Effect of feature discretization on classification performance of explainable scoring-based machine learning model

Arkadiusz Pajor[1,2][0000−0002−3032−292X], Jakub Żołnierek[3][0000−0003−2946−870X], Bartlomiej Sniezynski[2][0000−0002−4206−9052], and Arkadiusz Sitek[1][0000−0002−0677−4002]

[1]Sano Centre for Computational Medicine, Cracow, Poland
[2] AGH University of Science and Technology, Cracow, Poland
[3] Maria Skłodowska-Curie Memorial Cancer Center, Warsaw, Poland

Abstract. We improve the utility of the Risk-calibrated Supersparse Linear Integer Model (RiskSLIM). It is a scoring system that is an interpretable machine learning classification model optimized for performance. Scoring systems are commonly used in healthcare and justice. We implement feature discretization (FD) in the hyperparameter optimization process to improve classification performance and refer to the new approach as FD-RiskSLIM. We test the approach using two medical applications. We compare the results of FD-RiskSLIM, RiskSLIM, and other machine learning (ML) models. We demonstrate that scoring models based on RiskSLIM, in addition to being interpretable, perform at least on par with the state-of-the-art ML models such as Gradient Boosting in terms of classification metrics. We show the superiority of FD-RiskSLIM over RiskSLIM.

1 Introduction

Machine Learning (ML) starts to play an important role in domains like healthcare where making high stake decisions is common. Examples of such domains include cancer prognosis [8], hypertension outcomes [5], heart diseases [16], and many others. Typically researchers focus on the predictive performance of the ML models. Gradient Boosting, Random Forest, or Artificial Neural Networks are considered state-of-the-art algorithms that show superior performance to simpler models like Decision Tree or Linear Regression. However, over the last few years regulations outlined in the General Protection Data Regulation (GDPR) and in particular "right to explanation" [13] emphasized the importance of ML-algorithm trustability, transparency, and fairness and have sparked a discussion around important needs for interpretability of ML models. In our research, we focus on algorithms that generate models with a high level of predictive performance and low complexity which are interpretable or explainable [25].

This work focuses on improvements of the Risk-calibrated Supersparse Linear Integer Model (RiskSLIM) introduced in [23] which is an interpretable ML

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

model achieving accuracy comparable to black-box models mentioned before. We demonstrate that our improved interpretable RiskSLIM algorithm outperforms them for healthcare-related examples used here.

The main contributions of this paper: (1) we add feature discretization (FD) to RiskSLIM algorithm [23], (2) we compare FD-RiskSLIM with RiskSLIM and other classical ML techniques. We use two examples of medical applications, namely the prediction of heart-failure patient outcome and the prediction of the outcome of kidney cancer treatment. We demonstrate that the performance of our interpretable models (FD-RiskSLIM and RiskSLIM) is comparable or even superior to classical non-interpretable models such as Random Forest or Gradient Boosting. It is important to note that we optimize all algorithms used in this paper in the same way, using the algorithm for hyperparameter optimization (see section 3.3).

2 Related Research

Our research was focused on medical decision support based on scoring systems. Scoring models are sparse linear models with integer coefficients that make them interpretable. Many popular scoring systems like Simplified Acute Physiology Score (SAPS) [15], Systemic Inflammatory Response Syndrome (SIRS) [4], Acute Physiologic Assessment and Chronic Health Evaluation (APACHE) [14], Stroke Risk Assessment in Atrial Fibrillation ($CHADS_2$) [12], Thrombolysis in Myocardial Infarction (TIMI) [3] to name a few. The models were built by domain experts based at least partially on their experience. In Figure 1, a scoring model $CHADS_2$ is presented. The $CHADS_2$ index was created by including independent risk factors: prior cerebral ischemia, history of hypertension, diabetes mellitus, congestive heart failure, and age of 75 years. Other factors like high blood pressure or sex were not included based on domain expertise even though similar scoring systems developed by others take them into account. The points contributing to the overall score were assigned arbitrarily. However, they were validated with the use of an exponential survival model which measured how the rate of stroke was affected by 1-point increases in $CHADS_2$.

Other classical scoring systems such as SAPS [15] were built with the use of machine learning techniques. Features contributing to the overall SAPS score were selected with the use of cross-validation. The final model exposed twenty relevant features, but the complexity of the model was reduced by using logistic regression.

Our method is built on a novel approach to scoring systems introduced in [22]. The authors developed SLIM, the predecessor of RiskSLIM, which can build scoring systems directly from data with no necessity of using domain knowledge. With the use of SLIM, the authors managed to predict Obstructive Sleep Apnea using various information available for the patients using polysomnography results as the ground truth[24]. Data used for such prediction included standard medical information such as age, BMI, gender, diabetes, smoking, and past problems with the heart reported by patients. Authors demonstrated that the

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

Features	Points
Congestive heart failure	1
Hypertension: blood pressure consistently above 140/90 mmHg (or treated hypertension on medication)	1
Age >= 75 years	1
Diabetes mellitus	1
Prior Stroke or TIA or Thromboembolism	2

Score	0	1	2	3	4	5	6
Risk (%)	1.9	2.8	4.0	5.9	8.5	12.5	18.2

Fig. 1. Structure of $CHADS_2$ scoring system. Sum of points evaluates to the risk as presented.

scoring system built by them performed better than the commonly used STOP-BANG [7] scoring system, which in addition to features used by SLIM took into consideration symptoms reported by patients. The same scientific group used the successor of SLIM, RiskSLIM to predict seizures. The prediction was based on patterns in continuous electroencephalography (cEEG) [21]. Authors created a new scoring system 2HELPS2B and demonstrated that it performed as well as neural networks, but with important advantages of interpretability and transparency [20].

3 Methods

3.1 Risk-calibrated Supersparse Linear Integer Model (RiskSLIM)

RiskSLIM was introduced by Ustun and Rudin in 2019 [23]. It is a scoring system similar to the predictive models designed by humans over the last century (e.g. $CHADS_2$ [11]). However, contrary to the traditional models, RiskSLIM determines integer score points (which are coefficients of the linear model) relying solely on the data, using non-linear integer optimization, instead of obtaining score points from experts. Authors assumed that domain knowledge may be incorporated in a form of specific constraints to the input and output variables. It was shown that RiskSLIM allows the creation of scoring systems that give accurate and interpretable results for decisions related to medicine and criminal justice [24], [21], [26].

To train a RiskSLIM model, a mixed-integer, non-linear problem with hard computational complexity has to be solved. It follows that the computation for a highly-dimensional problem is time-consuming and often impractical in medical settings. The approach uses constraints to make score points to be small integers. The objective function is a log-loss (the same as used in the logistic regression) that is minimized during model training. As a result, we get risk-calibrated scoring systems, in which predicted risks agree with risks calculated directly from the data [19]. The formula which is used for the estimation of the risk of the event under consideration (stroke, cancer death) consists of an intercept value and a calculated integer score which is a sum of all score points provided

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

by the model (score points are added to the score if the corresponding feature is present in the analyzed event):

$$\frac{1}{1+\exp(-intercept-score)}, \tag{1}$$

where the intercept in the expression (1) is determined during the model training.

Table 1. Sample numerical feature discretized using 3 subspaces represented by columns with white background. Discretization is done with overlapping (top image) and with disjoint regions (bottom image). In FD-RiskSLIM a number of subspaces and threshold values are hyperparameters that are optimized (see section 3.3). For presented experiments (see section 4) we utilized disjoint regions.

feature	feature < 4.0	feature ≥ 4.0	feature ≥ 13.67
25	0	1	1
16	0	1	1
9	0	1	0
4	0	1	0
1	1	0	0
0	1	0	0
1	1	0	0
4	0	1	0
9	0	1	0
16	0	1	1
25	0	1	1

feature	0 ≤ feature < 4.0	4.0 ≤ feature < 13.67	13.67 ≤ feature < 25
25	0	0	1
16	0	0	1
9	0	1	0
4	0	1	0
1	1	0	0
0	1	0	0
1	1	0	0
4	0	1	0
9	0	1	0
16	0	0	1
25	0	0	1

3.2 RiskSLIM with Feature Discretization (FD-RiskSLIM)

The original RiskSLIM algorithm assigns at most a single coefficient (number of score points) to a given feature, so in a case when there is a potential non-linear relation between the feature and the output variable the model performance can be compromised. Such a single RiskSLIM coefficient is similar to a coefficient of a linear regression model where modeling non-linear relations is difficult. We designed the tool which allows us to divide the space of a numerical feature

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

into subspaces (discretize it), so every subspace can have a different coefficient assigned.

The subspaces may be overlapping or disjoint from each other. For overlapping subspaces, one can define binary relation (lesser than, lesser or equal than, equal to, greater or equal than, greater than) between the actual feature value and a bin edge found by the selected discretization method. Table 1 presents a feature that is discretized in two ways, with overlapping and disjoint ranges as the output. For overlapping subspaces (upper part of the table) bin edges are values 4.0 and 13.67 and binary relations are *lesser than* and *greater or equal* for 4.0 and *greater or equal* for 13.67. For this kind of discretization, there can be many 1s in a row. For disjoint subspaces (lower part of the table) such binary relations are not applied. For this kind of discretization, there can be only a single 1 in a row.

To find appropriate discretization, an optimization algorithm is applied. The resulting bins depend on chosen discretization strategy (e.g. uniform, quantile, k-means, MDLP, etc.). In the optimization process hyperparameters like a number of bins for quantile discretization or maximum depth of a tree for MDLP discretizer are tuned. We also optimize RiskSLIM-specific hyperparameters such as a number and a magnitude of output coefficients and an intercept value. In our case, the goal of the optimization was to achieve maximal accuracy, but one can specify other metrics such as F1-score, Matthews correlation coefficient, etc. The full algorithm of FD-RiskSLIM is the following:

1. Define hyperparameters for RiskSLIM, as a grid of parameters,
2. Define hyperparameters for feature discretization, as a grid of parameters,
3. Pass these grids to the hyperparameter optimization framework,
4. Select the best model which is found by hyperparameter optimization,
5. Using a data test set compute the metrics of the model's performance.

3.3 Implementation

To perform a fair comparison of the performance of different machine learning algorithms, we optimized hyperparameters for all ML methods used in this paper using the Optuna hyperparameter optimization framework [2]. We wrapped the Optuna in a class to create a reusable tool to optimize any model, with any number and type of hyperparameters.

For feature discretization, we designed a solution that allows one to apply a discretizer of a choice. We provided the functionality that wraps the discretizer and exposes its ability through the constant interface. As a proof of concept, we used KBinsDiscretizer from scikit-learn library [17], which can use a few strategies of discretization (uniform, quantile, k-means) and the MDLP (Minimum Description Length Principle) discretizer [10].

For RiskSLIM we optimized hyperparameters that specify the number of coefficients (number of features used for risk calculation), their value (small non-zero integers), and the intercept (see Fig 1). For FD-RiskSLIM we also optimized hyperparameters of feature discretization, which involved finding an

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

optimal number of bins and the position of their edges as described in section 3.2.

We used RiskSLIM implementation by Ustun and Rudin [23] available on https://github.com/ustunb/risk-slim.

4 Experiments

We conducted experiments using two real-world medical datasets to demonstrate the application of FD-RiskSLIM. In the first experiment described in section 4.1 we compare its performance to the performance of RiskSLIM. We also compared our results with results obtained by others on the same dataset. In the second experiment, we applied RiskSLIM, FD-RiskSLIM, and other ML methods that we implemented to the original dataset describing the survival of patients with kidney cancer (section 4.2).

4.1 Prediction of death of patients with heart failure

The first set of experiments involves the dataset containing the medical records (13 features) of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad in Punjab, Pakistan in 2015 [1]. We test FD-RiskSLIM on the classification task of patient death during the observation period as in [6] where authors used the same heart-failure dataset. We followed the same methodology for model training and performance evaluation as in [6]. We split the dataset into 80% for the training set and 20% for the test set. As in [6], we used the following metrics: accuracy, F1 score, Matthews correlation coefficient (MCC), and the area under the ROC curve to measure the performance of RiskSLIM and FD-RiskSLIM.

We built the following four models. We created RiskSLIM and FD-RiskSLIM models by using all the features and RiskSLIM and FD-RiskSLIM models using only 2 out of 13 features. This choice was inspired by authors of [6] who predicted the survival only from *serum creatinine* and *ejection fraction* features ignoring the other 11 features. We scaled the features using a min-max (0-1) scaler as it ensures small final values of a score calculated with a ready scoring system. The models are presented in Figure 2. Even if we passed all the features for training the output risk scoring systems A and B contain only part of them as the underlying algorithm apart from minimizing loss, it minimizes the number of outputted coefficients. The resulting risk formula differs between risk scoring systems as it also includes interception coefficient which can differ for different models.

Tables 2 and 3 show performance metrics. The results for models other than RiskSLIM and FD-RiskSLIM come from [6]. We followed the same methodology of performance evaluation. We assumed 50% of the risk threshold for positive evaluation to compute the accuracy. FD-RiskSLIM performed the best by far.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

Features	Points
serum creatinine	6
age	4
sex	-1
ejection fraction	-5
Risk formula: $\frac{1}{1+\exp(1-score)}$	**A**

Features	Points
age between **63.65 (inc.)** and **95 (inc.)**	2
ejection fraction between **0 (inc.)** and **20**	2
serum creatinine between **1 (inc.)** and **9.4 (inc.)**	1
serum creatinine between **0 (inc.)** and **0.5**	-1
Risk formula: $\frac{1}{1+\exp(2-score)}$	**B**

Features	Points
serum creatinine	7
ejection fraction	-5
Risk formula: $\frac{1}{1+\exp(-score)}$	**C**

Features	Points
serum creatinine between **7.7 (inc.)** and **9.4 (inc.)**	3
serum creatinine between **1 (inc.)** and **7.7**	-1
ejection fraction between **28 (inc.)** and **80 (inc.)**	-2
serum creatinine between **0 (inc.)** and **1**	-3
Risk formula: $\frac{1}{1+\exp(3-score)}$	**D**

Fig. 2. A and **B** are models trained on all features for RiskSLIM and FD-RiskSLIM, respectively. **C** and **D** are models trained on features including only serum creatinine and ejection fraction for RiskSLIM and FD-RiskSLIM, respectively. To obtain *score* we multiply the vector of *Features* by the vector of *Points*. A and C risk scoring systems contain continuous features only in the range 0-1. B and D risk scoring systems contain categorical features only with values 0 or 1.

Discussion We found that RiskSLIM models perform superior to other machine learning models. This is a surprising finding and initially, we suspected that we may simply compute the metrics differently, on a different test set, etc., compared to [6]. However, in the second experiment (section 4.2) we found similar superiority when we compared RiskSLIM with ML models that we implemented and optimized on identical test sets.

One of the greatest advantages of scoring systems is their clarity and interpretability. Only a small subset of features and simple formula are needed to quickly evaluate risk even using a piece of paper and a pen. This simplicity and interpretability cannot be achieved for models like Random Forest, Gradient Boosting, or Support Vector Machine. These scoring systems also directly expose feature importances. Essentially, deciding with the use of this system, one knows which feature contributes to the risk and by how much. The highest the absolute number of points assigned by a model to a given feature the biggest impact it has on the final risk.

When we consider interpretable ML models the Decision Tree model naturally comes to mind. We present the example in Figure 3. It consists of four leaf nodes and it is simple and easy to interpret due to the shallow depth of the tree. However, the performance (averaged) of this model is substantially worse than RiskSLIM (accuracy: 0.735, F1: 0.532, ROC AUC: 0.675, MCC: 0.372). Another disadvantage of the Decision Tree model is that it forces conditions to be checked in a given order, while RiskSLIM allows to evaluate them in any order.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

Table 2. Comparison of performances of the models trained on all the features. Results are sorted by MCC. All results except for RiskSLIM comes from [6] and represent means computed over the test set.

Method	MCC	F1 score	Accuracy	ROC AUC
FD-RiskSLIM	0.436	0.617	0.744	0.723
RiskSLIM	0.392	0.529	0.744	0.672
Random forests	0.384	0.547	0.740	0.800
Decision tree	0.376	0.554	0.737	0.681
Gradient boosting	0.367	0.527	0.738	0.754
Linear regression	0.332	0.475	0.730	0.643
One rule	0.319	0.465	0.729	0.637
Artificial neural network	0.262	0.483	0.680	0.559
Naive bayes	0.224	0.364	0.696	0.589
SVM radial	0.159	0.182	0.690	0.749
SVM linear	0.107	0.115	0.684	0.754
KNN	-0.025	0.148	0.624	0.493

Table 3. Comparison of performances of the models trained on serum creatinine and ejection fraction features only. Results are sorted by MCC. All results except for RiskSLIM and FD-RiskSLIM are reproduced from [6] and represent means computed over the test set.

Method	MCC	F1 score	Accuracy	ROC AUC
FD-RiskSLIM	0.435	0.616	0.746	0.719
Random forests	0.418	0.585	0.754	0.698
Gradient boosting	0.414	0.585	0.750	0.792
RiskSLIM	0.380	0.476	0.750	0.649
SVM radial	0.348	0.543	0.720	0.667

4.2 Prediction of outcomes of kidney cancer treatment

We also performed experiments on two datasets of medical records of patients with kidney cancer (metastatic renal-cell carcinoma) related to treatment using sunitinib and everolimus drugs. The datasets come from the National Institute of Oncology in Warsaw, Poland [9]. Most of the features in the datasets are binary or categorical (tumor grading, metastases, and other diseases), but there are also some numerical features such as age, weight, BMI, lymphocytes, leukocytes, and neutrophils. There are two outcome variables: (1) progression-free survival (PFS) time which is the length of time during and after the treatment of cancer, that a patient lives with cancer but it does not get worse, given in years, and overall survival (OS) time which is the length of time from the beginning of treatment to death, given in years. The data are right-censored but in this work, we ignored this as our main goal was to demonstrate the application of RiskSLIM algorithm.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

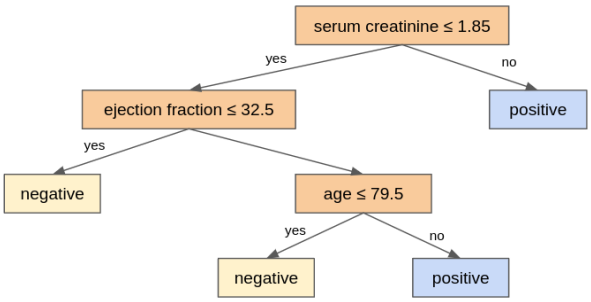

Fig. 3. Decision Tree model trained on all the features.

We selected two classification tasks to compare the performance of explainable methods (RiskSLIM and FD-RiskSLIM) with other state-of-the-art ML methods. We predict whether PFS and OS times are longer than their respective medians.

We transformed the original datasets by dropping irrelevant features and features with a substantial ratio of missing values (over 20%). We concatenated these two datasets to increase the total number of records. Only common features were kept in the final dataset. An additional categorical feature denoting the type of the treatment was added. The final dataset consisted of 149 events (rows) with 50 features.

For comparison, we also applied the following classification models: K-Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine with the radial kernel, and Support Vector Machine with linear kernel. We divided the dataset into subsets for training and testing, in proportions 80:20, respectively. Splits were repeated 100 times and results are provided as averages.

Figure 4 shows scoring systems obtained for both classification tasks. Most of the features from the original dataset are categorical or binary therefore the discretization was applied only to four of them (BMI, lymphocytes, leukocytes, and neutrophils). All the numerical features were scaled with the use of a min-max (0-1) scaler. Scoring systems denoted as **A** and **B** allow calculating the risk of OS being longer than the median. Scoring systems denoted as **C** and **D** are classification models for PFS longer than the median.

Tables 4 and 5 present comparison of the performance of the different classification models. As for experiments presented in the section 4.1 for RiskSLIM models we assumed that the risk greater or equal to 50% evaluates to the positive class. For models listed in both figures we performed hyperparameter optimization using Optuna implemented as described in section 3.

Discussion For both classification tasks, RiskSLIM performs as well as the state-of-the-art models like Gradient Boosting or Random Forest. FD-RiskSLIM performs better which is also consistent with findings of the experiment with the heart failure described in section 4.1.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

Features	Points
G1	2
NEUT > UT	-1
Heng1	-4
neutrophils	-5
MSKCC2	-5
Risk formula: $\frac{1}{1+\exp(-6-score)}$	**A**

Features	Points
lymphocytes between **2.13 (inc.)** and **5.21 (inc.)**	-1
NEUT > UT	-2
G3	-2
Heng1	-3
MSKCC2	-5
Risk formula: $\frac{1}{1+\exp(-5-score)}$	**B**

Features	Points
distant lymph nodes	1
number of other cancers	-1
AH	-1
T2	-1
NEUT > UT	-2
G3	-2
LDH > 1.5xUT	-7
Risk formula: $\frac{1}{1+\exp(-9-score)}$	**C**

Features	Points
leukocytes between **3.21 (inc.)** and **4.49 (inc.)**	1
AH	-1
HGB < LT	-1
LDH > 1.5xUT	-1
NEUT > UT	-1
G3	-1
Risk formula: $\frac{1}{1+\exp(-3-score)}$	**D**

Fig. 4. A and **B** are scoring models of risk for OS longer than median for RiskSLIM and FD-RiskSLIM. **C** and **D** are scoring models PFS longer than median for RiskSLIM and FD-RiskSLIM. To obtain the *score* we multiply the vector of *Features* by the vector of *Points*. Explanations for feature names are provided in the Appendix.

Most of the points present in risk scoring systems shown in Figure 4 are negative. This indicates that they decrease the chance that someone would live without progression or overall longer than the median time. Interestingly, RiskSLIM/FD-RiskSLIM use G1 and G3 features (tumor grades 1 and 3) but do not use G2 and G4. We speculate that if G1 increases the chance and G3 decreases then G2 might have no contribution (e.g. coefficient equal to 0). Also, class imbalance may play a role as, for example, only less than 5% of the samples in the dataset have cancer grade 4 (G4).

As for previous experiments in the section 4.1, apart from predictive accuracy, we are also interested in knowledge representation. RiskSLIM scoring systems have feature importance and their strength are provided explicitly. The evaluation of risk for new patients goes fast as there are just several coefficients that have to be summed and passed to the risk formula.

The clinical interpretation of the model is beyond the scope of this work. However, it is of great practical importance and we will pursue this direction of research in collaboration with our clinical colleagues in future work.

For comparison of explainability, we also built Decision Tree models for both classification tasks. They are presented in Figure 5 and Figure 6. These trees are not sparse and easy to work with due to a small number of rules. However, their performance is worse and they can not evaluate the risk, they allow for classification only.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

Table 4. Comparison of performances of the models built for predicting whether overall survival time is longer than a median. Results are sorted by MCC.

Method	MCC	F1 score	Accuracy	ROC AUC
FD-RiskSLIM	0.225 (0.149)	0.744 (0.059)	0.648 (0.069)	0.594 (0.066)
RiskSLIM	0.224 (0.144)	0.726 (0.064)	0.638 (0.073)	0.601 (0.065)
Gradient boosting	0.211 (0.135)	0.715 (0.057)	0.632 (0.063)	0.598 (0.065)
Random forests	0.181 (0.148)	0.739 (0.058)	0.629 (0.074)	0.571 (0.063)
Decision tree	0.152 (0.152)	0.690 (0.076)	0.606 (0.074)	0.572 (0.072)
SVM linear	0.131 (0.128)	0.672 (0.063)	0.592 (0.061)	0.564 (0.063)
KNN	0.052 (0.152)	0.688 (0.062)	0.572 (0.073)	0.523 (0.065)
SVM radial	0.000 (0.000)	0.742 (0.058)	0.597 (0.074)	0.500 (0.000)

Table 5. Comparison of performances of the models built for predicting whether progression-free survival time is longer than a median. Results are sorted by MCC.

Method	MCC	F1 score	Accuracy	ROC AUC
Random forests	0.397 (0.127)	0.707 (0.075)	0.688 (0.069)	0.696 (0.064)
FD-RiskSLIM	0.389 (0.122)	0.745 (0.053)	0.689 (0.063)	0.679 (0.073)
Gradient boosting	0.379 (0.108)	0.703 (0.064)	0.684 (0.056)	0.689 (0.054)
RiskSLIM	0.340 (0.150)	0.705 (0.070)	0.668 (0.071)	0.665 (0.074)
Decision tree	0.313 (0.136)	0.654 (0.086)	0.649 (0.069)	0.655 (0.068)
SVM linear	0.229 (0.153)	0.635 (0.079)	0.613 (0.074)	0.614 (0.077)
KNN	0.162 (0.131)	0.616 (0.067)	0.581 (0.066)	0.580 (0.065)
SVM radial	-0.007 (0.030)	0.669 (0.110)	0.519 (0.072)	0.498 (0.007)

5 Conclusions

RiskSLIM is capable of making accurate and interpretable predictions at the same time outperforming complex and non-interpretable black-box models in certain applications which we explored in this paper. Interpretability became one of the most important factors when it comes to making high stake decisions [18] e.g. in medical domains which underlines the importance of explainable models such as the ones considered here. Adding discretization makes this approach even better. The drawback of RiskSLIM is its high computational demand as model training is longer by orders of magnitude in comparison to models such as Decision Tree, Random Forest, or Gradient Boosting. Also feature discretization introduced by us here as a part of data preprocessing increases substantially model training time.

In future work, we plan to incorporate feature discretization into the model building stage to improve efficiency and reliability. Additionally, in the future work we will investigate results stability. We also plan to apply it to other domains.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

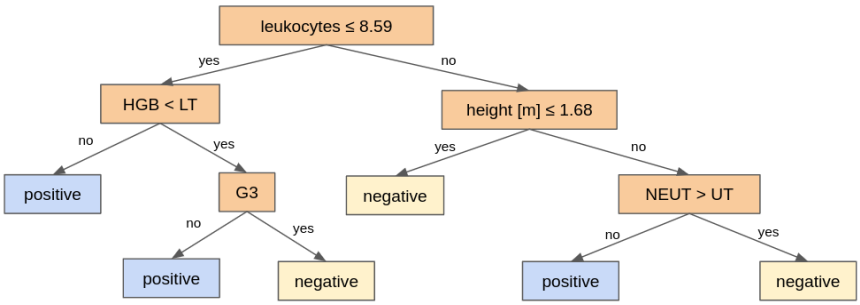

Fig. 5. Decision Tree model trained for predicting whether OS is longer than the median.

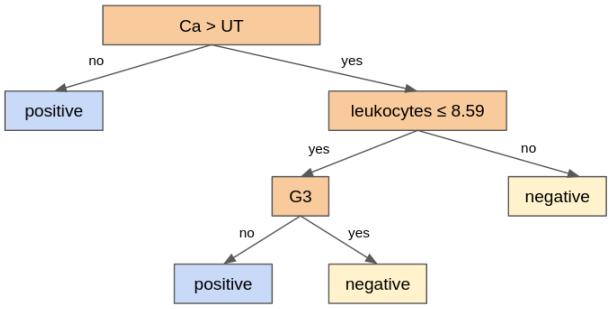

Fig. 6. Decision Tree model trained for predicting whether PFS is longer than the median.

6 Acknowledgements

This work is supported in part by the European Union's Horizon 2020 research and innovation programme under grant agreement Sano No. 857533 and the International Research Agendas programme of the Foundation for Polish Science, co-financed by the EU under the European Regional Development Fund.

References

1. Ahmad, T., Munir, A., Bhatti, S.H., Aftab, M., Raza, M.A.: Survival analysis of heart failure patients: A case study. PLoS ONE **12** (2017)
2. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2019)
3. Antman, E.M., Cohen, M., Bernink, P.J., McCabe, C., Horácek, T., Papuchis, G.C., Mautner, B., Corbalán, R., Radley, D.R., Braunwald, E.: The timi risk score for unstable angina/non-st elevation mi: A method for prognostication and therapeutic decision making. JAMA **284 7**, 835–42 (2000)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

4. Bone, R.C., Balk, R.A., Cerra, F.B., Dellinger, R.P., Fein, A.M., Knaus, W.A., Schein, R.M.H., Sibbald, W.J.: Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. the accp/sccm consensus conference committee. american college of chest physicians/society of critical care medicine. Chest **101 6**, 1644–55 (1992)
5. Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., Zhou, S.: A machine-learning-based prediction method for hypertension outcomes based on medical data. Diagnostics **9** (2019)
6. Chicco, D., Jurman, G.: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making **20** (2020)
7. Chung, F., Abdullah, H.R., Liao, P.: Stop-bang questionnaire: A practical approach to screen for obstructive sleep apnea. Chest **149**(3), 631–638 (2016). https://doi.org/10.1378/chest.15-0903
8. Cruz, J.A., Wishart, D.S.: Applications of machine learning in cancer prediction and prognosis. Cancer Informatics **2**, 59 – 77 (2007)
9. Dudek, A.Z., Żołnierek, J., Dham, A., Lindgren, B.R., Szczylik, C.: Sequential therapy with sorafenib and sunitinib in renal cell carcinoma. Cancer **115** (2009)
10. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: IJCAI (1993)
11. Gage, B.F., Waterman, A.D., Shannon, W., Boechler, M., Rich, M.W., Radford, M.J.: Validation of Clinical Classification Schemes for Predicting StrokeResults From the National Registry of Atrial Fibrillation. JAMA **285**(22), 2864–2870 (06 2001). https://doi.org/10.1001/jama.285.22.2864
12. Gage, B.F., Waterman, A.D., Shannon, W.D., Boechler, M., Rich, M.W., Radford, M.J.: Validation of clinical classification schemes for predicting stroke: Results from the national registry of atrial fibrillation. JAMA **285**, 2864–2870 (2001)
13. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation" (2016). https://doi.org/10.1609/aimag.v38i3.2741, http://arxiv.org/abs/1606.08813
14. Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., Bergner, M., Bastos, P.G., Sirio, C.A., Murphy, D.J., Lotring, T., Damiano, A.M.: The apache iii prognostic system. risk prediction of hospital mortality for critically ill hospitalized adults. Chest **100 6**, 1619–36 (1991)
15. Metnitz, P.G.H., Moreno, R., Almeida, E., Jordan, B., Bauer, P., Campos, R.A., Iapichino, G., Edbrooke, D.L., Capuzzo, M., Gall, J.R.L.: Saps 3—from evaluation of the patient to evaluation of the intensive care unit. part 1: Objectives, methods and cohort description. Intensive Care Medicine **31**, 1336 – 1344 (2005)
16. Mohan, S., Thirumalai, C., Srivastava, G.: Effective heart disease prediction using hybrid machine learning techniques. IEEE Access **7**, 81542–81554 (2019)
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
18. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019)
19. Rudin, C., Ustun, B.: Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. Interfaces **48**, 449–466 (2018)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9

20. Struck, A.F., Rodriguez-Ruiz, A.A., Osman, G., Gilmore, E.J., Haider, H.A., Dhakar, M.B., Schrettner, M., Lee, J.W., Gaspard, N., Hirsch, L.J., et al.: Comparison of machine learning models for seizure prediction in hospitalized patients (6 2019), https://onlinelibrary.wiley.com/doi/full/10.1002/acn3.50817
21. Struck, A.F., Ustun, B., Ruiz, A.A.R., Lee, J.W., Laroche, S.M., Hirsch, L.J., Gilmore, E.J., Vlachý, J., Haider, H.A., Rudin, C., Westover, M.B.: A practical risk score for eeg seizures in hospitalized patients (s11.002). Neurology **90** (2018)
22. Ustun, B., Rudin, C.: Supersparse linear integer models for optimized medical scoring systems. Machine Learning pp. 1–43 (2015). https://doi.org/10.1007/s10994-015-5528-6
23. Ustun, B., Rudin, C.: Learning Optimized Risk Scores. Journal of Machine Learning Research **20**(150), 1–75 (2019), http://jmlr.org/papers/v20/18-615.html
24. Ustun, B., Westover, M.B., Rudin, C., Bianchi, M.T.: Clinical prediction models for sleep apnea: The importance of medical history over symptoms. Journal of clinical sleep medicine : JCSM : official publication of the American Academy of Sleep Medicine **12 2**, 161–8 (2016)
25. Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. ArXiv **abs/2006.00093** (2020)
26. Wang, C.L., Han, B., Patel, B., Mohideen, F., Rudin, C.: In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. ArXiv **abs/2005.04176** (2020)

7 Appendix

Explanations for the features listed in Figure 4. **G1**: binary, from tumor grading, denotes low grade (tumor well differentiated), **NEUT > UT**: binary, 1 if concentration of neutrocytes exceeds upper threshold of a range of normal values, **Heng1**: binary, Heng scale, **neutrophils**: continuous, scaled with min-max (0-1) scaler, amount of neutrophils, **MSKCC2**: binary, MSKCC (Memorial Sloan Kettering Cancer Center) scale, **lymphocytes between 2.13 (inc.) and 5.21 (inc.)**: binary, 1 if amount of lymphocytes is inside this range, **G3**: binary, from tumor grading, denotes high grade (tumor poorly differentiated), **distant lymph nodes**: binary, 1 if tumor metastasis in the distant lymph nodes, **number of other cancers**: numeric, number of other cancers (metastasis) than: lungs, liver, bones and distant lymph nodes, **AH**: binary, denotes if someone suffers from arterial hypertension, **T2**: binary, cancer staging (T1 - T4), **LDH > 1.5xUT**: binary, lactate dehydrogenase activity, 1 if exceeds 1.5 x upper threshold of a range of normal values, **leukocytes between 3.21 (inc.) and 4.49 (inc.)**: binary, 1 if amount of leukocytes is inside this range, **HGB < LT**: binary, 1 if hemoglobin concentration exceeds lower threshold of a range of normal values, **Ca > UT**: binary, 1 if calcium concentration exceeds upper threshold of a range of normal values.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_9