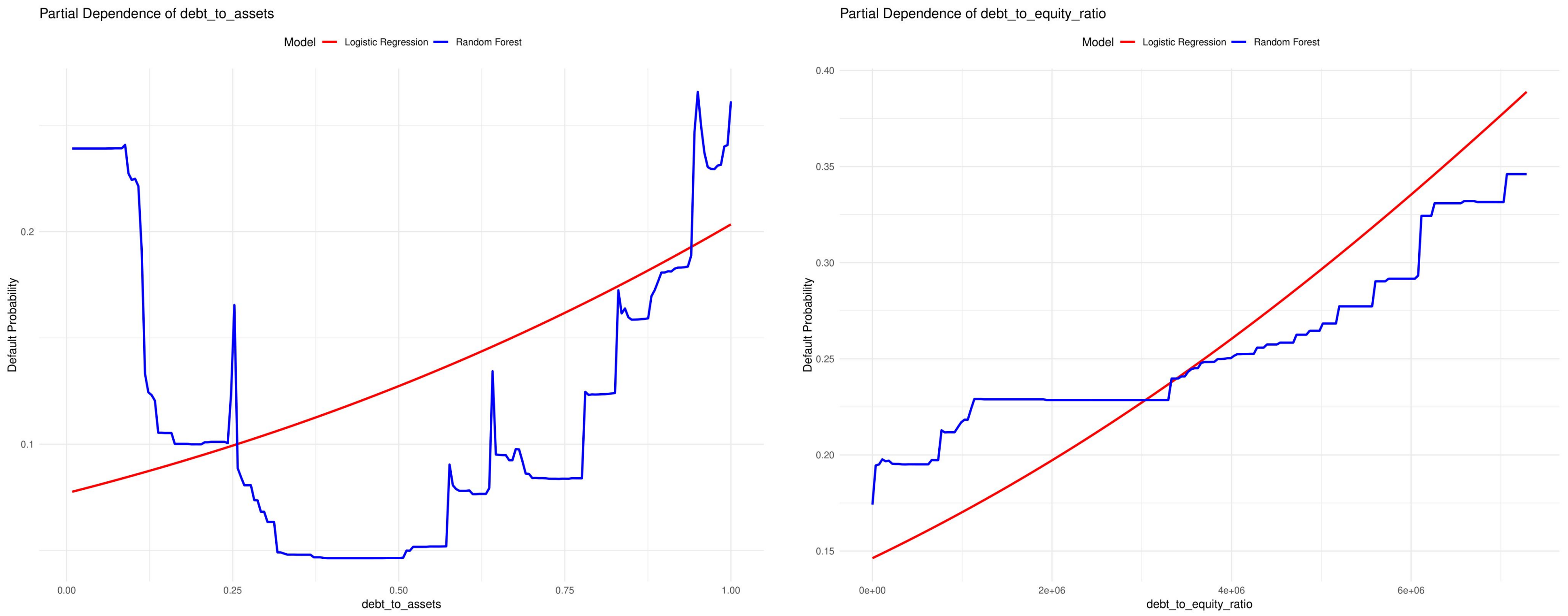
Logistic regression

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                1.219e+01  1.656e+01   0.736   0.4615
ebit_margin               -4.115e-01  4.114e-01  -1.000   0.3173
net_profit_margin         -2.586e-02  3.648e-01  -0.071   0.9435
ROA                       -5.109e-01  1.904e+00  -0.268   0.7884
ROE                       -2.804e-02  1.004e-01  -0.279   0.7800
debt_to_assets             2.238e+00  1.063e+00   2.106   0.0352 *
debt_to_equity_ratio       3.600e-07  1.403e-07   2.566   0.0103 *
interest_coverage         -4.330e-01  7.667e-01  -0.565   0.5722
cf_op_to_debt              6.299e-01  8.107e-01   0.777   0.4371
cf_op_to_sales             1.306e+00  9.418e-01   1.386   0.1656
cf_inv_to_assets           8.885e-01  2.095e+00   0.424   0.6714
cf_fin_to_assets          -5.108e-01  1.210e+00  -0.422   0.6728
payables_turnover_days     1.643e-03  3.133e-03   0.524   0.6000
receivables_turnover_days -6.121e-03  7.053e-03  -0.868   0.3854
current_ratio              4.056e-02  1.097e-01   0.370   0.7117
return_on_operating_profit -1.108e-01  6.109e-01  -0.181   0.8561
short_term_debt_ratio     -1.492e+01  1.400e+01  -1.066   0.2866
medium_term_debt_ratio    -1.445e+01  1.441e+01  -1.003   0.3160
long_term_debt_ratio      -1.678e+01  1.411e+01  -1.189   0.2344
bank_debt_ratio            4.379e-01  8.365e-01   0.524   0.6006
trade_payables_ratio       1.999e+00  9.035e-01   2.212   0.0269 *
bond_debt_ratio            3.758e+00  3.796e+00   0.990   0.3221
cash_ratio                -1.151e+00  8.040e-01  -1.431   0.1524
cash_flow_coverage_ratio   1.975e-01  4.264e-01   0.463   0.6432
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

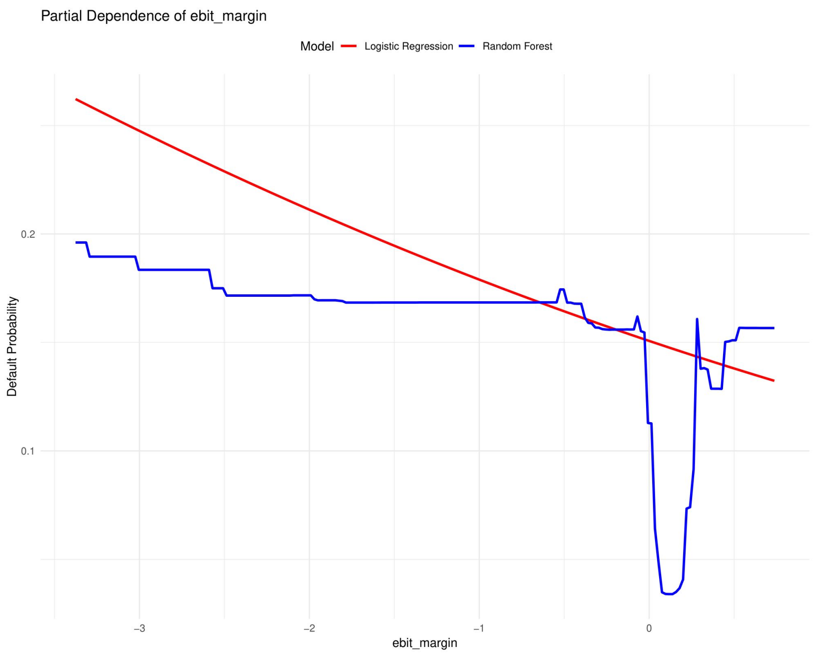
Random forests

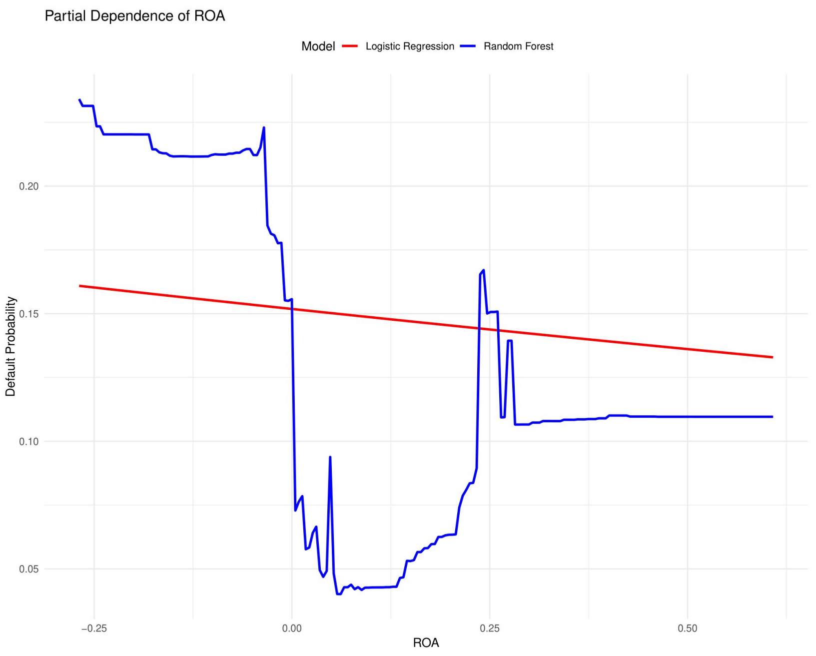


Partial dependence plots. Debt ratios

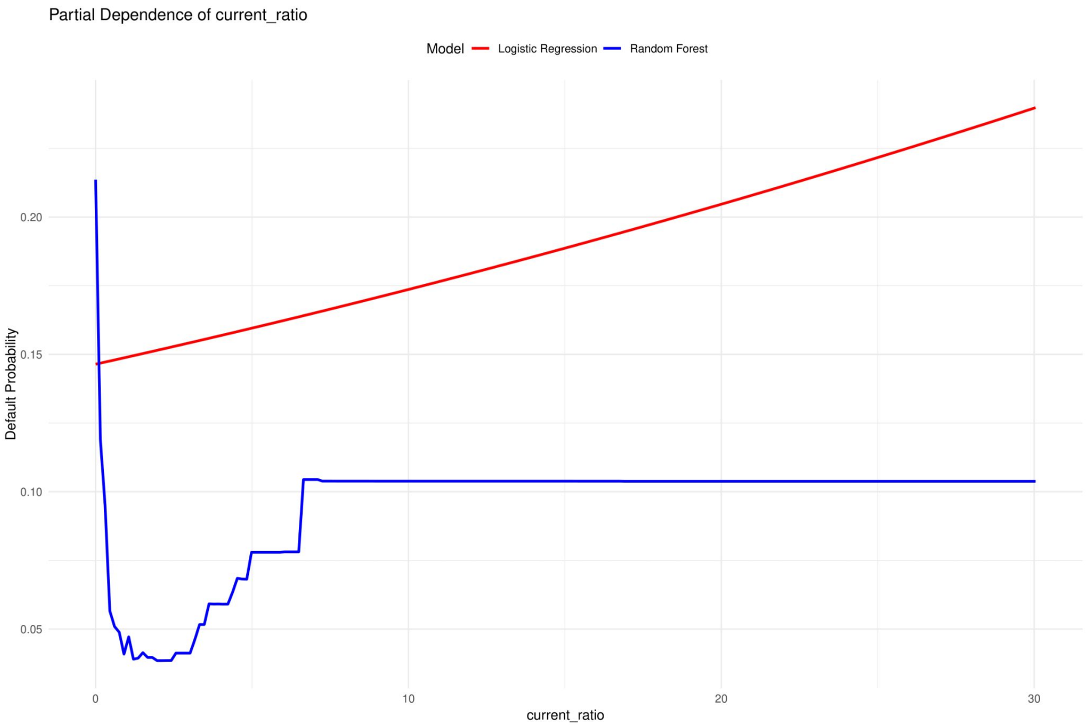


Partial dependence plots. Profitability ratios

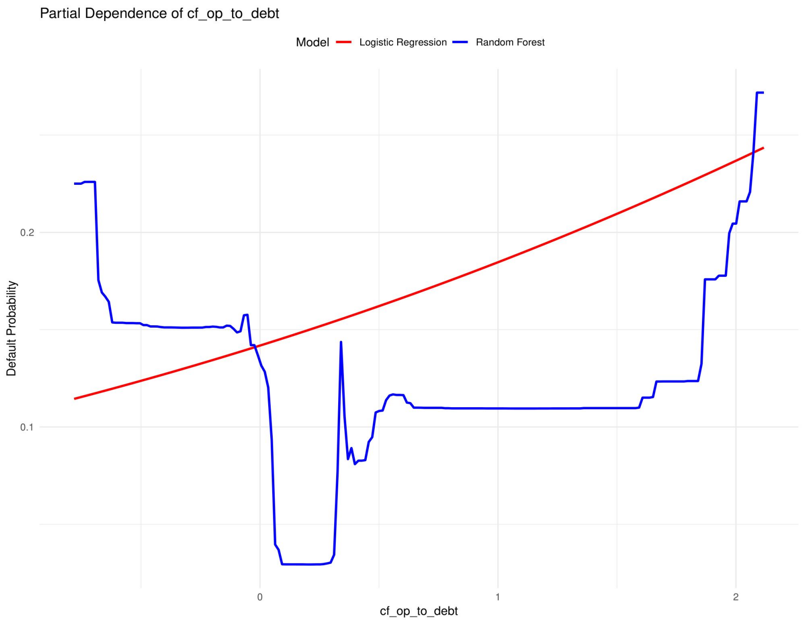


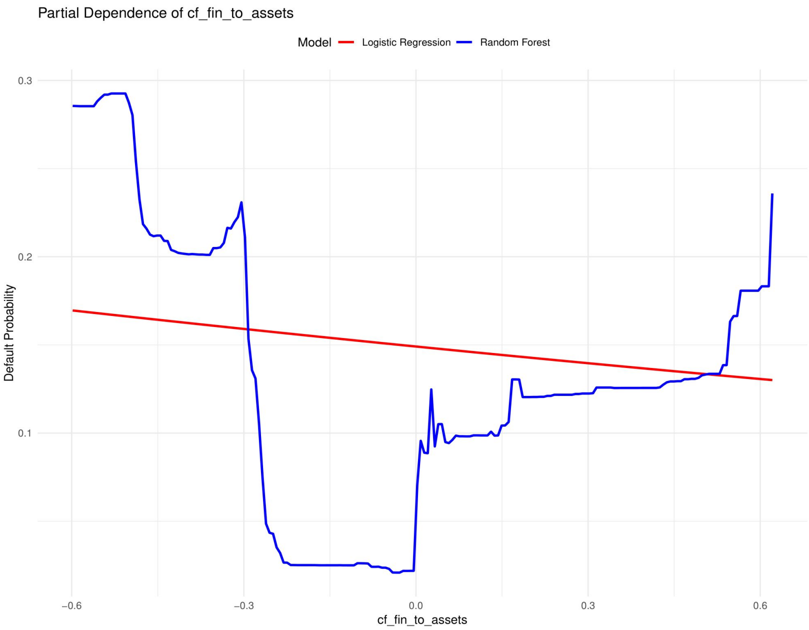


Partial dependence plots. Liquidity ratios

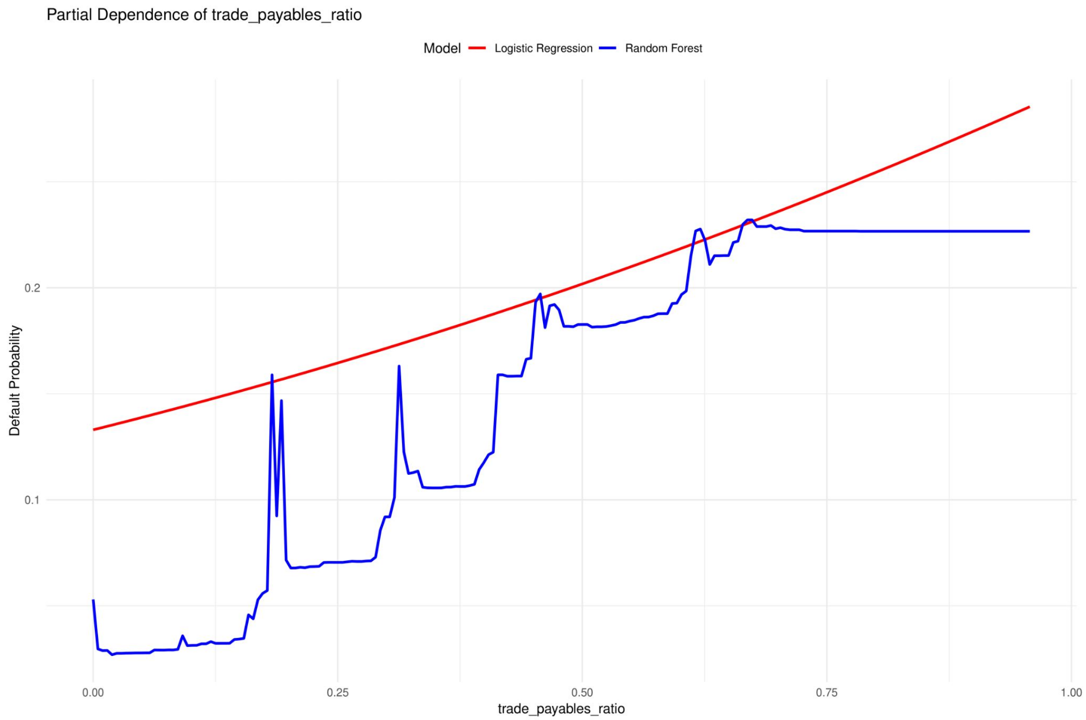


Partial dependence plots. CF ratios





Partial dependence plots. Turnover ratios



Q&A

Session

Correlation between Ratios