

Trustworthy Decision-Making via Optimal Scoring Systems

How Discrete Optimization Can Help Create Models that are Easier to Understand and Validate

Berk Ustun

Scoring systems are sparse linear classification models with small integer coefficients. Starting with the work of Burgess in 1928, such models have been extensively used for data-driven decision-making in domains where humans have traditionally made decisions. Scoring systems are currently used for numerous applications in medicine (e.g. to predict mortality of various medical conditions¹), criminal justice (e.g. to assess recidivism risk²), and finance (e.g. to assess creditworthiness and support investment decisions³).

The widespread deployment of scoring systems is inherently related to the fact that simple linear models with small integer coefficients are easy for humans to understand, validate, and trust. Sparsity and small integer coefficients make quick predictions through simple arithmetic, without a computer or a calculator. These qualities also address key limitations in human cognition, such as limits in our ability to handle 4+ items in working memory [1], and to track associations between 3+ entities [2]. Considering these limitations, sparsity and small integer coefficients can help users understand how multiple input variables are used in the prediction. This allows users to easily validate the model, and provides them with the option to overrule the model in an informed manner when needed.

In spite of extensive deployment over the past century, there has been no standardized approach to build scoring systems. This is partially due to the fact that models have to satisfy domain specific *operational constraints* to be deployed. Such constraints are difficult to address in a systematic manner because they are related to ill-defined model qualities (e.g. usability, understandability, and alignment with domain expertise) that vary significantly across applications. As a result, scoring systems are still developed *ad hoc*. In some cases, models are built by combining traditional statistical methods with heuristics and expert judgement. In others, models are hand-crafted by a panel of experts and data is used for validation purposes only.

New Methods to Build Scoring Systems: In this white paper, we describe two new machine learning methods to create scoring systems through discrete optimization:

- SLIM (Supersparse Linear Integer Models) to create optimized scoring systems for decision-making [6] (see Figure 1);
- RISKSLIM (Risk-calibrated Supersparse Linear Integer Models) to create optimized scoring systems for risk assessment [5, 7] (see Figure 2).

Unlike traditional machine learning methods, SLIM and RISKSLIM aim to solve hard NP-hard optimization problems to optimize and constrain exact measures of performance and form (e.g., the number of mistakes and variables): SLIM, for instance, requires the solution to a mixed-integer program (MIP), while RISKSLIM requires the solution to a mixed-integer nonlinear program (MINLP). Solving these problems with off-the-shelf commercial solvers (e.g., CPLEX) only works consistently for small datasets. As such, a key part of this work involves the development of new techniques that can solve these optimization problems for the largest possible datasets. Paired with such techniques, SLIM and RISKSLIM can produce scoring systems that are fully optimized for performance, sparsity, integer coefficients, and other real-world constraints.

Benefits in Model Development and Decision Making: Since SLIM and RISKSLIM fit scoring systems using discrete optimization, they can produce simple models that fit on an index card and yet perform in line with modern ML models. This approach also allows users can easily customized models

¹See <https://www.mdcalc.com> for an extensive list of medical scoring systems.

²See the risk assessment tool developed by the [Pennsylvania Sentencing Commission](#).

³See the [Piotroski F-score](#) to assess the strength of a company's balance sheet.

References

- [1] Nelson Cowan. The magical mystery four how is working memory capacity limited, and why? *Current directions in psychological science*, 19(1):51–57, 2010.
- [2] Dennis Jennings, Teresa M Amabile, and Lee Ross. Informal covariation assessment: Data-based vs. theory-based judgments. In D. Kahneman, P. Slovic, and A. Tversky, editors, *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.
- [3] Aaron F. Struck, Berk Ustun, Andres Rodriguez Ruiz, Jong Woo Lee, Suzette LaRoche, Lawrence J. Hirsch, Emily J. Gilmore, Cynthia Rudin, and Brandon M Westover. A practical risk score for EEG seizures in hospitalized patients. *JAMA Neurology*, 2017.
- [4] Berk Ustun, Lenard A Adler, Cynthia Rudin, Stephen V Faraone, Thomas J Spencer, Patricia Berglund, Michael J Gruber, and Ronald C Kessler. The World Health Organization Adult Attention-Deficit / Hyperactivity Disorder Self-Report Screening Scale for DSM-5. *JAMA Psychiatry*, 74(5):520–526, 2017.
- [5] Berk Ustun and Cynthia Rudin. Learning Optimized Risk Scores for Large-Scale Datasets. *arXiv:1610.00168*, 2016.
- [6] Berk Ustun and Cynthia Rudin. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016.
- [7] Berk Ustun and Cynthia Rudin. Optimized Risk Scores. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.
- [8] Berk Ustun, M.B. Westover, Cynthia Rudin, and Matt T. Bianchi. Clinical prediction models for sleep apnea: The importance of medical history over symptoms. *Journal of Clinical Sleep Medicine*, 12(2):161–168, 2016.
- [9] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable Classification Models for Recidivism Prediction. *Journal of the Royal Statistical Society: Series A*, 2016.