FH MÜNSTER
University of Applied Sciences

MSB FB Wirtschaft
Münster School of Business

Transparency of Machine Learning Models in Credit Scoring

CRC Conference XVI

Michael Bücker, Gero Szepannek, Przemyslaw Biecek, Alicja Gosiewska and Mateusz Staniak

28 August 2019


SENORGIF.COM

Introduction

Introduction

Michael Bücker

Professor of Data Science at Münster School of Business

2003-2011

technische universität dortmund

2011-2018

McKinsey & Company

2018-today

FH MÜNSTER University of Applied Sciences

SCALEWORK MAKE SCALE WORK

hawk:AI

Introduction

- Main requirement for Credit Scoring models: provide a risk prediction that is **as accurate as possible**
- In addition, regulators demand these models to be **transparent and auditable**
- Therefore, very **simple predictive models** such as Logistic Regression or Decision Trees are still widely used (Lessmann, Baesens, Seow, and Thomas 2015; Bischl, Kühn, and Szepannek 2014)
- Superior predictive power of modern **Machine Learning algorithms** cannot be fully leveraged
- A lot of **potential is missed**, leading to higher reserves or more credit defaults (Szepannek 2017)

Research Approach

- For an open data set we build a traditional and still state-of-the-art Score Card model
- In addition, we build alternative Machine Learning Black Box models
- We use model-agnostic methods for interpretable Machine Learning to showcase transparency of such models
- For computations we use R and respective packages (Biecek 2018; Molnar, Bischl, and Casalicchio 2018)

The incumbent: Score Cards

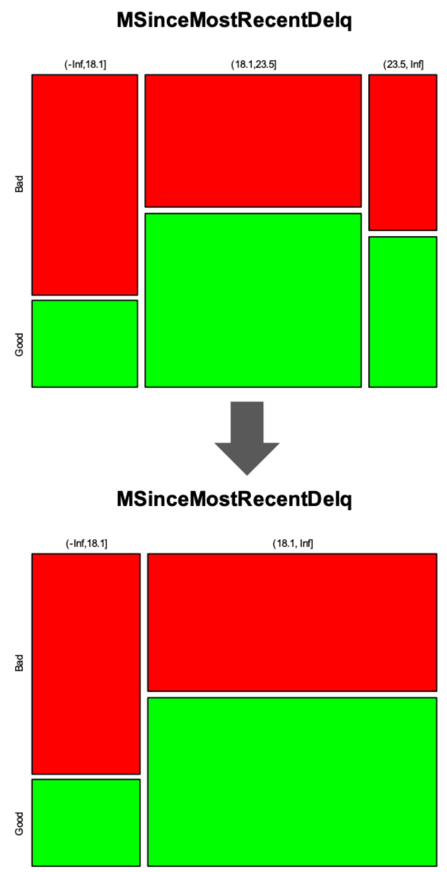
Steps for Score Card construction using Logistic Regression (Szepannek 2017)

1. Automatic binning
2. Manual binning
3. WOE/Dummy transformation
4. Variable shortlist selection
5. (Linear) modelling and automatic model selection
6. Manual model selection

The incumbent: Score Cards

Steps for Score Card construction using Logistic Regression (Szepannek 2017)

1. Automatic binning
2. Manual binning
3. WOE/Dummy transformation
4. Variable shortlist selection
5. (Linear) modelling and automatic model selection
6. Manual model selection



Score Cards: Manual binning

Manual binning allows for

- (univariate) non-linearity
- (univariate) plausibility checks
- integration of expert knowledge for binning of factors

...but: only univariate effects (!)

... and means a lot of manual work

The challenger models

We tested a couple of Machine Learning algorithms ...

- Random Forests (`randomForest`)
- Gradient Boosting (`gbm`)
- XGBoost (`xgboost`)
- Support Vector Machines (`svm`)
- Logistic Regression with spline based transformations (`rms`)

... and also two AutoML frameworks to beat the Score Card

- h2o AutoML (`h2o`)
- mljar.com (`mljar`)

Data set for study: xML Challenge by FICO

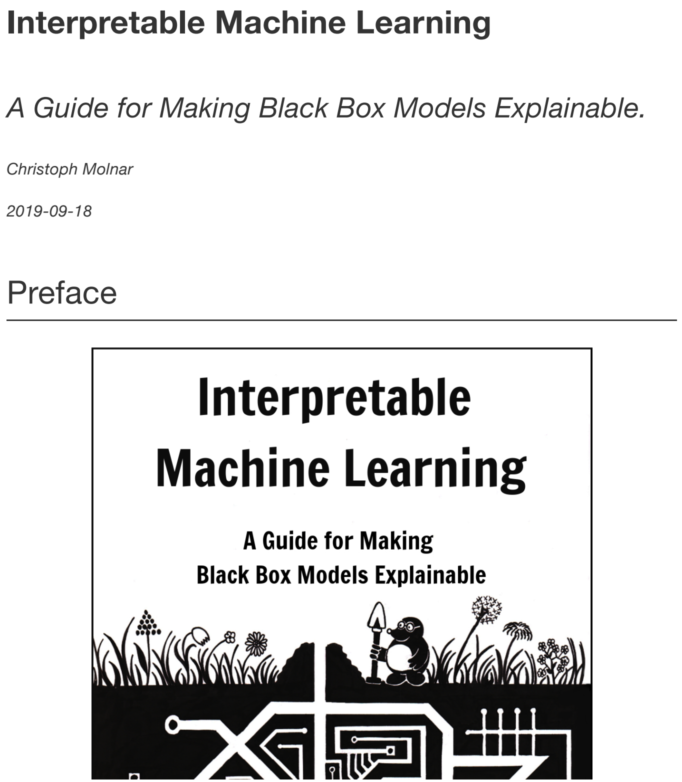
- Explainable Machine Learning Challenge by FICO (2019)
- Focus: Home Equity Line of Credit (HELOC) Dataset
- Customers requested a credit line in the range of $5,000 - $150,000
- Task is to predict whether they will repay their HELOC account within 2 years
- Number of observations: 2,615
- Variables: 23 covariates (mostly numeric) and 1 target variable (risk performance "good" or "bad")

FICO®

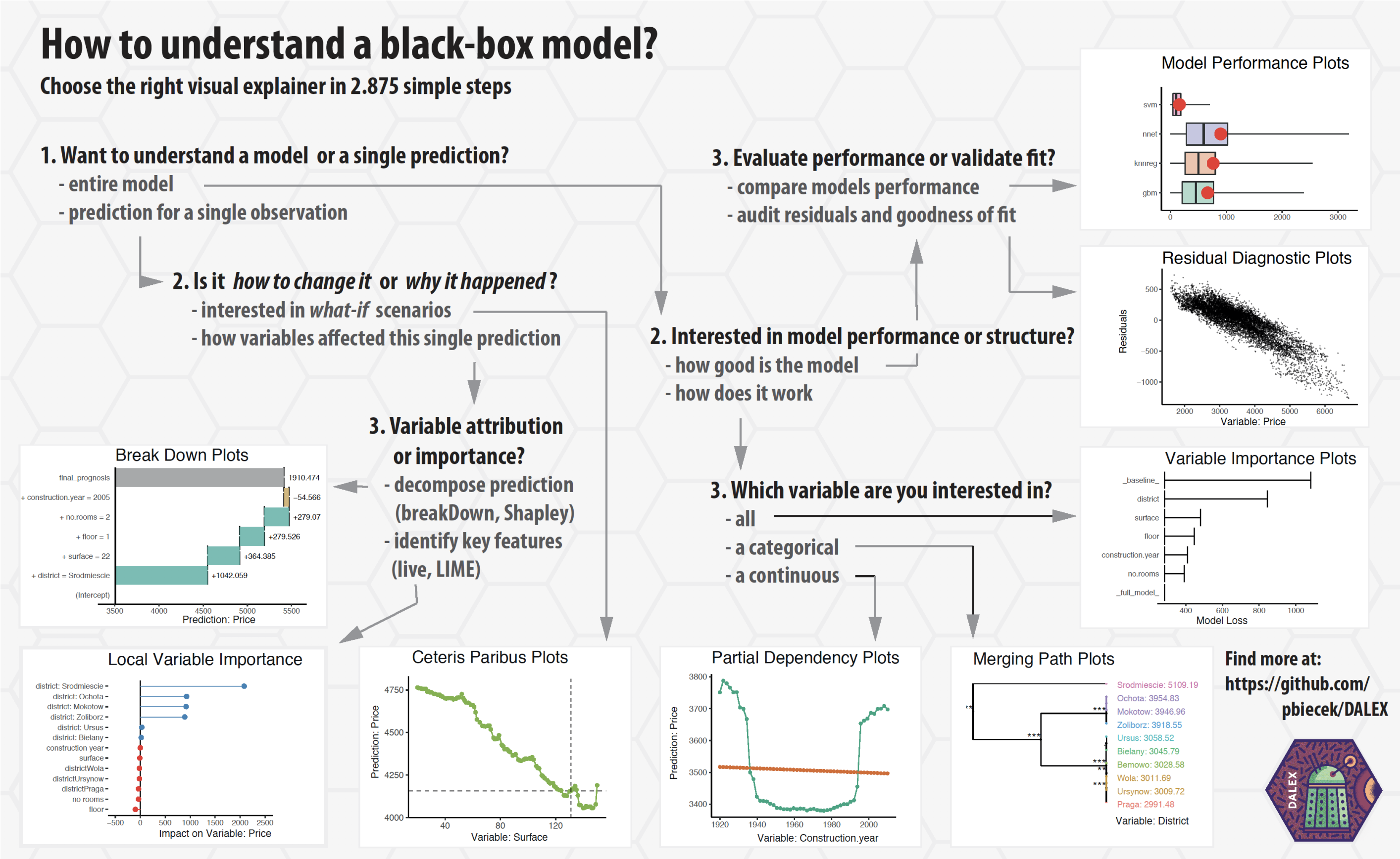
Explainability of Machine Learning models

There are many model-agnostic methods for interpretable ML today; see Molnar (2019) for a good overview.

- Partial Dependence Plots (PDP)
- Individual Conditional Expectation (ICE)
- Accumulated Local Effects (ALE)
- Feature Importance
- Global Surrogate and Local Surrogate (LIME)
- Shapley Values, SHAP
- ...



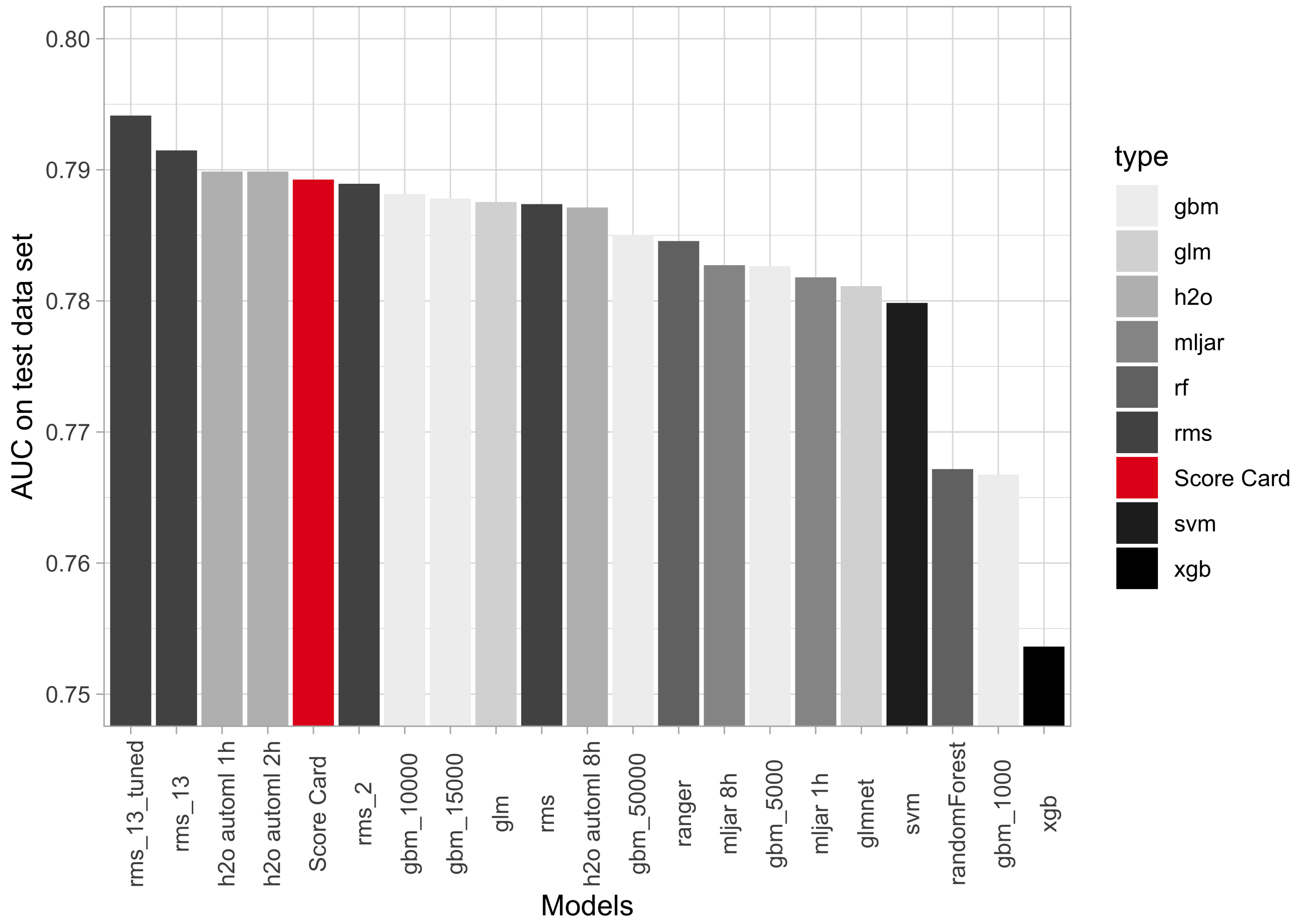
Implementation in R: DALEX



- Descriptive mAchine Learning EXplanations
- DALEX is a set of tools that help to understand how complex models are working

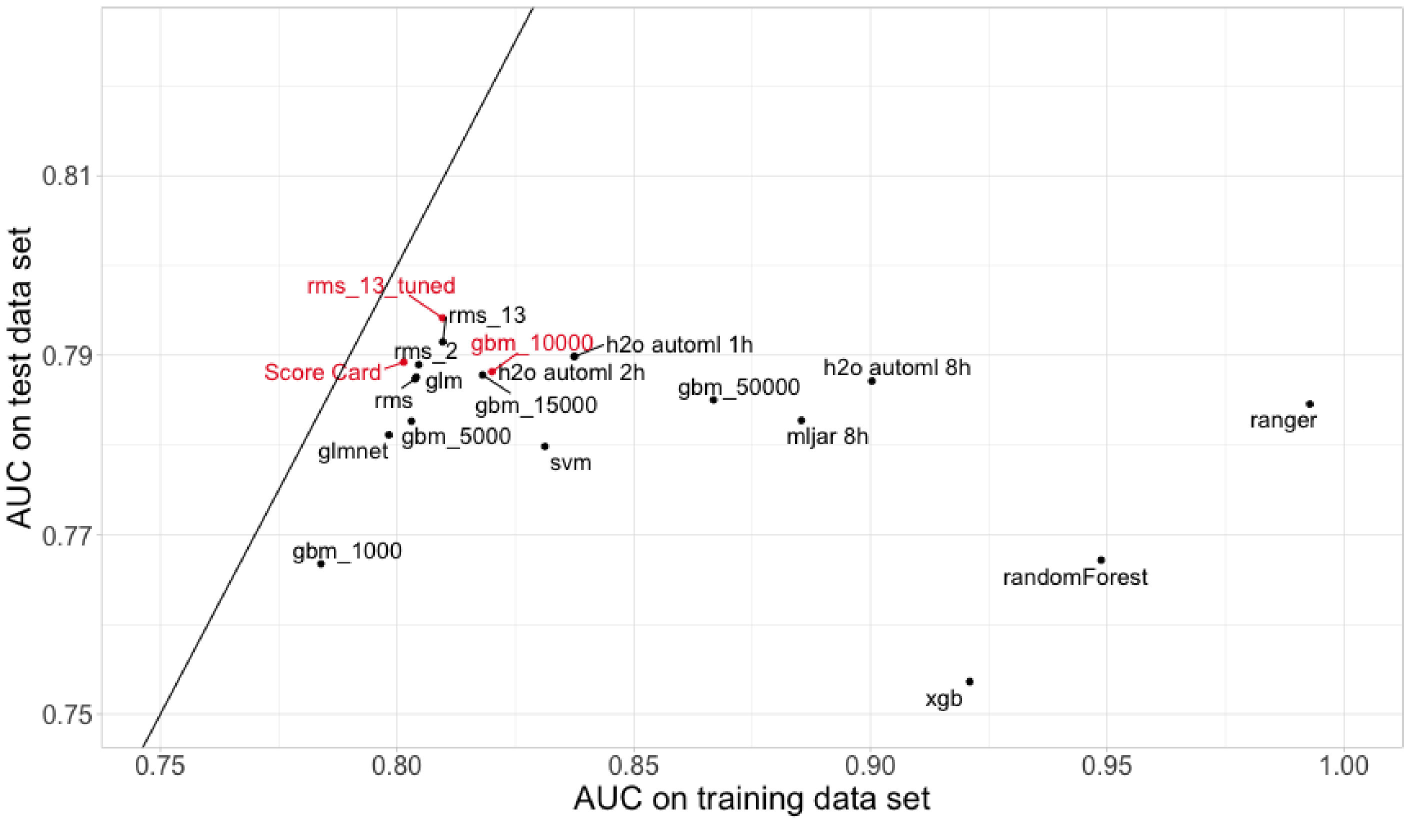
Results: Model performance

Results: Comparison of model performance



- Predictive power of the traditional Score Card model surprisingly good
- Logistic Regression with spline based transformations best, using `rms` by Harrell Jr (2019)

Results: Comparison of model performance

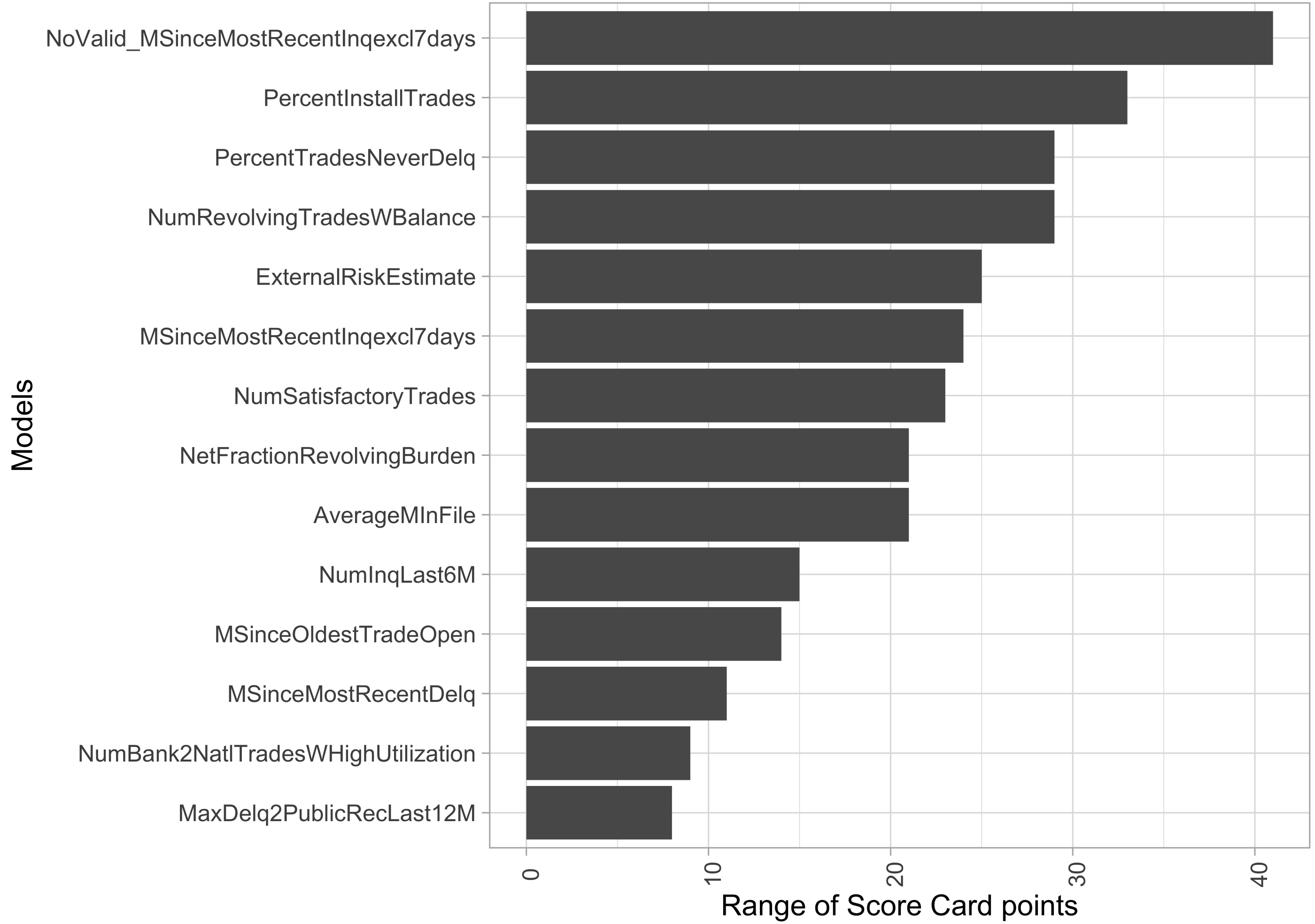


For comparison of explainability, we choose

- the Score Card,
- a Gradient Boosting model with 10,000 trees,
- a tuned Logistic Regression with splines using 13 variables

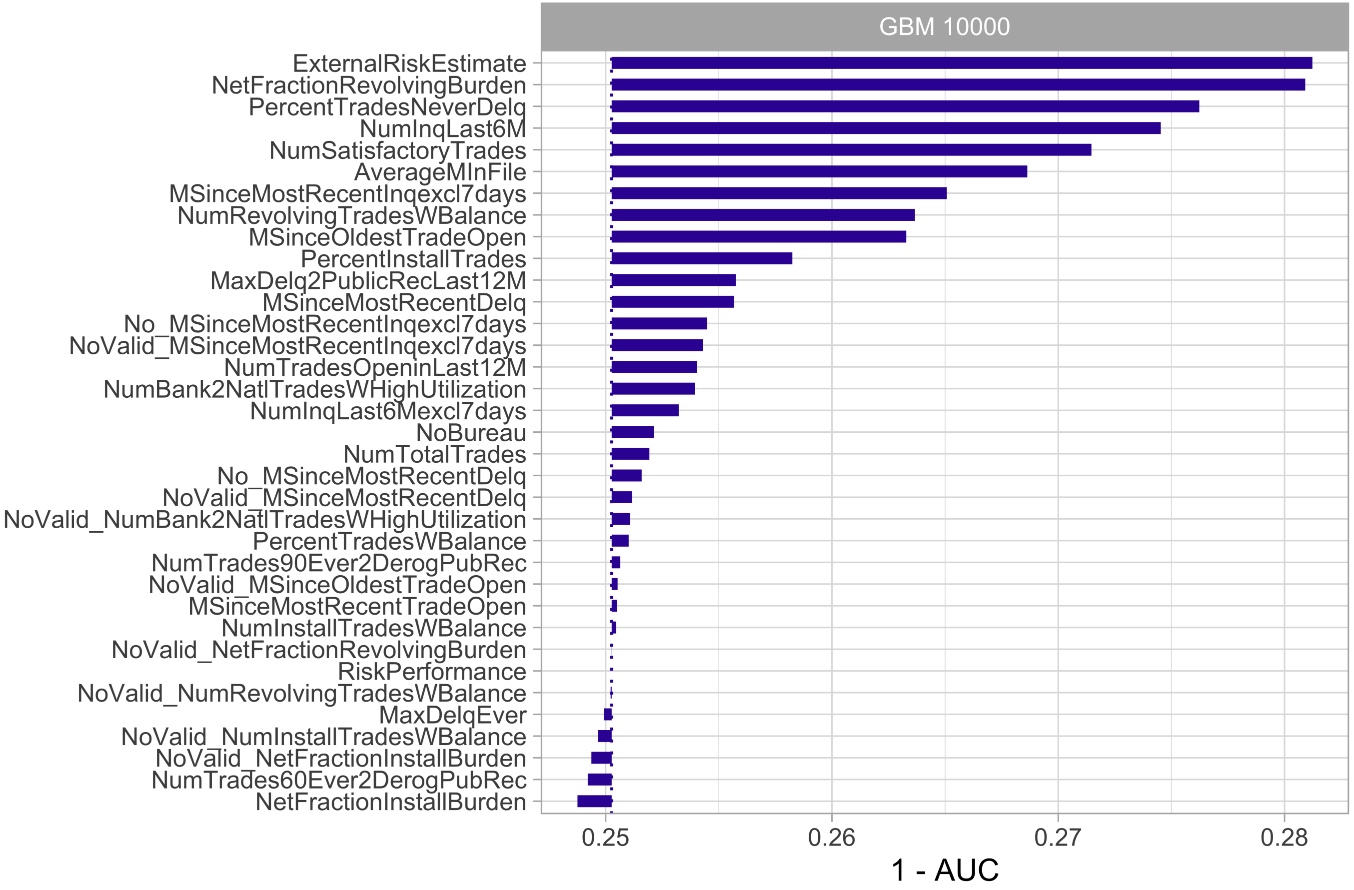
Results: Global explanations

Score Card: Variable importance as range of points



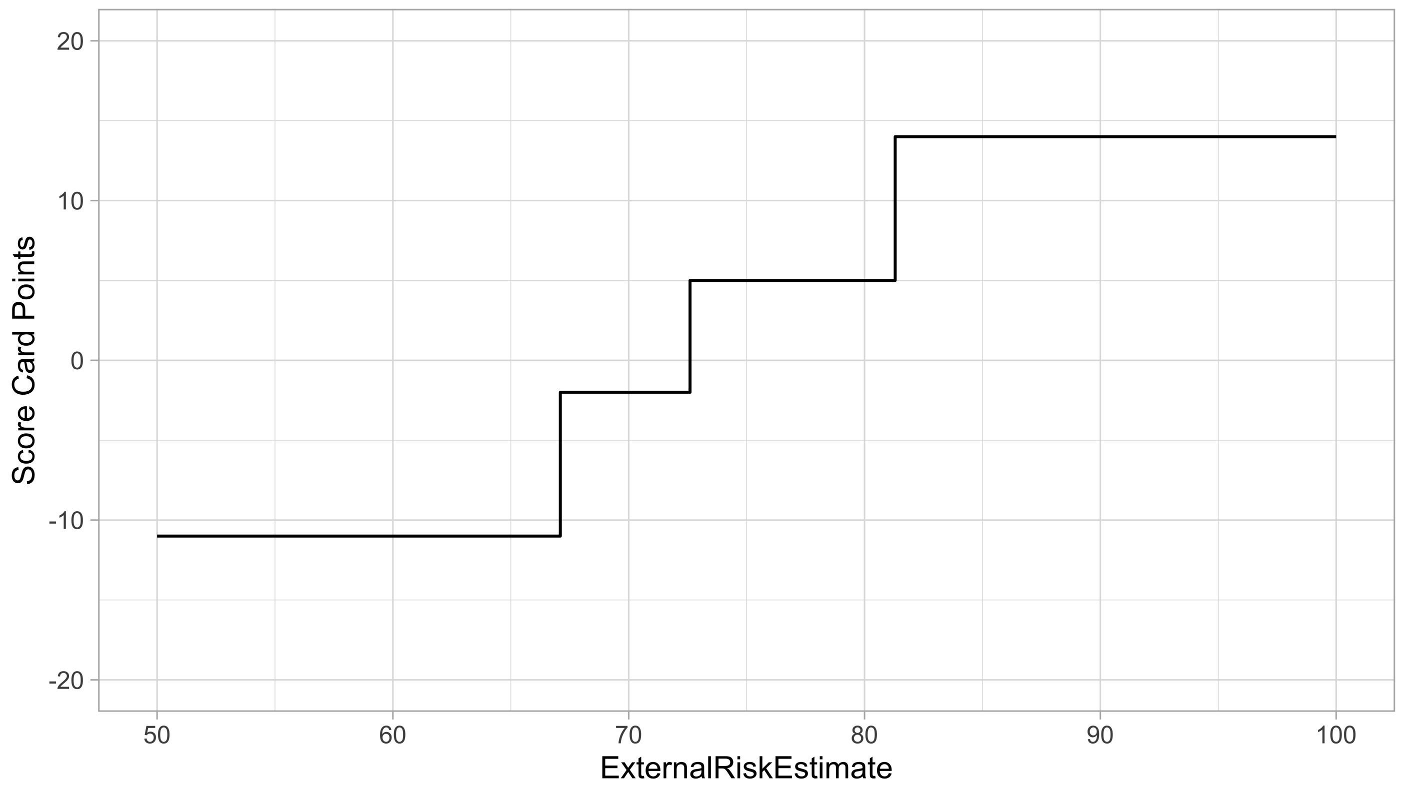
- Range of Score Card point as an indicator of relevance for predictions
- Alternative: variance of Score Card points across applications

Model agnostic: Importance through drop-out loss



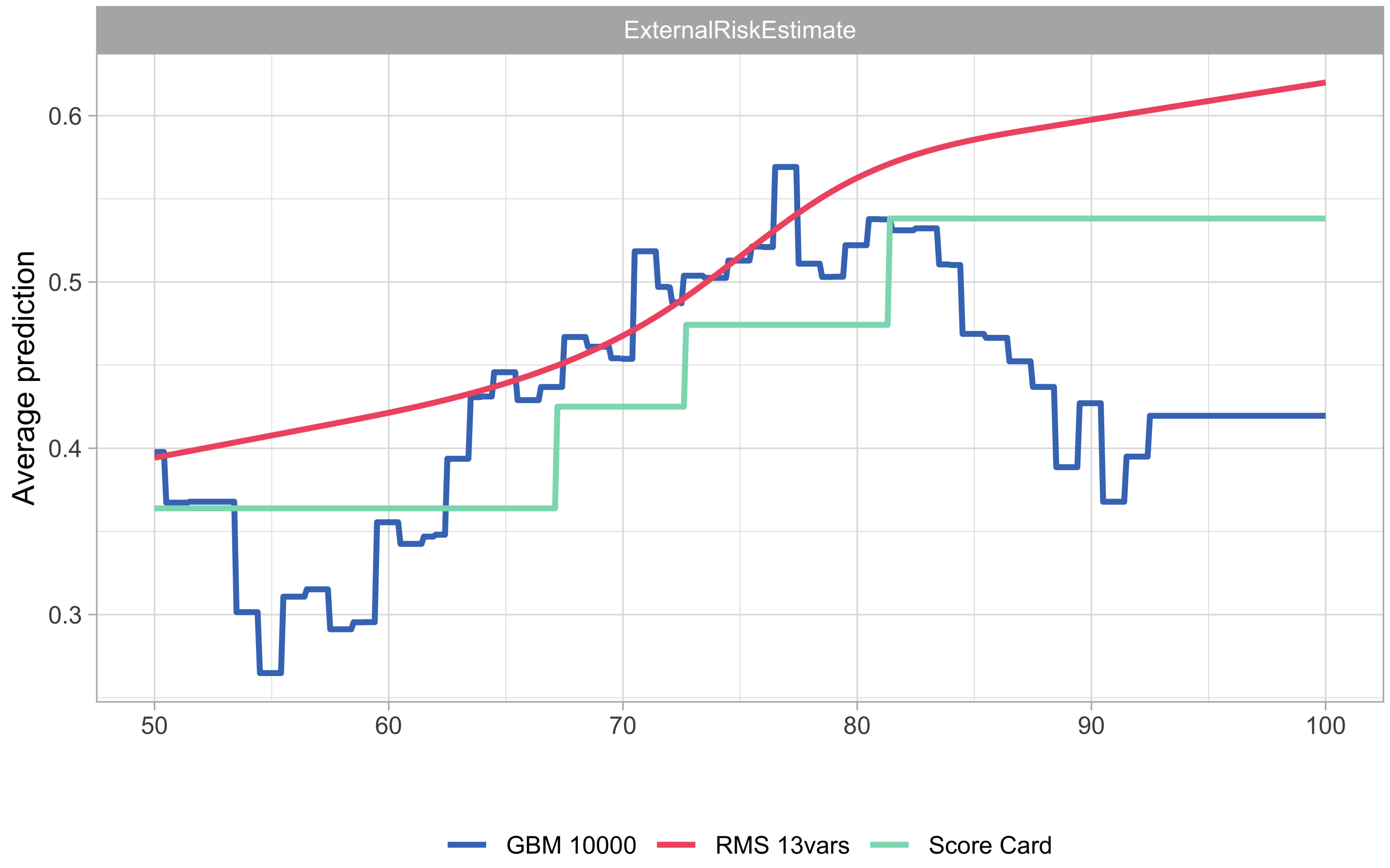
- The drop in model performance (here AUC) is measured after permutation of a single variable
- The more siginficant the drop in performance, the more important the variable

Score Card: Variable explanation based on points



- Score Card points for values of covariate show effect of single feature
- Directly computed from coefficient estimates of the Logistic Regression

Model agnostic: Partial dependence plots



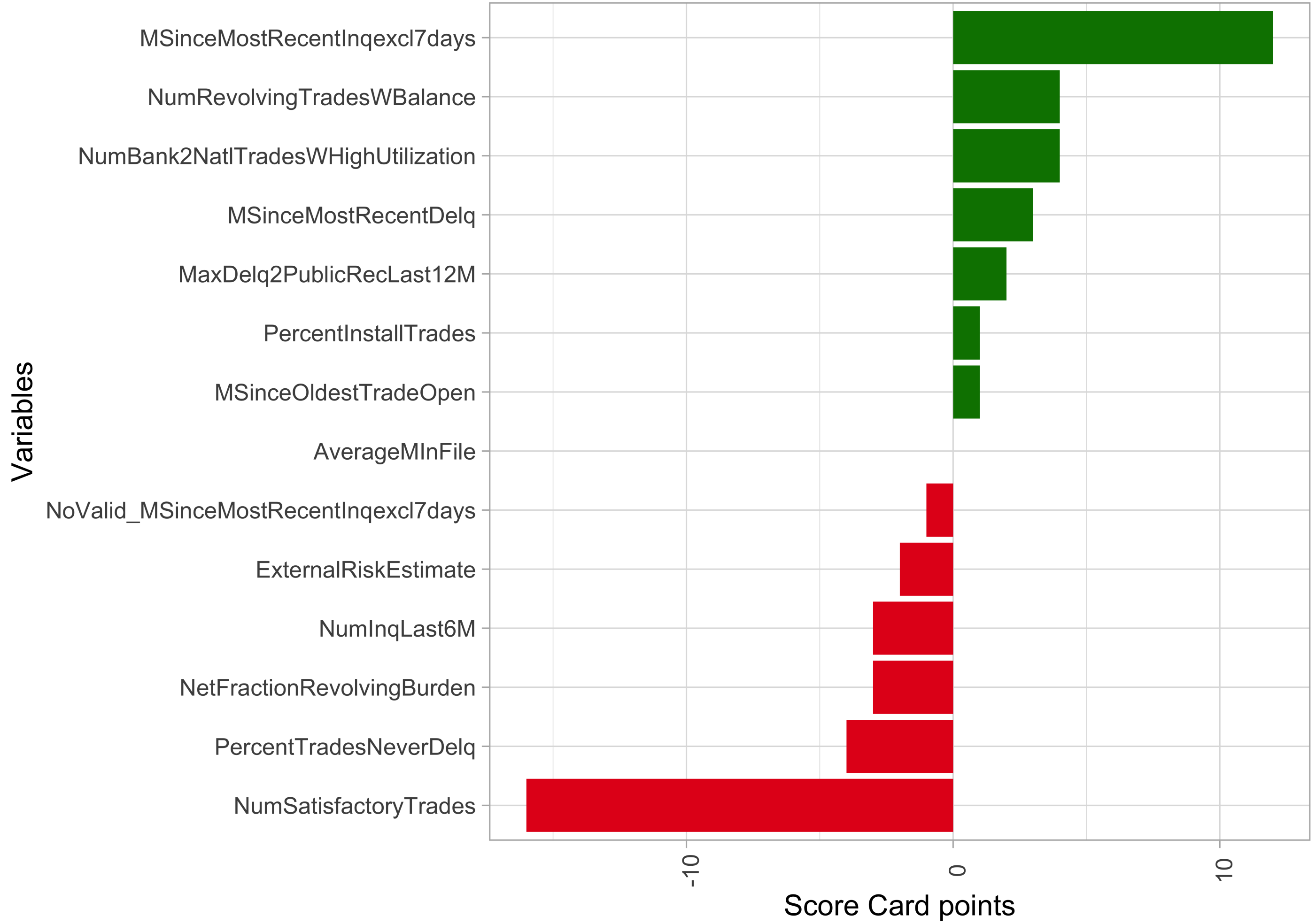
- Partial dependence plots created with (Biecek 2018)
- Interpretation very similar to marginal Score Card points

Results: Local explanations

Instance-level explanations

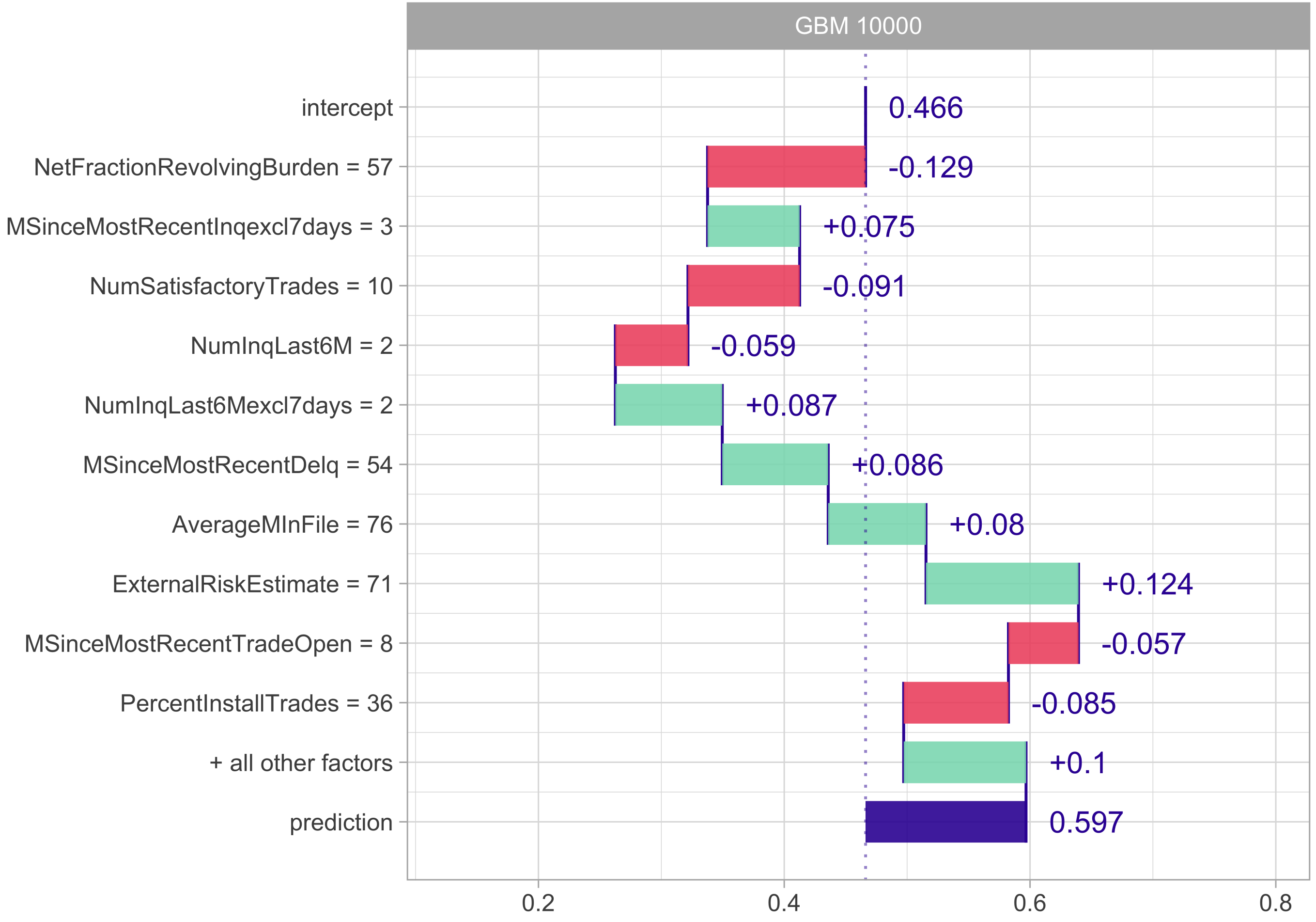
- Instance-level exploration helps to understand how a model yields a prediction for a single observation
- Model-agnostic approaches are
  - additive Breakdowns
  - Shapley Values, SHAP
  - LIME
- In Credit Scoring, this explanation makes each credit decision transparent

Score Card: Local explanations



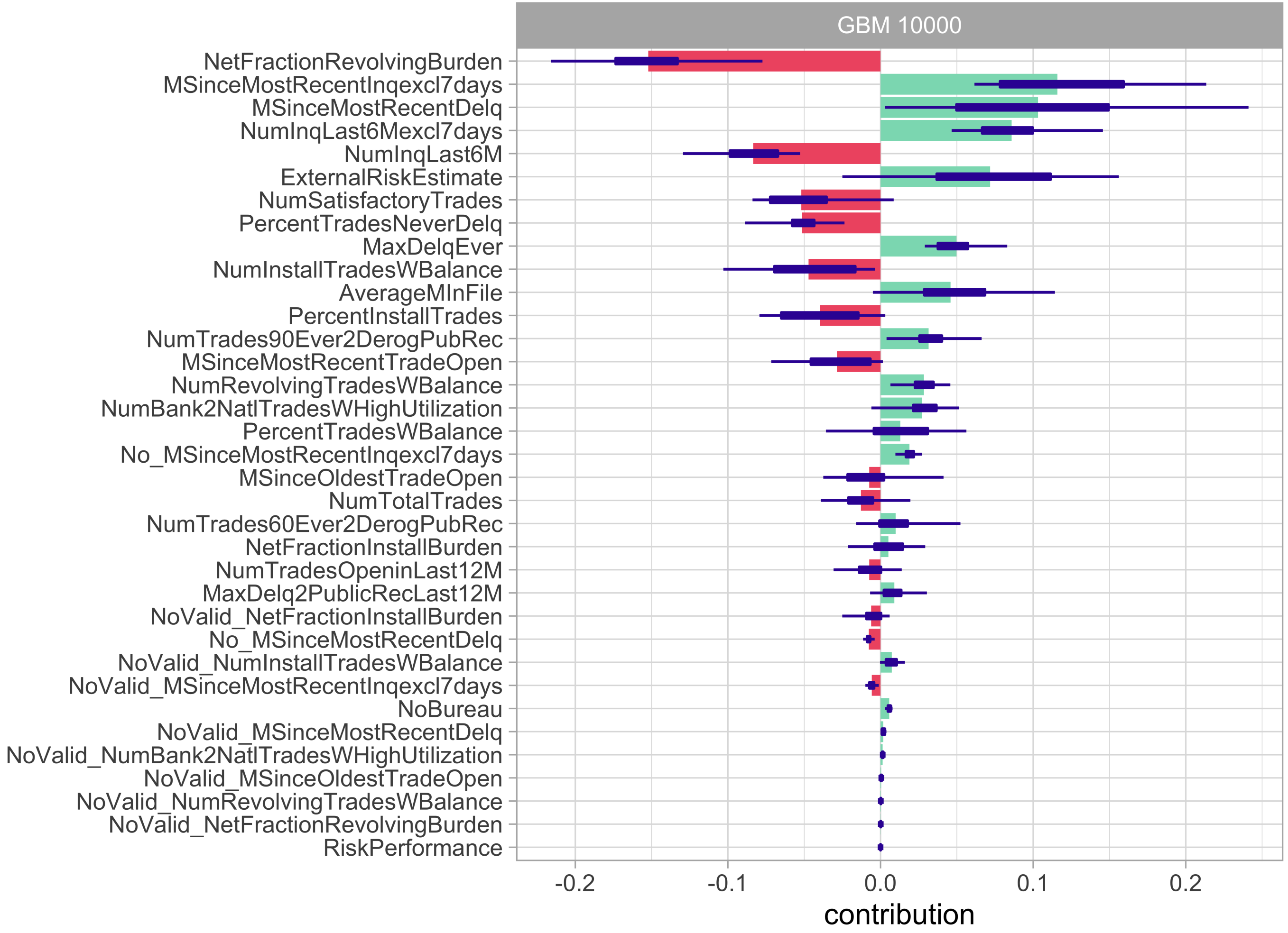
- Instance-level exploration for Score Cards can simply use individual Score Card points
- This yields a breakdown of the scoring result by variable

Model agnostic: Variable contribution break down



- Such instance-level explorations can also be performed in a model-agnostic way
- Unfortunately, for non-additive models, variable contributions depend on the ordering of variables

Model agnostic: SHAP



- Shapley attributions are averages across all (or at least large number) of different orderings
- Violet boxplots show distributions for attributions for a selected variable, while length of the bar stands for an average attribution

Conclusion

Modeldown: HTML summaries for predictive Models

Rf. Biecek, Tatarynowicz, Romaszko, and Urbański (2019)

modelDown

Explore your model!

Basic data information

- 2615 observations
- 35 columns

Explainers

- RMS 13vars (download) (explainers/RMS 13vars.rda)
- GBM 10000 (download) (explainers/GBM 10000.rda)
- Score Card (download) (explainers/Score Card.rda)

Summaries for numerical variables

	vars	n	mean	sd	median	trimmed	mad	min	max	range

Scan me

Conclusion

- We have built models for Credit Scoring using Score Cards and Machine Learning
- Predictive power of Machine Learning models was superior (in our example only slightly, other studies show clearer overperformance)
- Model agnostic methods for interpretable Machine Learning are able to meet the degree of explainability of Score Cards and may even exceed it

References (1/3)

Biecek, P. (2018). "DALEX: explainers for complex predictive models". In: *Journal of Machine Learning Research* 19.84, pp. 1-5.

Biecek, P, M. Tatarynowicz, K. Romaszko, and M. Urbański (2019). *modelDown: Make Static HTML Website for Predictive Models*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=modelDown.

Bischl, B., T. Kühn, and G. Szepannek (2014). "On Class Imbalance Correction for Classification Algorithms in Credit Scoring". In: *Operations Research Proceedings*. Ed. by M. Löbbecke, A. Koster, L. P., M. R., P. B. and G. Walther. , pp. 37-43.

FICO (2019). *xML Challenge*. Online. URL: https://community.fico.com/s/explainable-machine-learning-challenge.

References (2/3)

Harrell Jr, F. E. (2019). *rms: Regression Modeling Strategies*. R package version 5.1-3.1. URL: https://CRAN.R-project.org/package=rms.

Lessmann, S, B. Baesens, H. Seow, and L. Thomas (2015). "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research". In: *European Journal of Operational Research* 247.1, pp. 124-136.

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. URL: https://christophm.github.io/interpretable-ml-book/.

Molnar, C, B. Bischl, and G. Casalicchio (2018). "iml: An R package for Interpretable Machine Learning". In: *Journal Of Statistical Software* 3.26, p. 786. URL: http://joss.theoj.org/papers/10.21105/joss.00786.

References (3/3)

Szepannek, G. (2017b). *A Framework for Scorecard Modelling using R*. CSCC 2017.

Szepannek, G. (2017a). "On the Practical Relevance of Modern Machine Learning Algorithms for Credit Scoring Applications". In: *WIAS Report Series* 29, pp. 88-96.

Thank you!

Prof. Dr. Michael Bücker

Professor of Data Science
Münster School of Business

FH Münster - University of Applied Sciences -
Corrensstraße 25, Room C521
D-48149 Münster

Tel: +49 251 83 65615
E-Mail: michael.buecker@fh-muenster.de
http://prof.buecker.ms