

# Interpretable Machine Learning with application to Credit Scoring

Alicja Gosiewska

16.11.2020

● Interpretable Machine Learning  
Wyszukiwane hasło

● Explainable Artificial Intellige...  
Wyszukiwane hasło

+ Dodaj porównanie

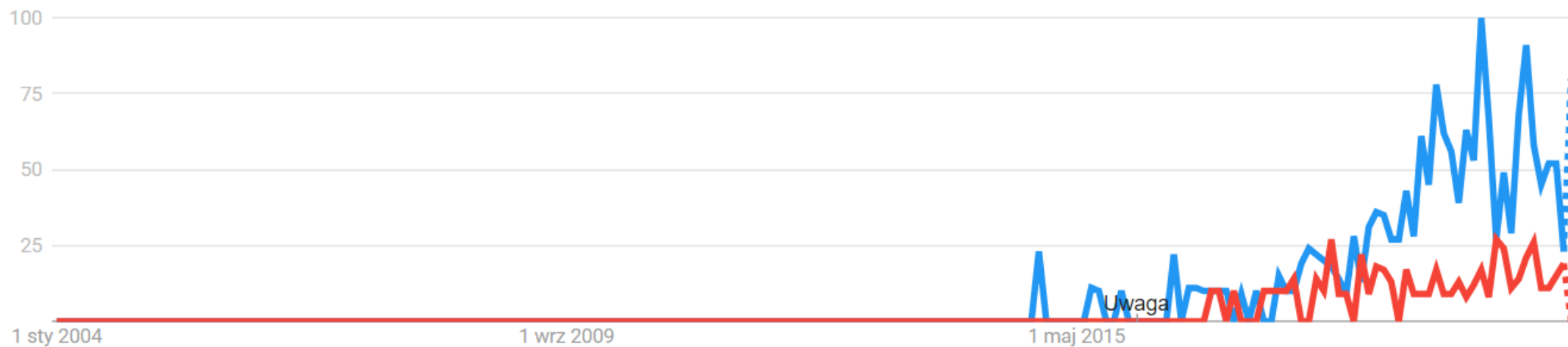
Cały świat ▼

2004 – dziś ▼

Wszystko ▼

Wyszukiwarka Google ▼

Zainteresowanie w ujęciu czasowym ?



Średnia



Bez ograniczenia  
czasowego

Od 2020

Od 2019

Od 2016

Zakres  
niestandardowy...

Wg trafności

Wg daty

Dowolny język

Tylko język polski

☒ uwzględnij patenty

☒ uwzględnij cytaty

☒ Utwórz alert

### [KSIĄŻKA] Interpretable Machine Learning

[C Molnar](#) - 2020 - [books.google.com](#)

This book is about making machine learning models and their decisions interpretable. After exploring the concepts of interpretability, you will learn about simple, interpretable models such as decision trees, decision rules and linear regression. Later chapters focus on general ...

☆ Cytowane przez 277 Powiązane artykuły

### A Survey on **Explainable Artificial Intelligence** (XAI): Toward Medical XAI

[E Tjoa](#), [C Guan](#) - [IEEE Transactions on Neural Networks and ...](#), 2020 - [ieeexplore.ieee.org](#)

Recently, artificial intelligence and machine learning in general have demonstrated remarkable performances in many tasks, from image processing to natural language processing, especially with the advent of deep learning (DL). Along with research progress ...

☆ Cytowane przez 48 Powiązane artykuły Wszystkie wersje 3

### [HTML] **Explainable Artificial Intelligence** (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI

[AB Arrieta](#), [N Díaz-Rodríguez](#), [J Del Ser](#), [A Bennetot](#)... - [Information ...](#), 2020 - Elsevier

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in ...

☆ Cytowane przez 222 Powiązane artykuły Wszystkie wersje 15

[PDF] [ieee.org](#)

[HTML] [sciencedirect.com](#)

Bez ograniczenia  
czasowego

Od 2020

Od 2019

Od 2016

Zakres  
niestandardowy...

Wg trafności

Wg daty

Dowolny język

Tylko język polski

☒ uwzględnij patenty


☒ uwzględnij cytaty

☒ Utwórz alert

## [KSIĄŻKA] Interpretable Machine Learning

[C Molnar](#) - 2020 - [books.google.com](#)

This book is about making machine learning models and their decisions interpretable. After exploring the concepts of interpretability, you will learn about simple, interpretable models such as decision trees, decision rules and linear regression. Later chapters focus on general ...


☆  Cytowane przez 277 Powiązane artykuły

## A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI

[PDF] [ieee.org](#)

[E Tjoa](#), [C Guan](#) - IEEE Transactions on Neural Networks and ..., 2020 - [ieeexplore.ieee.org](#)

Recently, artificial intelligence and machine learning in general have demonstrated remarkable performances in many tasks, from image processing to natural language processing, especially with the advent of deep learning (DL). Along with research progress ...


☆  Cytowane przez 48 Powiązane artykuły Wszystkie wersje 3

## [HTML] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI

[HTML] [sciencedirect.com](#)

[AB Arrieta](#), [N Díaz-Rodríguez](#), [J Del Ser](#), [A Bennetot](#)... - Information ..., 2020 - Elsevier

In the last few years, Artificial Intelligence (AI) has achieved a notable momentum that, if harnessed appropriately, may deliver the best of expectations over many application sectors across the field. For this to occur shortly in Machine Learning, the entire community stands in ...

☆  Cytowane przez 222 Powiązane artykuły Wszystkie wersje 15

Citation report for 1 640 results from Web of Science Core Collection between 1900 and 2021 Go

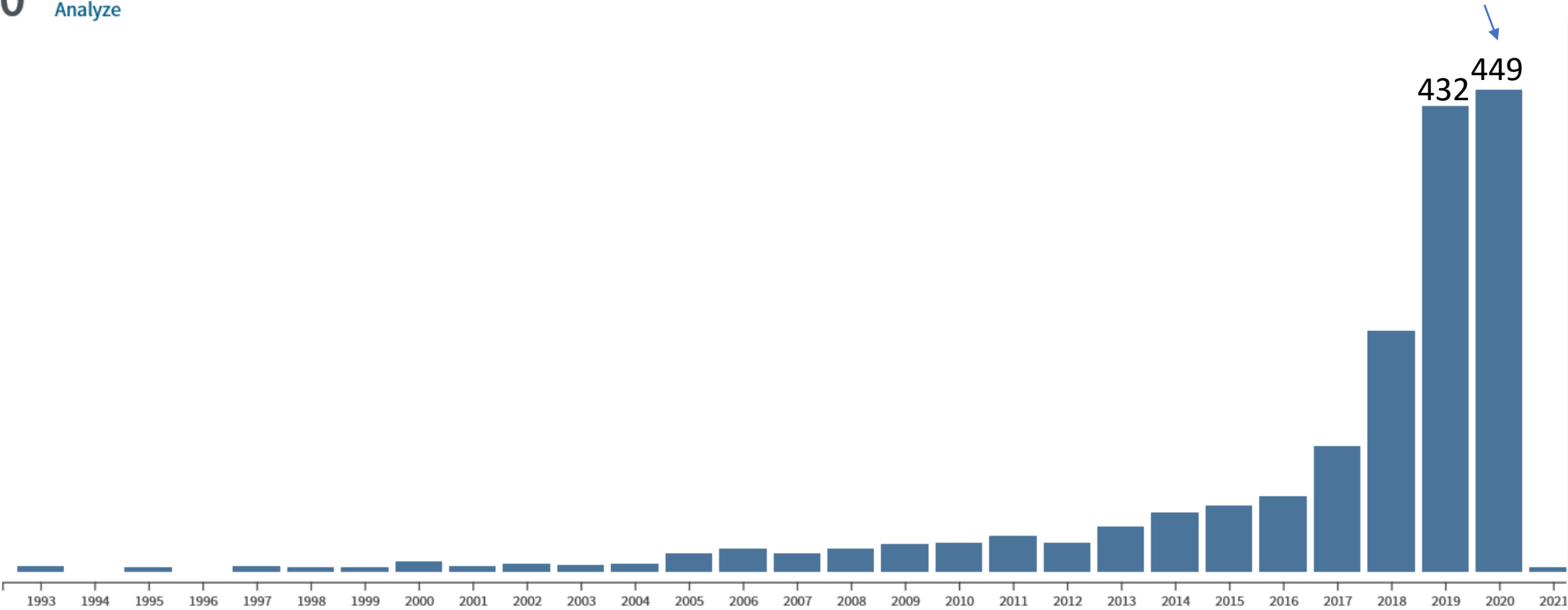
You searched for: TOPIC: (interpretable machine learning) OR TOPIC: (explainable artificial intelligence)  
Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC.  
[...Less](#)

This report reflects citations to source items indexed within Web of Science Core Collection. Perform a Cited Reference Search to include citations to items not indexed within Web of Science Core Collection.

Export Data: Save to Excel File


Total Publications  
1 640 [Analyze](#)

only till November

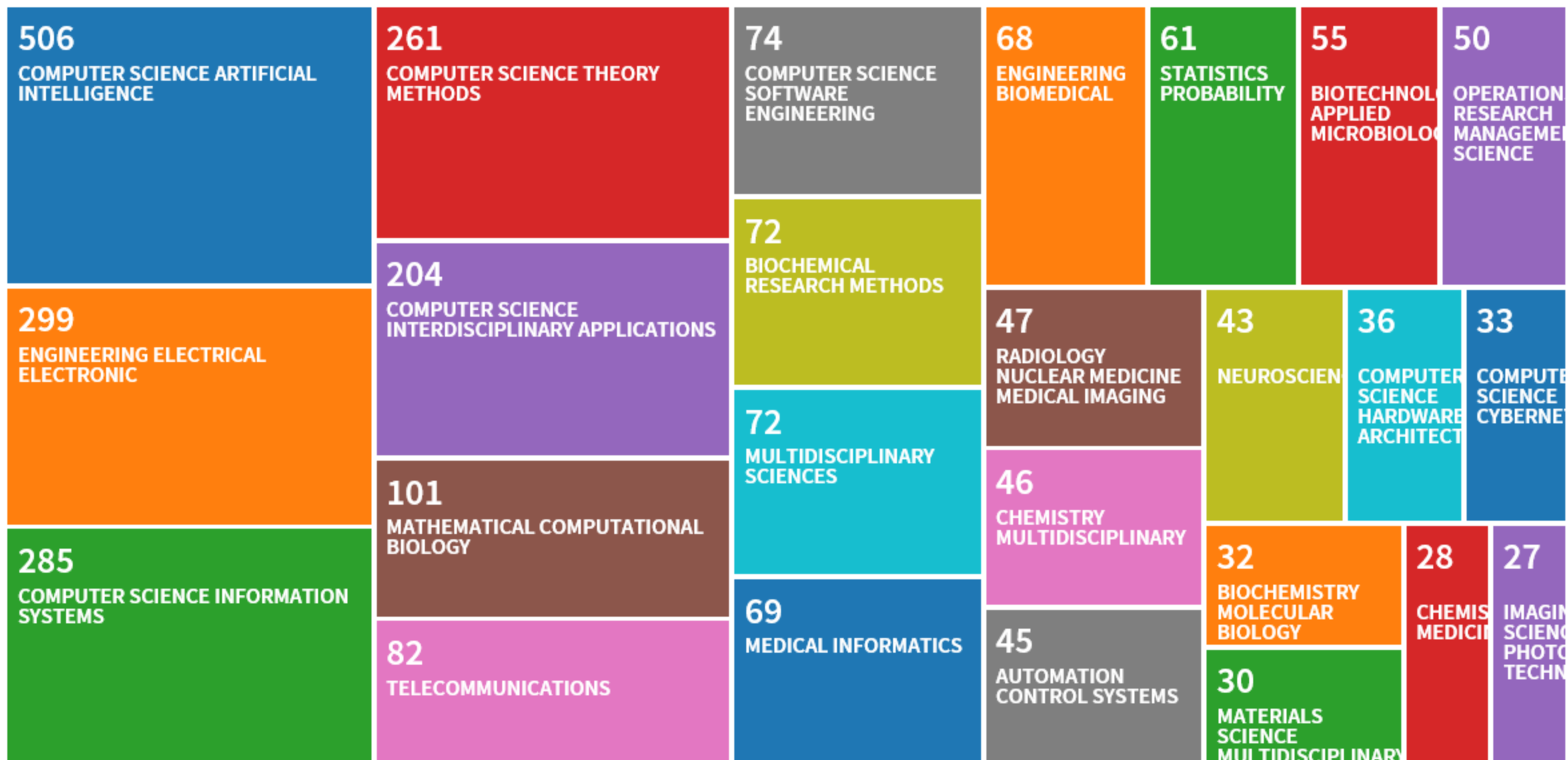


Visualization **Treemap**

Number of results **25**

 Download

Hide



# Przeglądówka przeglądówek

### A survey of surveys on the use of visualization for interpreting machine learning models

Angelos Chatzimpampas , Rafael M. Martins, Ilir Jusufi, more...

[Show all authors](#) ▾

First Published March 19, 2020 | Research Article |



<https://doi.org/10.1177/1473871620904671>

[Article information](#) ▾



#### Abstract

Research in machine learning has become very popular in recent years, with many types of models proposed to comprehend and predict patterns and trends in data originating from different domains. As these models get more and more complex, it also becomes harder for users to assess and trust their results, since their internal operations are mostly hidden in black boxes. The interpretation of machine learning models is currently a hot topic in the information visualization community, with results showing that insights from machine learning models can lead to better predictions and improve the trustworthiness of the results. Due to this, multiple (and extensive) survey articles have been published recently trying to summarize the high number of original research papers published on the topic. But there is not always a clear definition of what these surveys cover, what is the overlap between them, which types of machine learning models they deal with, or what exactly is the scenario that the readers will find in each of them. In this article, we present a meta-analysis (i.e. a “survey of surveys”) of manually collected survey papers that refer to the visual interpretation of machine learning models, including the papers discussed in the selected surveys. The aim of our article is to serve both as a detailed summary and as a guide through this survey ecosystem by acquiring, cataloging, and presenting fundamental knowledge of the state of the art and research opportunities in the area. Our results confirm the increasing trend of interpreting machine learning with visualizations in the past years, and that visualization can assist in, for example, online training processes of deep learning models and enhancing trust into machine learning. However, the question of exactly how this assistance should take place is still considered as an open challenge of the visualization community.

Article available

Vol 19, Issue 1

Related Articles

Similar Articles

[Task Cube: A conceptual space for visualization](#)

[Show details](#)

[Interactive tool for collaborative visualization](#)

[Show details](#)

[Collaborative visualization challenges, a survey](#)

[Show details](#)

Articles Citing

[The State of Trust in Machine Learning](#)

[Show details](#)

[Scenario-Based Elicitation for Explainable AI](#)

[Show details](#)



# https://arxiv.org/abs/2009.13248



Cornell University

the S

arXiv.org > cs > arXiv:2009.13248

Search...

Help | Advanc

Computer Science > Machine Learning

[Submitted on 24 Sep 2020 (v1), last revised 22 Oct 2020 (this version, v2)]

## Landscape of R packages for eXplainable Artificial Intelligence

Szymon Maksymiuk, Alicja Gosiewska, Przemyslaw Biecek

The growing availability of data and computing power fuels the development of predictive models. In order to ensure the safe and effective functioning of such models, we need methods for exploration, debugging, and validation. New methods and tools for this purpose are being developed within the eXplainable Artificial Intelligence (XAI) subdomain of machine learning. In this work (1) we present the taxonomy of methods for model explanations, (2) we identify and compare 27 packages available in R to perform XAI analysis, (3) we present an example of an application of particular packages, (4) we acknowledge recent trends in XAI. The article is primarily devoted to the tools available in R, but since it is easy to integrate the Python code, we will also show examples for the most popular libraries from Python.

Comments: 20 pages

Subjects: **Machine Learning (cs.LG)**; Machine Learning (stat.ML)

Cite as: [arXiv:2009.13248](https://arxiv.org/abs/2009.13248) [cs.LG]

(or [arXiv:2009.13248v2](https://arxiv.org/abs/2009.13248v2) [cs.LG] for this version)

### Submission history

From: Przemyslaw Biecek [[view email](#)]

[v1] Thu, 24 Sep 2020 16:54:57 UTC (1,538 KB)

[v2] Thu, 22 Oct 2020 09:28:06 UTC (1,644 KB)



# https://arxiv.org/abs/2010



Cornell University

arXiv.org > stat > arXiv:2010.09337

Search...

Help | Adv

**Statistics > Machine Learning**

*[Submitted on 19 Oct 2020]*

## Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges

Christoph Molnar, Giuseppe Casalicchio, Bernd Bischl

We present a brief history of the field of interpretable machine learning (IML), give an overview of state-of-the-art interpretation methods, and discuss challenges. Research in IML has boomed in recent years. As young as the field is, it has over 200 years old roots in regression modeling and rule-based machine learning, starting in the 1960s. Recently, many new IML methods have been proposed, many of them model-agnostic, but also interpretation techniques specific to deep learning and tree-based ensembles. IML methods either directly analyze model components, study sensitivity to input perturbations, or analyze local or global surrogate approximations of the ML model. The field approaches a state of readiness and stability, with many methods not only proposed in research, but also implemented in open-source software. But many important challenges remain for IML, such as dealing with dependent features, causal interpretation, and uncertainty estimation, which need to be resolved for its successful application to scientific problems. A further challenge is a missing rigorous definition of interpretability, which is accepted by the community. To address the challenges and advance the field, we urge to recall our roots of interpretable, data-driven modeling in statistics and (rule-based) ML, but also to consider other areas such as sensitivity analysis, causal inference, and the social sciences.

Subjects: **Machine Learning (stat.ML)**; Machine Learning (cs.LG)

Cite as: [arXiv:2010.09337](#) **[stat.ML]**

(or [arXiv:2010.09337v1](#) **[stat.ML]** for this version)

### Submission history

From: Christoph Molnar [\[view email\]](#)

**[v1]** Mon, 19 Oct 2020 09:20:03 UTC (214 KB)

# Challenges

## Definition of Interpretability

A lack of definition for the term “interpretability” is a common critique of the field

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)

Lipton, Z.C.: The mythos of model interpretability. Queue 16(3), 31–57 (2018)

# Challenges

## Definition of Interpretability

How to measure the quality of an explanation?

- No ground truth
- We can measure, for example:
  - fidelity (how well an explanation approximates the ML model),
  - sensitivity to perturbations,
  - sparsity,
  - interaction strength.

But still no established best practice on how to evaluate explanations.

# Challenges

## Statistical Uncertainty and Inference

- The model itself, but also its explanations are subject to uncertainty.
  - First research is working towards quantifying uncertainty of explanations
    - Feature importance

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20(177), 1–81 (2019)
    - Shapley values

Williamson, B.D., Feng, J.: Efficient nonparametric statistical inference on population feature importance using Shapley values. *arXiv preprint arXiv:2006.09481* (2020)

# Challenges

## Statistical Uncertainty and Inference

- The model itself, but also its explanations are subject to uncertainty.
  - First research is working towards quantifying uncertainty of explanations
    - Feature importance

Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research* 20(177), 1–81 (2019)
    - Shapley values

Williamson, B.D., Feng, J.: Efficient nonparametric statistical inference on population feature importance using Shapley values. *arXiv preprint arXiv:2006.09481* (2020)
- What about statistical inference?
  - clearly stated assumptions to get better diagnostic tools for testing.

# Challenges

## Causal Interpretation

- When are we allowed to make causal interpretations of an ML model?
- First steps:
  - Permutation feature importance

Konig, G., Molnar, C., Bischl, B., Grosse-Wentrup, M.: Relative feature importance. arXiv preprint arXiv:2007.08283 (2020)
  - Shapley values

Ma, S., Tourani, R.: Predictive and causal implications of using Shapley value for model interpretation. In: Proceedings of the 2020 KDD Workshop on Causal Discovery. pp. 23–38. PMLR (2020)

# Challenges

## A more dynamic and holistic view

- ML models are usually not used in a static and isolated way, but are embedded in some process or product, and interact with people.
- How to explain predictions to individuals with diverse knowledge and backgrounds?
- The need of interpretability on the level of an institution or society in general.
- A work for human-computer interaction, psychology and sociology.

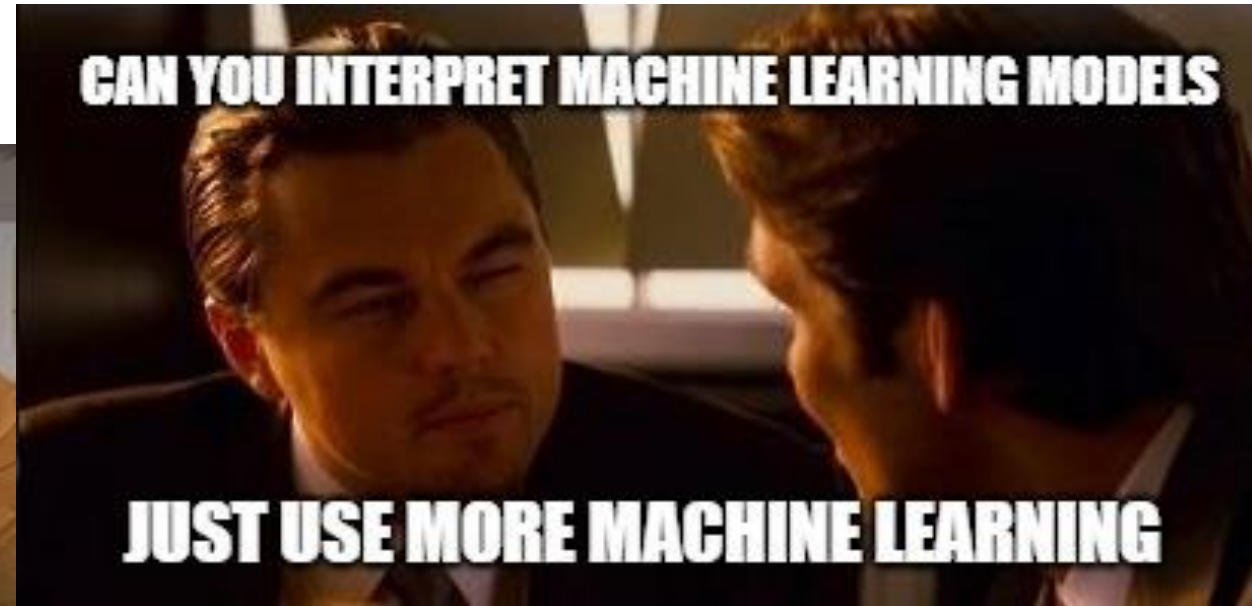


# Challenges

## A lack of XAI/IML memes



<https://mlr-org.com/docs/2018-04-30-interpretable-machine-learning-impl-and-mlr/>



<https://medium.com/towards-artificial-intelligence/show-me-the-black-box-3495dd6ff52c>



Alicja Gosiewska

OBSERWUJESZ

[Warsaw University of Technology](#)

Zweryfikowany adres z mini.pw.edu.pl - [Strona główna](#)

<input type="checkbox"/> TYTUŁ	+	⋮	CYTOWANE PRZEZ	ROK
<input type="checkbox"/> <a href="#">iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models</a>			6	2019
A Gosiewska, P Biecek arXiv preprint arXiv:1903.11420				
<input type="checkbox"/> <a href="#">Do Not Trust Additive Explanations</a>			4	2019
A Gosiewska, P Biecek arXiv preprint arXiv:1903.11420				
<input type="checkbox"/> <a href="#">auditor: an R Package for Model-Agnostic Visual Validation and Diagnostics</a>			4	2019
A Gosiewska, P Biecek The R Journal 11 (2), 85–98				
<input type="checkbox"/> <a href="#">Models in the wild: On corruption robustness of neural nlp systems</a>			3	2019
B Rychalska, D Basaj, A Gosiewska, P Biecek International Conference on Neural Information Processing, 235-247				
<input type="checkbox"/> <a href="#">SAFE ML: Surrogate Assisted Feature Extraction for Model Learning</a>			1	2019
A Gosiewska, A Gacek, P Lubon, P Biecek arXiv preprint arXiv:1902.11035				
<input type="checkbox"/> <a href="#">Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring</a>				2020
M Bücke, G Szepannek, A Gosiewska, P Biecek arXiv preprint arXiv:2009.13384				
<input type="checkbox"/> <a href="#">Landscape of R packages for eXplainable Artificial Intelligence</a>				2020
S Maksymiuk, A Gosiewska, P Biecek arXiv preprint arXiv:2009.13248				
<input type="checkbox"/> <a href="#">Interpretable Meta-Measure for Model Performance</a>				2020
A Gosiewska, K Woznica, P Biecek arXiv preprint arXiv:2006.02293				
<input type="checkbox"/> <a href="#">Lifting Interpretability-Performance Trade-off via Automated Feature Engineering</a>				2020
A Gosiewska, P Biecek arXiv preprint arXiv:2002.04267				
<input type="checkbox"/> <a href="#">EPP: interpretable score of model predictive power</a>				2019
A Gosiewska, M Bakala, K Woznica, M Zwolinski, P Biecek arXiv preprint arXiv:1908.09213				
<input type="checkbox"/> <a href="#">survxai: an R package for structure-agnostic explanations of survival models</a>				2018
A Grudziak, A Gosiewska, P Biecek Journal of Open Source Software 3 (31), 961				

Cytowane przez

	Wszystkie	Od 2015
Cytowania	18	18
h-indeks	3	3
i10-indeks	0	0



Współautorzy

EDYTUJ



Przemysław Biecek  
Warsaw University of Technology

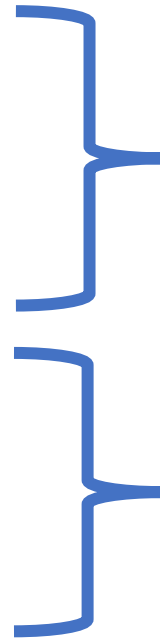


# Nieopublikowane

- iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models
- SAFE ML: Surrogate Assisted Feature Extraction for Model Learning
- Landscape of R packages for eXplainable Artificial Intelligence
- Interpretable Meta-Measure for Model Performance
- Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring

# Nieopublikowane

- iBreakDown: Uncertainty of Model Explanations for Non-additive Predictive Models
- SAFE ML: Surrogate Assisted Feature Extraction for Model Learning
- Landscape of R packages for eXplainable Artificial Intelligence
- Interpretable Meta-Measure for Model Performance
- Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring



Każdy wysłany w odpowiednio  
5 miejsc i 4 miejsca.

Było.

# https://arxiv.org/abs/2009.13384



Cornell University

the Sim

arXiv.org > stat > arXiv:2009.13384

Search...

Help | Advanced

Statistics > Machine Learning

[Submitted on 28 Sep 2020]

## Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring

Michael Bücker, Gero Szepannek, Alicja Gosiewska, Przemyslaw Biecek

A major requirement for credit scoring models is to provide a maximally accurate risk prediction. Additionally, regulators demand these models to be transparent and auditable. Thus, in credit scoring, very simple predictive models such as logistic regression or decision trees are still widely used and the superior predictive power of modern machine learning algorithms cannot be fully leveraged. Significant potential is therefore missed, leading to higher reserves or more credit defaults. This paper works out different dimensions that have to be considered for making credit scoring models understandable and presents a framework for making "black box" machine learning models transparent, auditable and explainable. Following this framework, we present an overview of techniques, demonstrate how they can be applied in credit scoring and how results compare to the interpretability of score cards. A real world case study shows that a comparable degree of interpretability can be achieved while machine learning techniques keep their ability to improve predictive power.

Subjects: **Machine Learning (stat.ML)**; Machine Learning (cs.LG); General Economics (econ.GN); Applications (stat.AP); Methodology (stat.ME)

Cite as: [arXiv:2009.13384](#) [stat.ML]

(or [arXiv:2009.13384v1](#) [stat.ML] for this version)

### Submission history

From: Michael Bücker [[view email](#)]

[v1] Mon, 28 Sep 2020 15:00:13 UTC (1,795 KB)



Michael Bücker  
University of Applied Sciences,  
Münster School of Business



Gero Szepannek  
Stralsund University of Applied  
Sciences







Przemysław Biecek  
University of Warsaw,  
Warsaw University of Technology

# It all started from the presentation at the conference



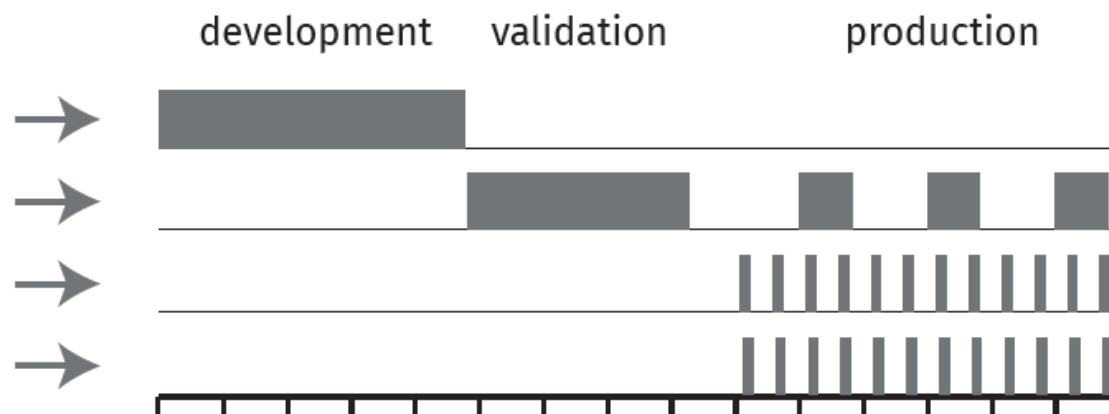
## Who?

### Stakeholders

-  model developer - internal
-  model auditor - internal or external
-  credit officer - internal
-  bank customer - external

## When?

### Model lifetime



## What?

### Specific needs

- Q1. Is model A better than B?
- Q2. Which variable is most important?
- Q3. What a customer can do to improve?
- Q4. Why a model prediction was wrong?

## How?

### XAI pyramid

- metric
- parts
- profile
- diagnostic



# Comparative study of Scorecards and Explainable Machine Learning



## **Congratulations to the winners!**

"Sanjeeb Dash, Oktay Günlük and Dennis Wei, representing IBM Research, were this year's challenge winners. The winning team received the highest score in an empirical evaluation method that considered how useful explanations are for a data scientist with the domain knowledge in the absence of model prediction, as well as how long it takes for such a data scientist to go through the explanations. For their achievements, the IBM team earned a \$5,000 prize. Receiving Honorable Mention and overall second place was New York University's team, comprised of Steffen Holter, Oscar Gomez, and Enrico Bertini. The NYU team took home \$2,000."

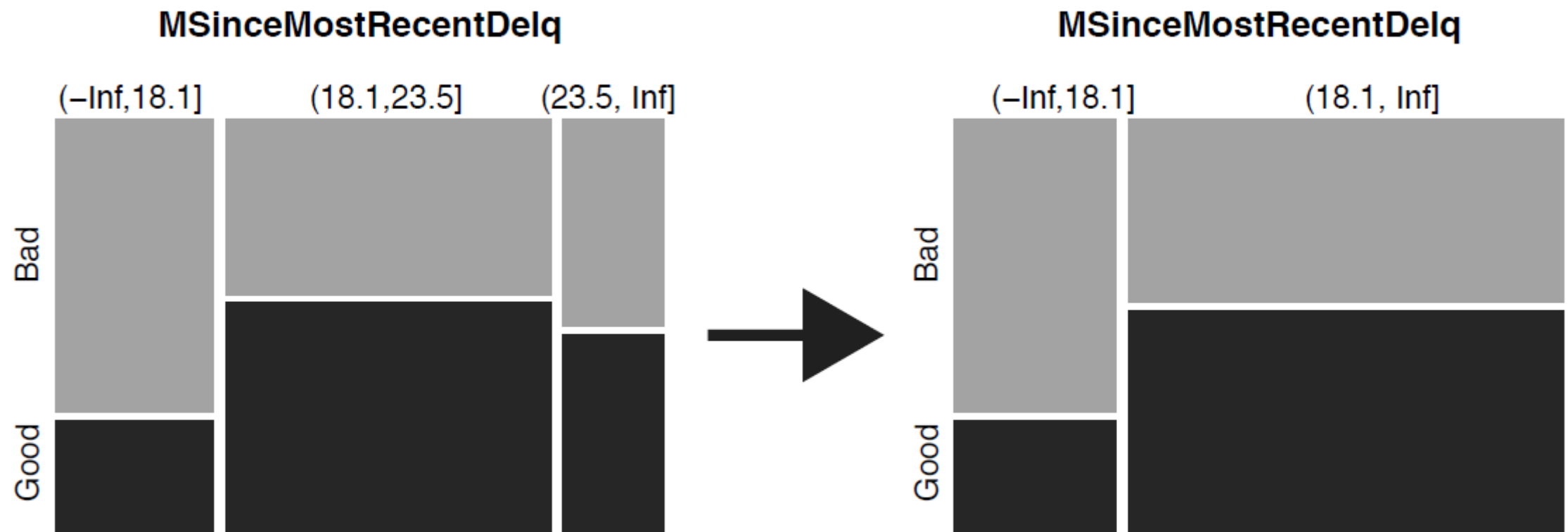
[Read the full press release.](#)

## **Read about the solution from the FICO Recognition Award winning team at Duke University.**

[We Didn't Explain the Black Box – We Replaced it with an Interpretable Model](#)

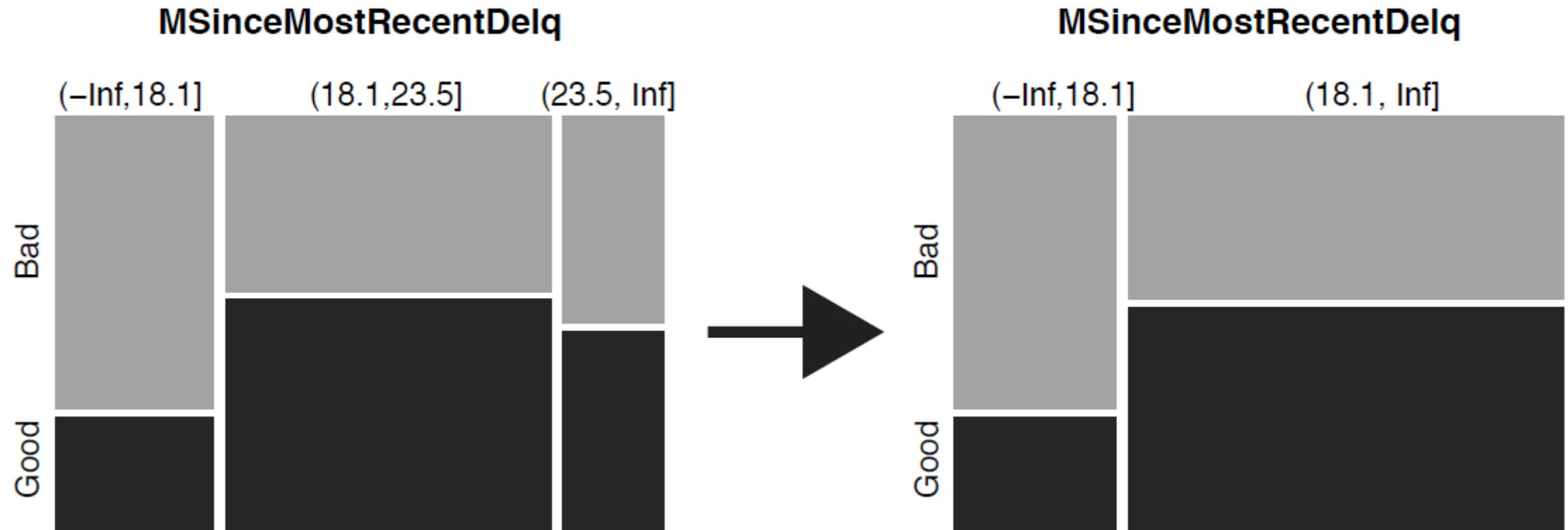
# A scorecard model

Level	Points	% Population	% Default
Total population		1.00	0.52
Intercept	385		
ExternalRiskEstimate			
(-Inf,67.1]	-11	0.333	0.78
(67.1,72.6]	-2	0.216	0.58
(72.6,81.3]	5	0.253	0.39
(81.3, Inf]	14	0.196	0.19
AverageMInFile			
(-Inf,59.3]	-13	0.259	0.71
(59.3,80.4]	0	0.327	0.52
(80.4, Inf]	8	0.412	0.40
NetFractionRevolvingBurden			
(59.4, Inf]	-13	0.204	0.78
(26.3,59.4]	-3	0.359	0.59
(-Inf,26.3]	8	0.436	0.35
PercentTradesNeverDelq			
(-Inf,58.4]	-24	0.204	0.90
(58.4,83.2]	-12	0.131	0.75
(83.2,95.4]	-4	0.308	0.60
(95.4, Inf]	5	0.534	0.40
MSinceMostRecentInqexcl7days			
(-Inf,2.68]	-3	0.813	0.56
(2.68,10.5]	12	0.129	0.38
(10.5, Inf]	21	0.057	0.29



**Figure 2.** Example: automatic vs. manual binning for the variable months since the most recent delinquency.

# Problems?



**Figure 2.** Example: automatic vs. manual binning for the variable months since the most recent delinquency.

# Challengers



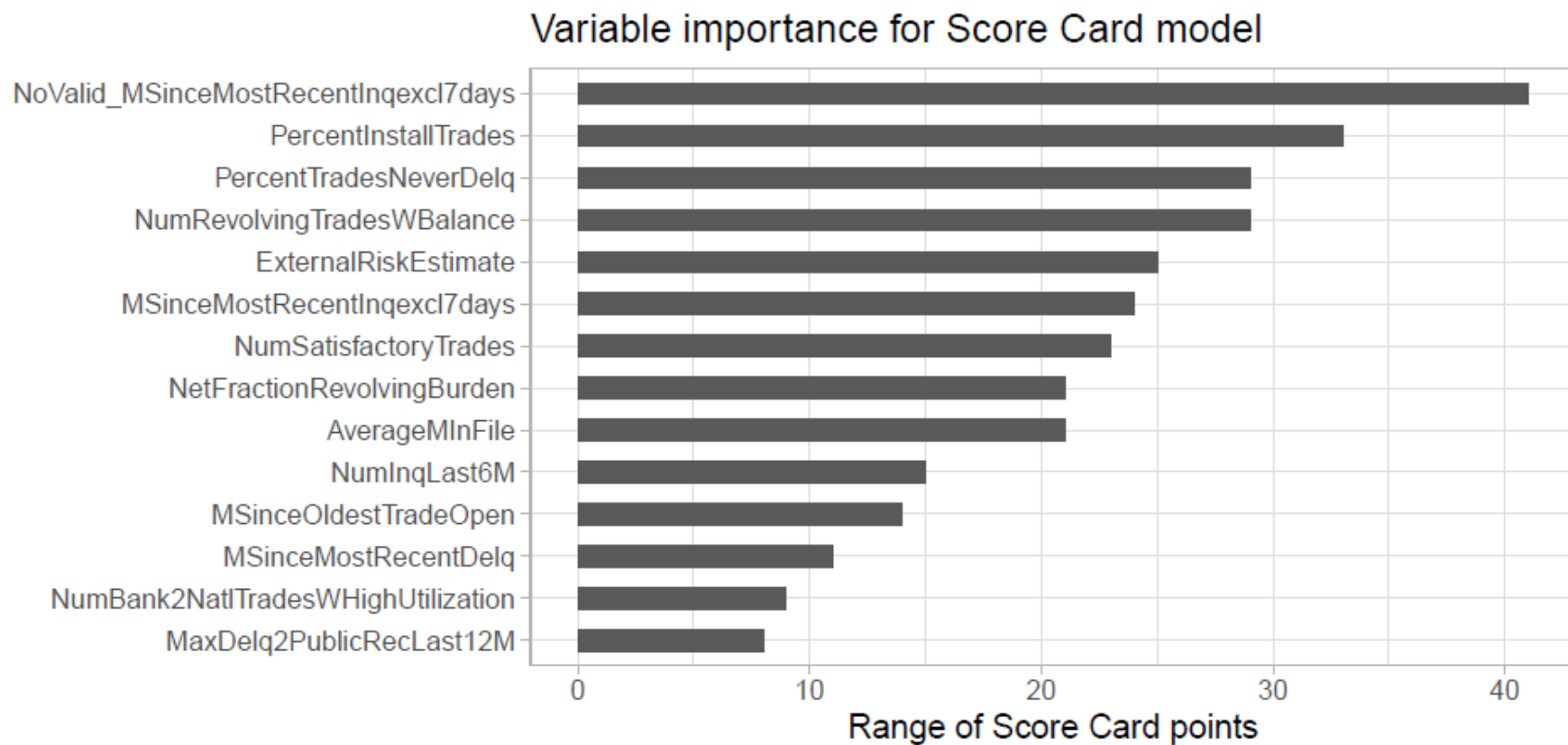
**Figure 3.** Model performance measured by AUC on test data

# Challengers



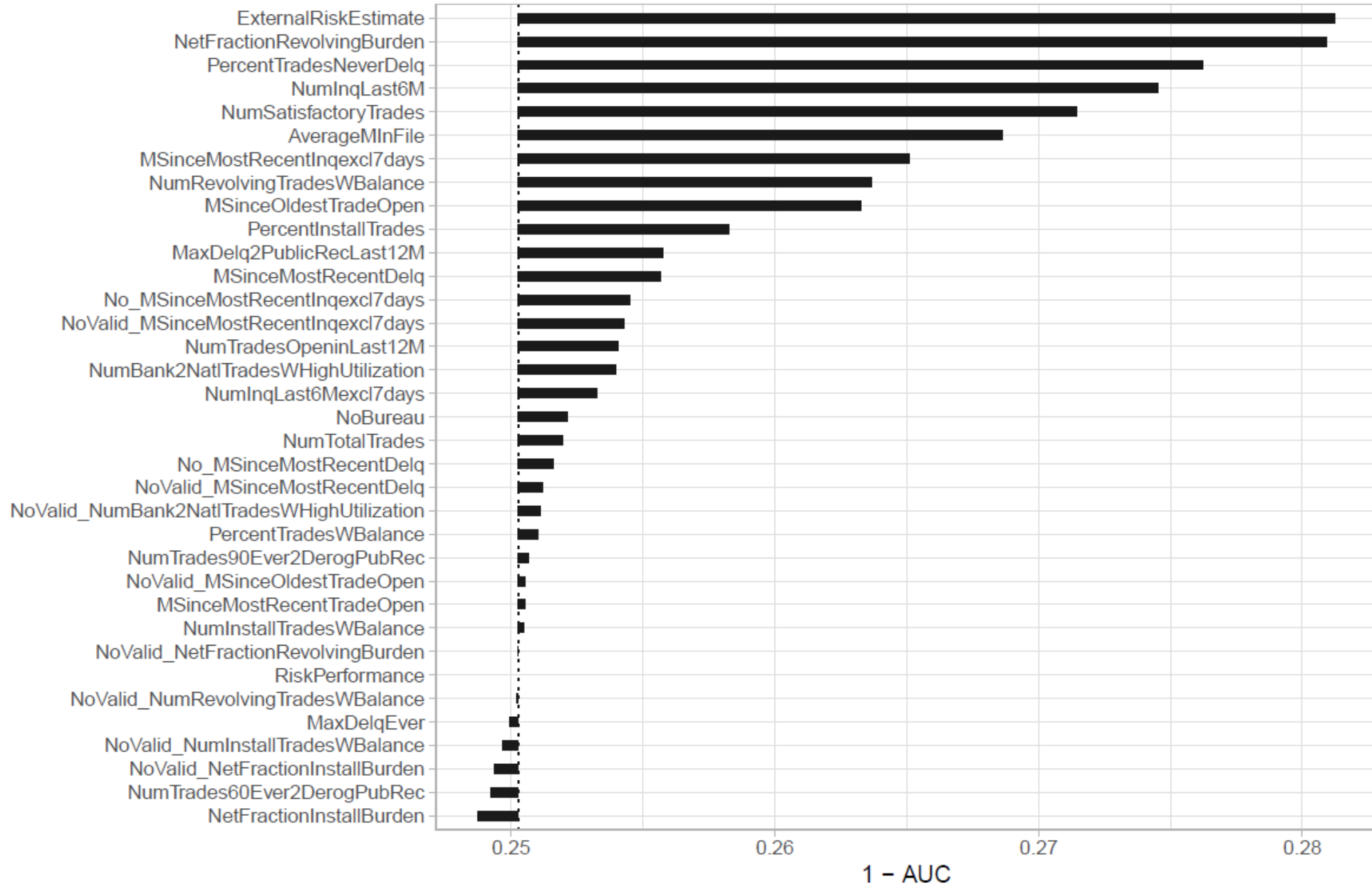


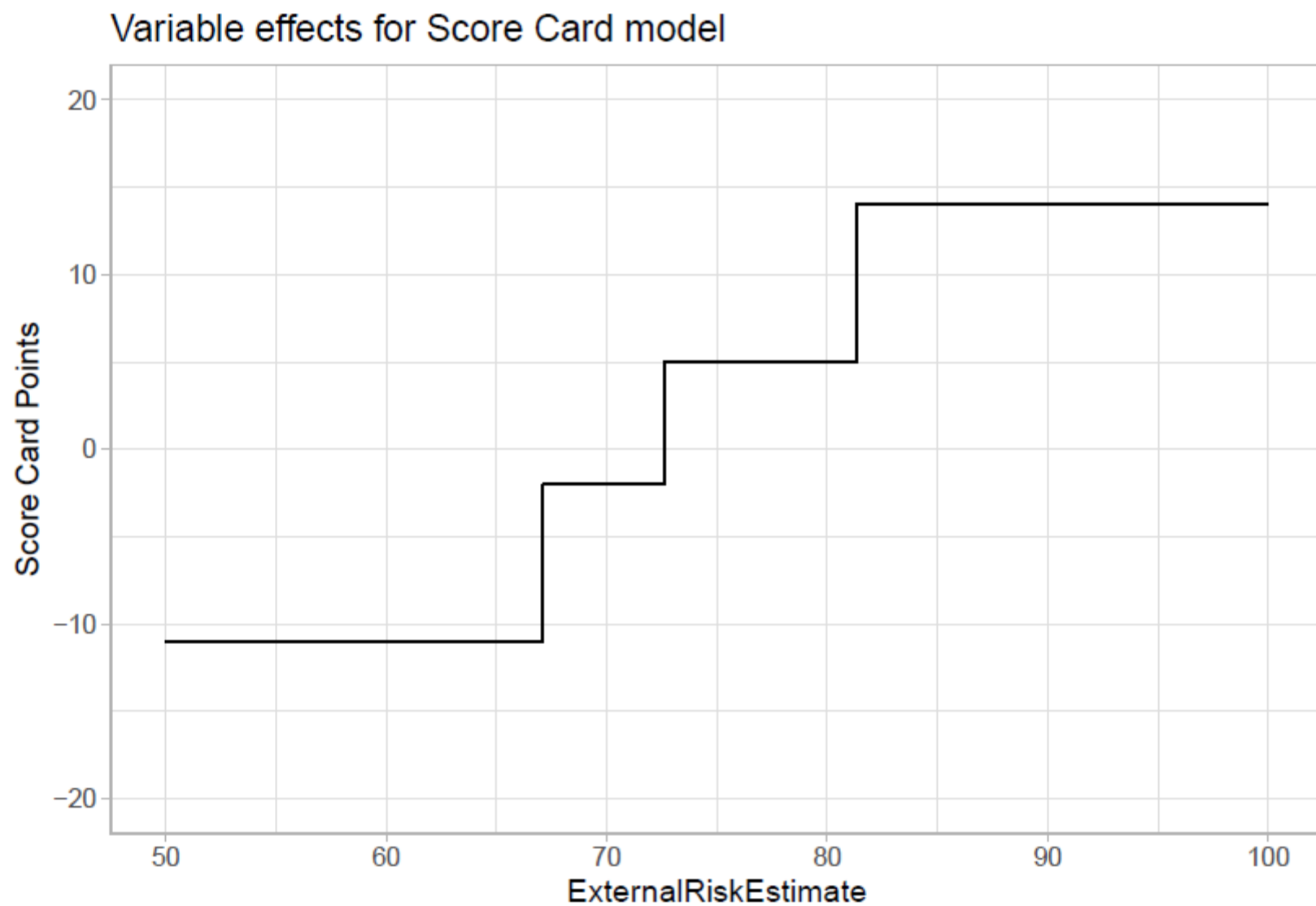
Why XAI?



**Figure 5.** Range of scorecard points as measure of variable importance for the Score Card

Variable importance for GBM 10000 model





**Figure 7.** Marginal effect of variable 'ExternalRiskEstimate' based on Score Card points

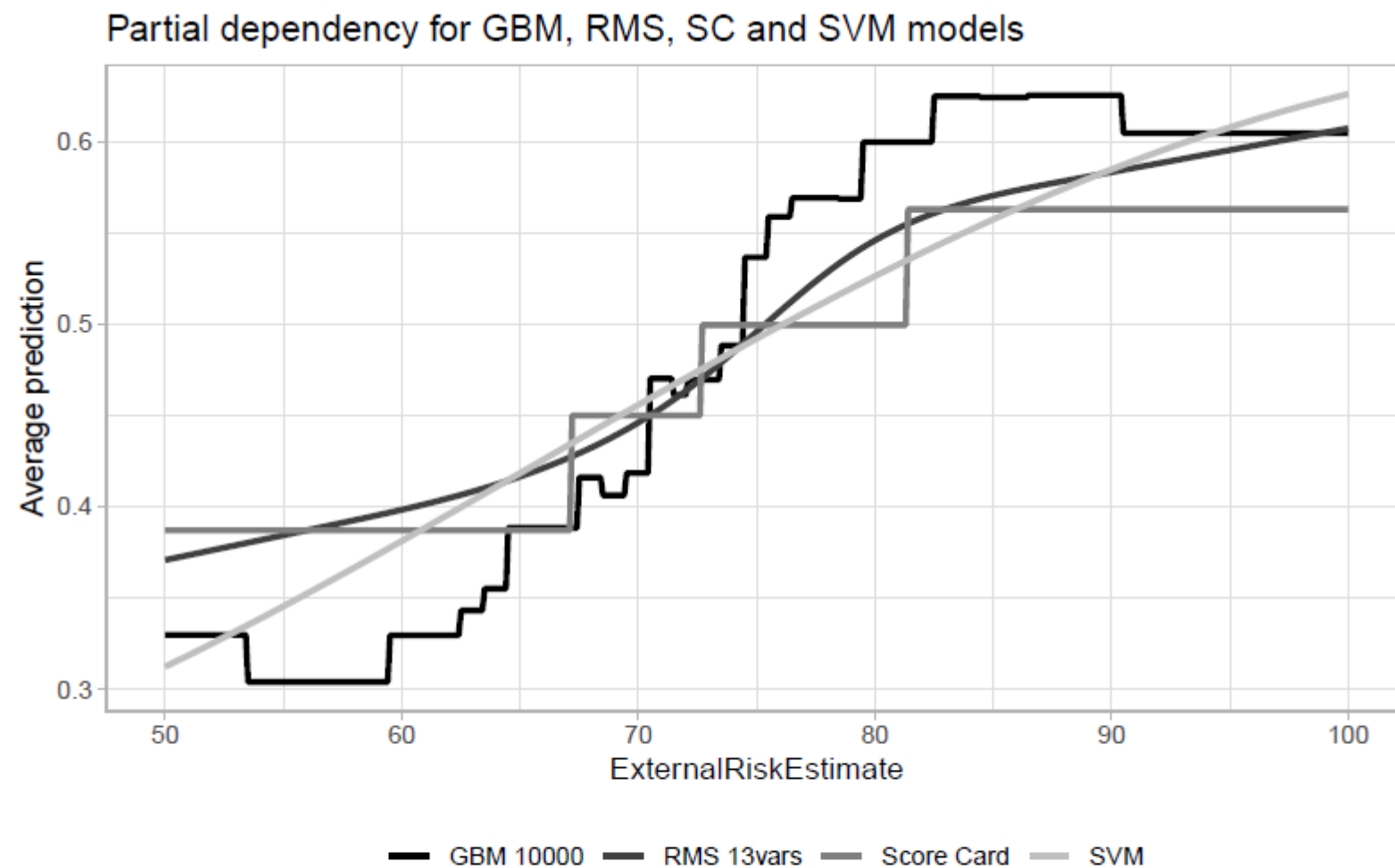
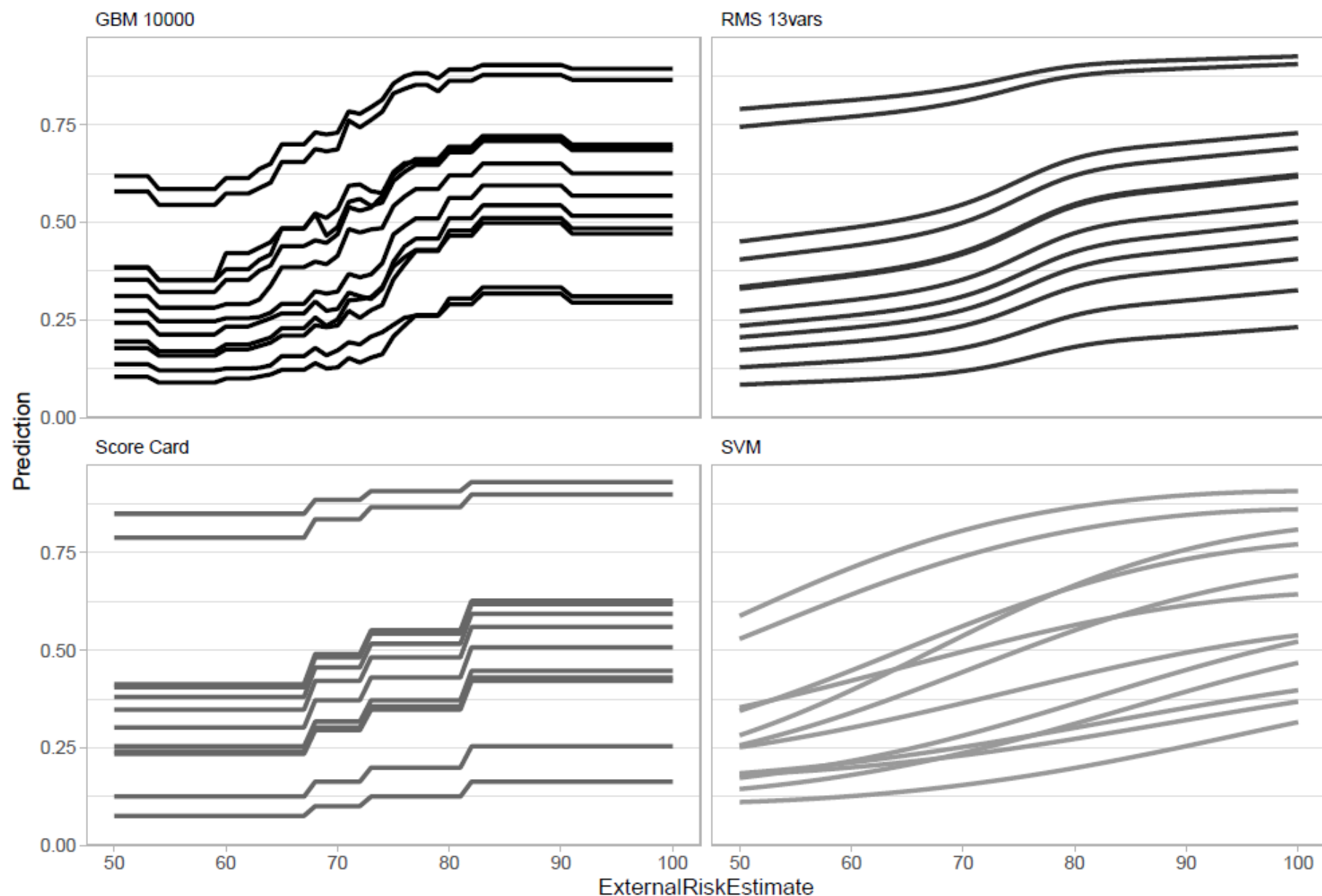


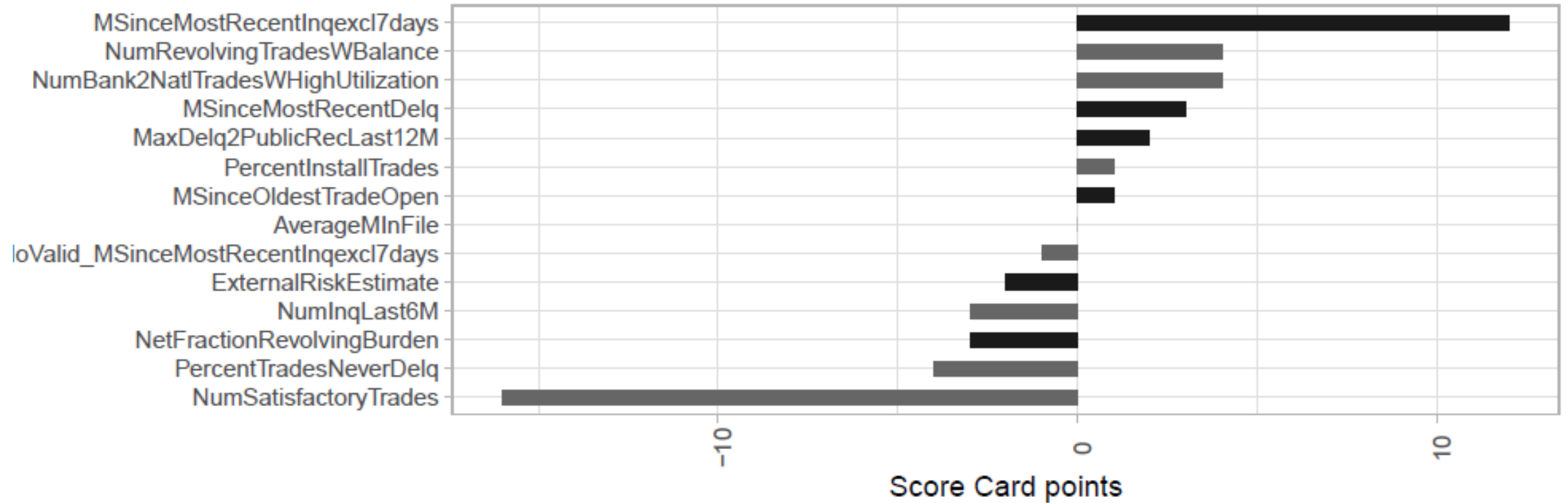
Figure 10. Partial Dependence plots for the variable 'ExternalRiskEstimate' based on the selected Machine Learning models

# Individual model profiles for GBM, RMS, SC and SVM models

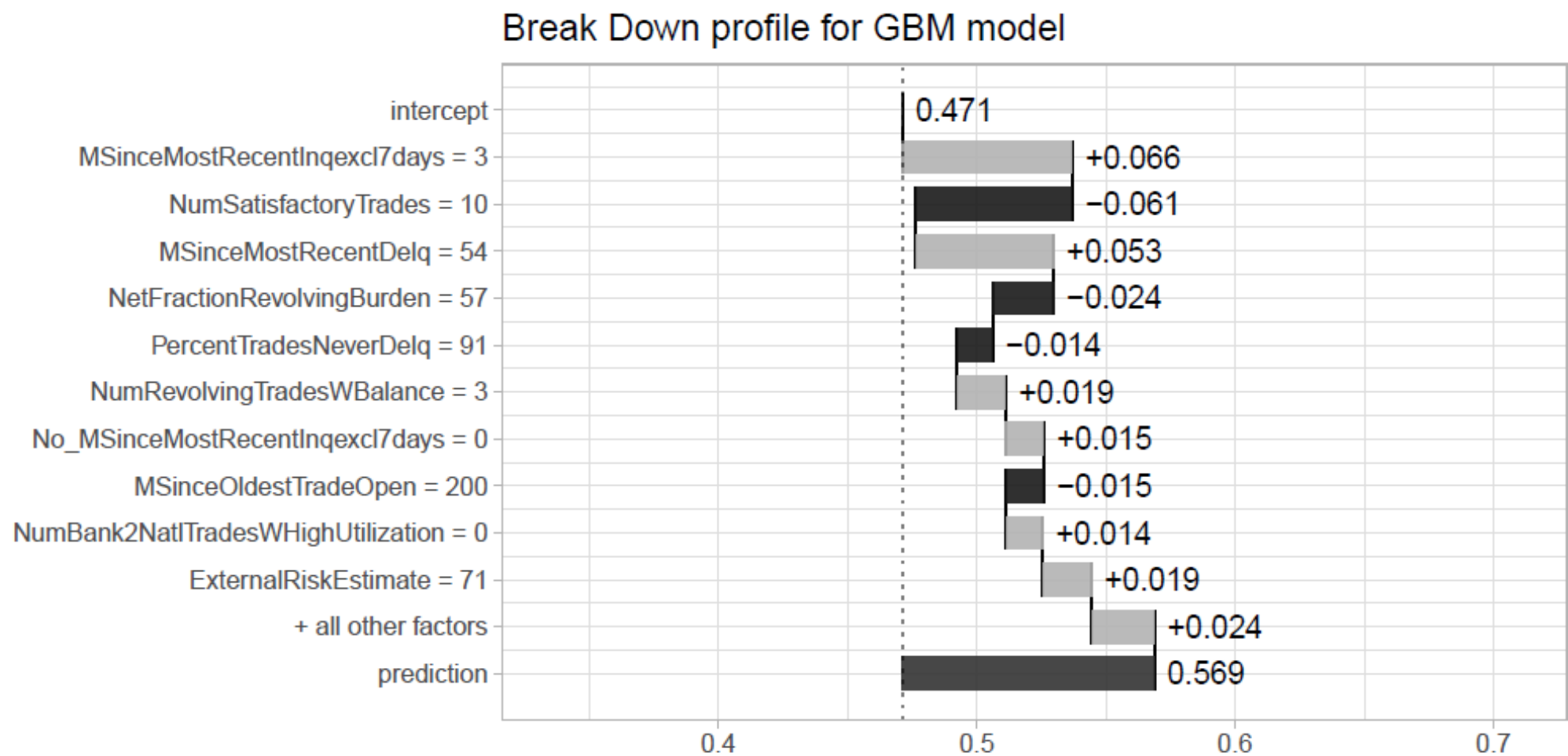


GBM 10000 RMS 13vars Score Card SVM

Individual variable importance for Score Card model







**Figure 12.** Additive breakdown for a single prediction as individual model agnostic explanation of variable importance for Machine Learning models (here: Gradient Boosting)