Improving Random Forest models for predicting Credit Risk

Supervisor: Univ. Prof. Dipl.-Ing. Dr. techn. Stefan Gerhold

Co-Supervisor: Dr. Arpad Pinter

TECHNISCHE UNIVERSITÄT WIEN

Credit Risk modeling

Assessing borrower creditworthiness is critical for financial stability.
Especially since the World Financial Crisis 2007 – 2008.

Contrasting Goals in Credit Risk modeling

Statistician's Perspective

Model Accuracy
Maximize predictive performance with high-complexity, models.

Flexibility:
Adapt models to dynamic datasets with regular updates to enhance robustness.

Assumptions and Imputation:
Handle missing/imbalanced data through statistical assumptions, interpolation, or imputation.

Regulator's Perspective

Interpretability:
Ensure transparency and traceability of decision-making processes (white-box models).

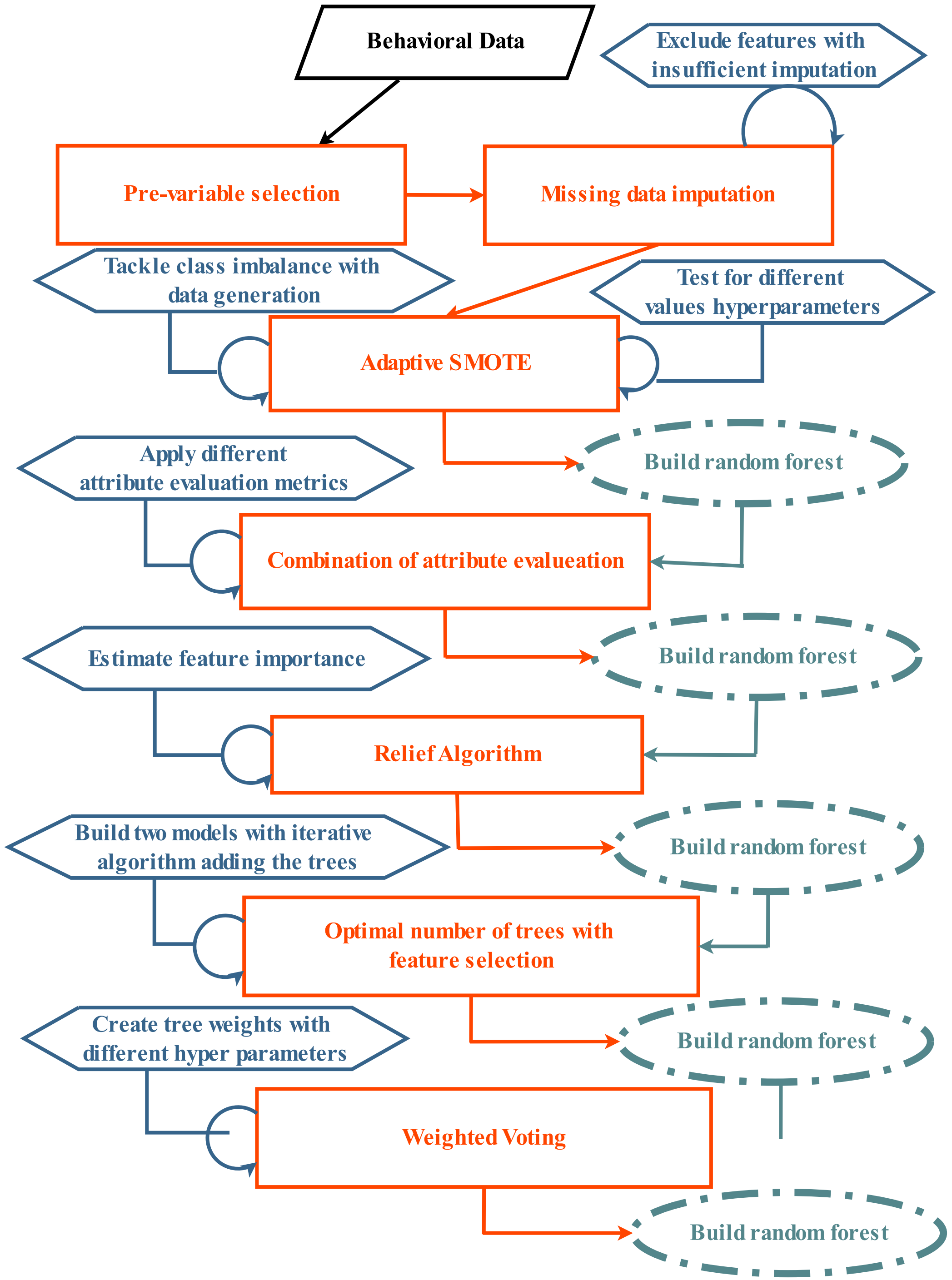
Simplicity:
Maintain stability with minimal updates for auditable and validated models.

Data Accuracy:
Prioritize accurate, minimally manipulated data to ensure compliance and reliability.

Logistic Regression

Why prefer Random Forests?

- Robustness to Overfitting
- Handles High-Dimensional and Imbalanced data
- Feature Importance Evaluation
- Untransformed Data Handling



- Robustness to Overfitting
- Handles High-Dimensional and Imbalanced data
- Feature Importance Evaluation
- Untransformed Data Handling

Attribute Evaluation & Relief Algorithm

Improvement in two ways

1. Prediction Accuracy of Individual Trees
2. Correlation between Trees

Partition data by split variable and split value for high homogeneity.

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

$$Entropy = -\sum_{i=1}^{C} p_i \log_2(p_i)$$

$$MDL\ (D) = min_{H \in \mathcal{H}}\ L(D|H) + L(H)$$

- Accuracy / Specificity, F_{β}-Score, Precision and Recall

Estimate feature importance based on *10*-nearest neighbors.

Measure conditional dependencies among attributes.

$$\omega_j = \left[\frac{1}{m}\sum_{i=1}^{m}\left(\omega_j - \frac{(x_{ij} - Hit_j)^2}{norm} + \frac{(x_{ij} - Miss_j)^2}{norm}\right)\right]^{\alpha},\ j = 1, \dots p,\ \ \alpha > 0,\ m \leq n$$

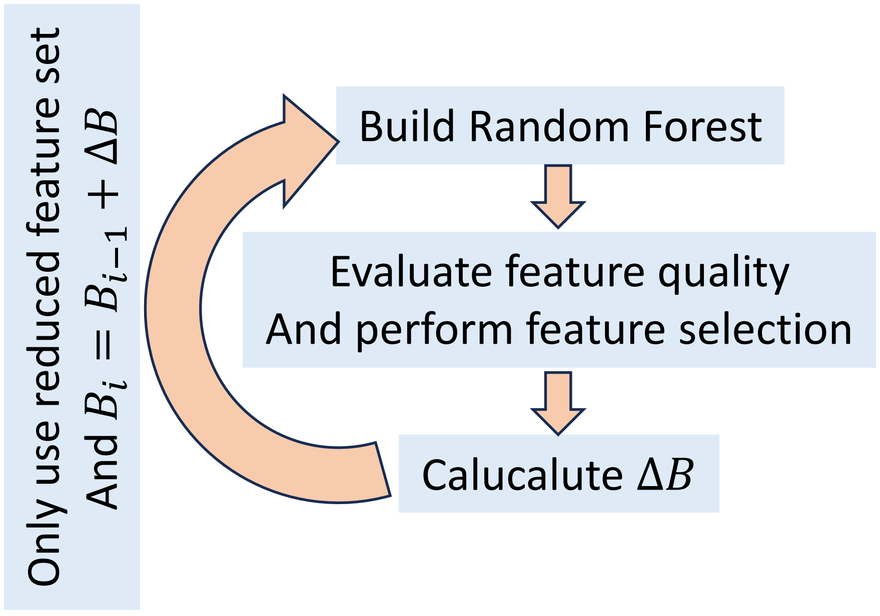
Optimal Number of Trees with feature selection

Too few trees: Model underfits the data.
Too many trees: Increased computation time and model complexity with diminishing returns.

1. Measure feature importance and perform selection.
2. Calculate number of trees for next model
3. Build next Random Forest

Improvement in two ways

1. Prediction Accuracy of Individual Trees
2. Correlation between Trees



$$\eta = \lambda(\eta_s\text{-}\eta_c)$$

$$d\eta = \lambda\left[\frac{\partial(\eta_s - \eta_c)}{\partial B} + \frac{\partial(\eta_s - \eta_c)}{\partial u} + \frac{\partial(\eta_s - \eta_c)}{\partial v}\right] > 0$$

$\eta_s(B, u, v)$ … strength of the trees
$\eta_c(B, u, v)$ … intra – tree correlation
$\lambda \in \mathbb{R}_+$ … scaling factor

Implementation

The code basis is provided by Breiman and Cutler's Random Forests for Classification and Regression package from CRAN.

The R-package utilizes the following programming languages:

R, C, Fortran77

All discussed methods are fully integrated in the R-package with focus on run time efficient implementation.

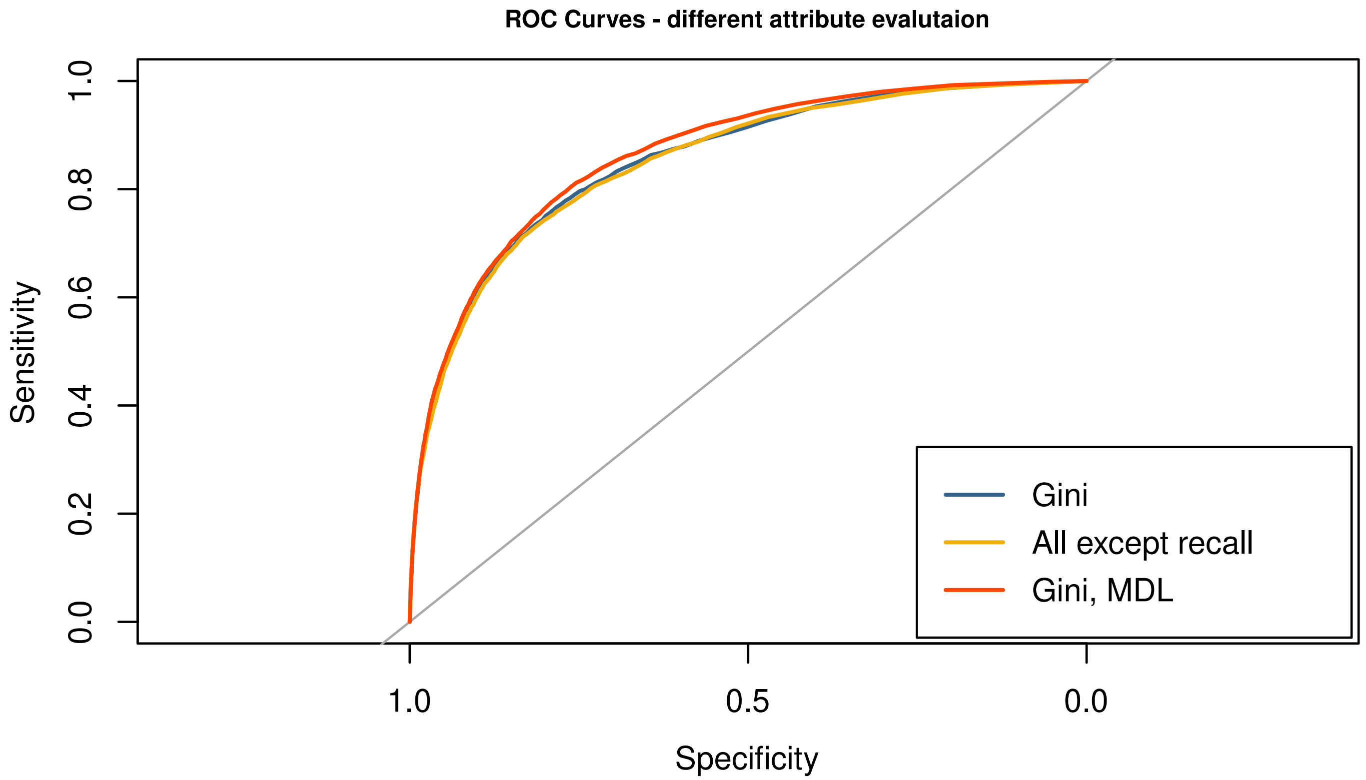
Result I

Improvement of

~13% compared to Gini

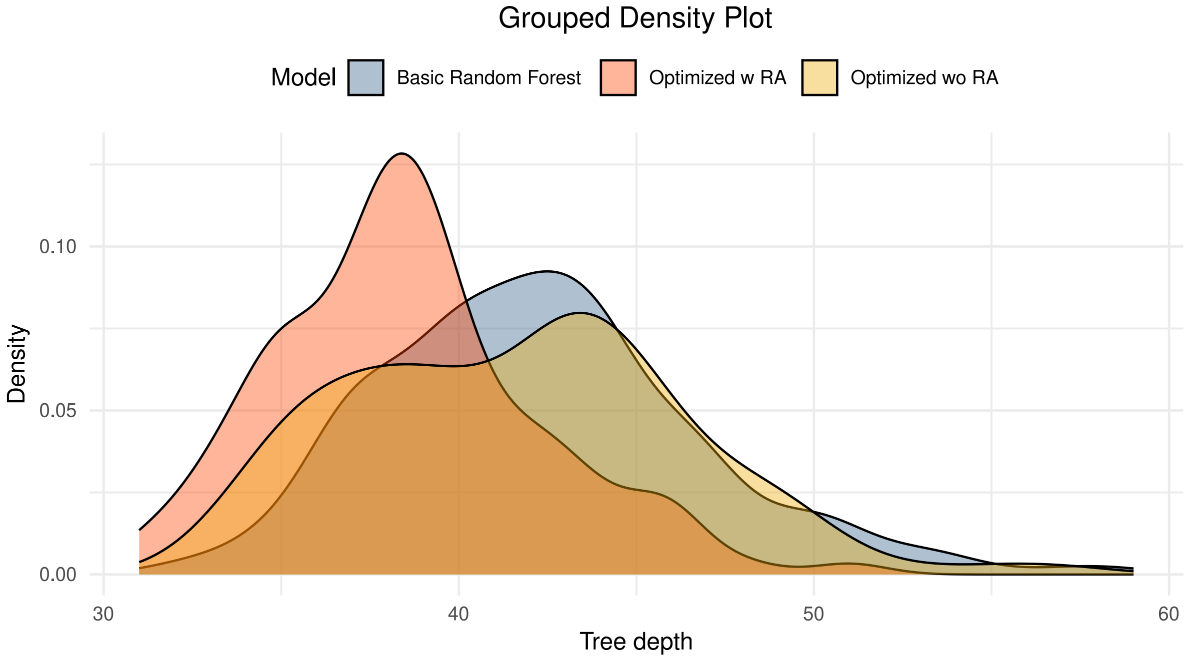
~ 6% compared to MDL

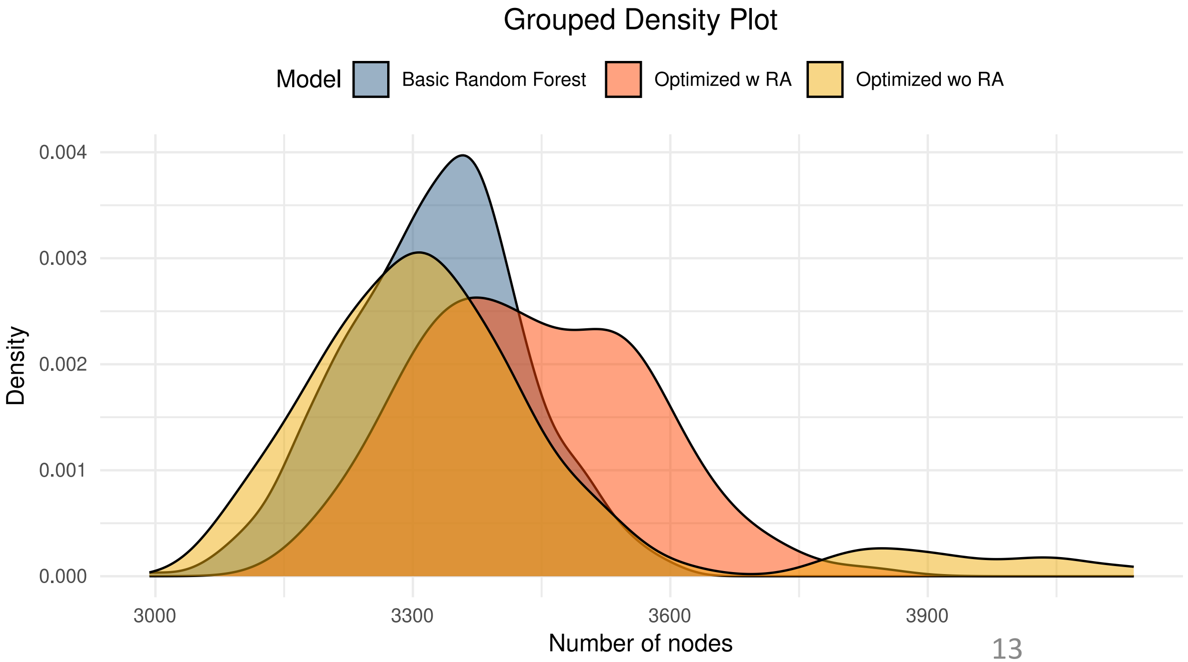
Metrics Used	AUC
Gini	0.8673
All except recall	0.8650
Gini & MDL	**0.8790**



Result II

Model	# of Trees	Avg. Tree depth	# of features	Avg. # of nodes
Basic Random Forest	500	42	110	3325
Optimized with reconstruct all	**141**	**38**	84	**3438**
Optimized with construct new trees	147	41	110	3353





Outlook

Promising results towards **enhanced performance** and **predictive accuracy** while balancing regulatory requirements, **interpretability** and **feasibility**.

Limitations

Missing at Random – MaR assumption in Data Imputation

Computational Complexity with respect to frequent model updates.

Model Improvement Methods only implemented for a two-class problem.

Key findings

Combined Attribute Evaluation and weighted voting improved prediction accuracy.

Optimal Number of Trees with feature selection allows computational efficiency and lower model complexity.

- Attribute Evaluation
- Hyperparameter optimization
- Additional Metrics
- Considering cross-dependencies and interactions between parameters

Appendix I

	$\hat{y} = \mathbf{0}$	$\hat{y} = \mathbf{1}$
$\boldsymbol{y} = \mathbf{0}$	TN	FP
$\boldsymbol{y} = \mathbf{1}$	FN	TP

$$Accuracy = \frac{TN + TP}{TN + FP + FN + TP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specifity = \frac{TN}{TN + FP}$$

$$F_\beta - score = \frac{(1+\beta)^2 TP}{(1+\beta)^2 TP + \beta^2 FN + FP}$$

$$MDL_i(j) = \frac{1}{n}\left[\log_2\binom{n}{n_1, n_2, \ldots n_c} - \log_2\binom{n_{.<j}}{n_{1<j}, n_{2<j}, \ldots n_{c<j}} - \log_2\binom{n_{.\geq j}}{n_{1\geq j}, n_{2\geq j}, \ldots n_{c\geq j}} + \log_2\binom{n+C-1}{C-1} - \log_2\binom{n_{.<j}+C-1}{C-1} - \log_2\binom{n_{.\geq j}+C-1}{C-1}\right]$$

n is the total number of data points.
$n_1, n_2, \ldots n_c$ are the number of data points in each class.
$n_{.<j}$ is the number of data points less than j.
$n_{.\geq j}$ is the number of data points greater than or equal j.
$n_{i\geq j,}$ $n_{i<j}$ are the counts of data points in each class i less than resp. greater than or equal to j.

Appendix II

Feature Importance:

$Entropy\ E_l(i,j), E_r(i,j)\ of\ node\ i, split\ j$

$$Q(i,j) = e^{-E_l(i,j)-E_r(i,j)}$$

$$\omega^\tau(j) = \frac{\sum_{i=1}^{N} Q(i,j)}{N}, \qquad j = 1,\dots,p$$

$$\gamma_\tau = \frac{\frac{1}{\delta_\tau}}{\max_\tau(\frac{1}{\delta_\tau})}, \qquad \tau = 1,\dots,B$$

$$\omega(j) = \frac{\sum_{\tau=1}^{B} \omega^\tau(j)\gamma_\tau}{\max_j \sum_{\tau=1}^{B} \omega^\tau(j)\gamma_\tau}, j = 1,\dots p$$

Calculating Optimal Number of Trees

$u, v\ \dots number\ of\ (un)-important\ features$

$\Delta u, \Delta v\ \dots change\ in\ u, v\ per\ iteration\ step$

$f\ \dots number\ of\ features\ selected\ for\ tree\ building$

$B\ \dots number\ of\ trees\ in\ Random\ Forest$

$N_{av}\ \dots avarage\ number\ of\ nodes\ in\ each\ tree$

$$q = 1 - \frac{\binom{v}{f}}{\binom{u+v}{f}}$$

$$q_u \approx \left(\frac{\Delta q}{\Delta u}\right)_v = \frac{v!\,(u+v-1-f)!\,f}{(v-f)!\,(u+v)!}$$

$$q_v \approx \left(\frac{\Delta q}{\Delta v}\right)_u = \frac{(v-1)!\,(u+v-1-f)!\,fu}{(v-f)!\,(u+v-1)!\,(u+v)}$$

$$\rho = \left(1 - \frac{\binom{u+v-f}{f}}{\binom{u+v}{f}}\right)^{N_{av}}$$

$$\eta_c(B,u,v) = 1 - (1-\rho)^{\frac{B}{2}}$$

$$\eta_s(B,u,v) = 1 - (1-q^{N_{av}})^B$$

$$|\Delta B| < \left|\frac{BN_{av}q^{N_{av}-1}(1-q^{N_{av}})^{B-1}(q_u\Delta u + q_v\Delta v)}{\frac{\partial(\eta_s-\eta_c)}{\partial B}}\right|$$