UNDERGRADUATE RESEARCH EXPERIENCE (UREx)

PROJECT REPORT

Comparing Risk Scoring Models for Predicting Infections, Rebleeding and Mortality in Cirrhotic Patients

Chow Jie Seth (A0233441R)

Supervised by: Mehul Motani

Department of Electrical Engineering, College of Design and Engineering
National University of Singapore

1 Abstract

This project investigates the use of interpretable risk scoring models to predict clinical outcomes in patients with liver cirrhosis, focusing on infection within 14 days, rebleeding within 14 days, and mortality within 42 days. Using a real-world dataset, we trained three transparent scoring systems—RiskSLIM, FasterRisk, and AutoScore—and compared them to black-box models selected via PyCaret's automated machine learning. Data preprocessing included processes such as discretisation of continuous variables, encoding of categorical features. Model thresholds were optimised on the training set using the Youden Index, and performance was assessed across multiple metrics including sensitivity, specificity, precision, F1 score, and AUC. Results show that interpretable models, particularly FasterRisk, achieve strong discriminative performance while preserving transparency, making them more suitable for clinical use. In contrast, PyCaret-selected models achieved high accuracy but failed to identify positive cases, emphasising the importance of tailored and interpretable modeling in healthcare applications.

2 Acknowledgements

I would like to express my gratitude to Prof. Mehul Motani for his guidance and continuous support on this research project throughout the academic year 2024/2025. I would also like to thank Kei Sen for his advice as well.

3 Table of Contents

4 Introduction

Patients with liver cirrhosis often face serious problems like infections, rebleeding, and death. Being able to predict these events early is important for treatment and planning. Machine learning (ML) models have become popular for medical predictions because they can handle complex data and often give good results. However, they are not easy to understand and interpret, which can make doctors less confident in using them.

Risk scoring models like RiskSLIM (Ustun & Rudin, 2016), AutoScore (Xie et al., 2020), and FasterRisk (Lee & Rudin, 2022) are much easier to understand as they use simple rules and scores that doctors can interpret easily. Even though these models are simpler than ML models, they might still perform well, especially when there are more negative cases than positive ones, which is common in medical data.

This project compares risk scoring models with ML models for predicting infection within 14 days, rebleeding within 14 days, and mortality within 42 days in cirrhotic patients, a type of liver disease. The focus is on how well these models can detect positive cases while keeping false positives and false negatives low. In medical settings, false negatives (missing a true positive case) can be more dangerous than false positives, so I also looked at how these models handle that trade-off.

5 Hypothesis

I hypothesised that interpretable risk scoring models can still perform well, even though they are simpler than ML models. They can detect most positive cases and give useful predictions without being too complex. While ML models may potentially have slightly better performance, risk scoring models might be good enough—especially since they are easier to understand and use in practice.

6 Methodology

This project compares three interpretable risk scoring models — RiskSLIM, AutoScore, and FasterRisk — with various ML models from PyCaret (PyCaret, n.d), across three clinical prediction tasks: infection, rebleeding, and 42-day mortality in cirrhotic patients.

6.1 Risk Score Models

Each risk scoring model produces a simple point-based score table that adds up individual points for selected features. To illustrate their output, an example score table for **predicting infection within 14 days** is shown beneath each corresponding model subsection below. Only features that received non-zero scores are included in these illustrations, as features with a score of zero do not contribute to the final risk calculation. Score tables for the other two prediction tasks—rebleeding within 14 days and mortality within 42 days—are provided in the Appendix for reference.

6.1.1 RiskSlim

RiskSLIM (Fig. 1) uses integer linear programming to create a sparse linear model with small integer coefficients (e.g. -5 to +5). It directly optimises for model sparsity and accuracy, making it highly interpretable and constrained by design.

==========	==========	=========
Blood_transfused_in_48_hours_u__bin_0	5 points	+ .....
Blood_transfused_in_48_hours_u__bin_1	5 points	+ .....
Hemoglobin_g_L__bin_0	5 points	+ .....
Hemoglobin_g_L__bin_1	5 points	+ .....
Bilirubin_mg_dL__bin_1	5 points	+ .....
Hospitalization_day__bin_1	5 points	+ .....
Bilirubin_mg_dL__bin_0	5 points	+ .....
Treatment_APC	3 points	+ .....
ICU_admission	2 points	+ .....
WBC_x10_3_uL__bin_0	1 points	+ .....
Sex	1 points	+ .....
Platelet_count_x10_3_uL__bin_0	1 points	+ .....
Albumin_g_dL__bin_0	1 points	+ .....
Creatinine_mg_L__bin_1	1 points	+ .....
Systolic_blood_pressure_mmHg__bin_0	1 points	+ .....
Antibiotic_prophylaxis	1 points	+ .....
Etiology_of_cirrhosis_HCV	1 points	+ .....
Na_mEq_L__bin_1	-1 points	+ .....
Etiology_of_cirrhosis_BC	-1 points	+ .....
Etiology_of_bleeding_peptic_ulcer	-2 points	+ .....
==========	==========	=========
ADD POINTS FROM ROWS 1 to 20	SCORE	= .....

Figure 1. RiskSlim Scoring Table for Predicting Infection Within 14 Days

6.1.2 AutoScore

AutoScore (Fig. 2) is a modular framework that ranks discretised variables using importance scores (e.g. from random forests), and then maps them to point values using logistic regression. It uses a stepwise selection process to build a score table in a data-driven but interpretable way. In the AutoScore output, features include suffixes such as "0", "1", "not_0", and "not_1". These denote logical conditions: "0" and "not_1" typically indicate the absence of a condition, while "1" and "not_0" indicate its presence. These suffixes result from how categorical variables are encoded and can be interpreted similarly based on context.

	Variable_Category	Score
0	HCCnot_0	1
1	ICU_admissionnot_0	12
2	INR_bin_0not_0	1
3	Sex1	4
4	Heart_rate_beats_min__bin_11	2
5	Etiology_of_cirrhosis_HCVnot_0	4
6	Bilirubin_mg_dL__bin_11	1
7	Na_mEq_L__bin_1not_1	6
8	Ascitesnot_0	2
9	INR_bin_1not_1	9
10	Heart_rate_beats_min__bin_0not_0	9
11	Antibiotic_prophylaxisnot_0	8
12	Systolic_blood_pressure_mmHg__bin_01	3
13	Bilirubin_mg_dL__bin_0not_0	9
14	Treatment_APCnot_0	8
15	ALT_IU_L__bin_1not_1	1
16	Platelet_count_x10_3_uL__bin_10	8
17	Albumin_g_dL__bin_1not_1	8
18	Hemoglobin_g_L__bin_11	1
19	Age_bin_1not_0	3

Figure 2. AutoScore Scoring Table for Predicting Infection Within 14 Days

6.1.3 FasterRisk

FasterRisk (Fig. 3) is a greedy algorithm that builds a scoring model by selecting features and assigning integer weights to minimise a loss function (typically logistic loss). It's faster than RiskSLIM but doesn't guarantee global optimality.

	Feature	Score Contribution
0	Antibiotic_prophylaxis	1.0
1	Sex	1.0
2	ICU_admission	1.0
3	Platelet_count_x10_3_uL__bin_1	-1.0
4	WBC_x10_3_uL__bin_1	-1.0
5	Na_mEq_L__bin_1	-1.0
6	Creatinine_mg_L__bin_0	-1.0
7	Bilirubin_mg_dL__bin_1	1.0
8	Albumin_g_dL__bin_1	-1.0
9	Hospitalization_day__bin_0	-5.0
10	Etiology_of_cirrhosis_BC	-2.0
11	Etiology_of_cirrhosis_NBNC	-1.0
12	Etiology_of_bleeding_portal_hypertension	2.0
13	Treatment_APC	3.0
14	Treatment_EIS	1.0

Figure 3. FasterRisk Scoring Table for Predicting Infection Within 14 Days

After generating the score tables, I applied each model to the training set to compute total scores for each patient. I then iterated through all possible total integer scores (from minimum to maximum) to find the best threshold with the highest Youden Index. The Youden Index is a common metric used in evaluating binary classification models, defined as sensitivity + specificity - 1. It ranges from -1 to 1, where 1 indicates perfect classification, 0 indicates performance no better than chance, and negative values suggest worse-than-random classification. The threshold that maximised this index on the training set was then applied to the test set to generate binary predictions.

From these predictions, I computed the confusion matrix (TP, FP, TN, FN) and derived the following metrics in Table 1 below:

Metric	Description
Accuracy	Overall proportion of correct predictions
Sensitivity (Recall)	Proportion of actual positives correctly predicted
Specificity	Proportion of actual negatives correctly predicted
Precision	Proportion of actual negatives correctly predicted
Negative Predictive Value (NPV)	Proportion of predicted negatives that are actually negative
F1 Score	Harmonic mean of precision and sensitivity
Youden Index	Sensitivity + Specificity - 1
AUC (Area Under the ROC Curve)	Computed using sklearn's roc_auc_score function

Table 1. Metrics Derived From Confusion Matrix and Used For Model Evaluation

6.2 PyCaret AutoML Models

To benchmark performance, I used PyCaret's classification module, which automates the training, tuning, and evaluation of multiple ML models such as logistic regression, decision trees, random forests, k-nearest neighbors, etc. It also handles model comparison, and selection of the best-performing model. The goal of including PyCaret is to provide a reference for how well modern ML models can perform on this task. If even these models struggle on the task (e.g. low F1 scores or AUC), it suggests the

classification problem is inherently difficult. This context helps us evaluate whether a risk scoring model is still a reasonable choice due to its interpretability and simplicity, even if it had a lower metric than the machine learning models on the prediction tasks.

7 Data Preprocessing and Cleaning

Before modeling, the dataset was cleaned and prepared to ensure it was suitable for both interpretable risk score models and standard machine learning models. Some basic fixes were applied, such as correcting a typo in the column name (“releeding” to “rebleeding”) and cleaning column names to remove special characters or spaces. Additionally, certain binary features (e.g., Sex, HCC, Antibiotic prophylaxis) were mapped to 0/1 values, and categorical variables (e.g., Etiology of cirrhosis, Etiology of bleeding, and Treatment) were one-hot encoded. The following subsections 7.1 to 7.3 are the justifications for the other preprocessing steps that I have conducted on the dataset before modelling.

7.1. Dropping Irrelevant Columns and Missing Values

Irrelevant columns such as Date, Patient No., Child-Pugh score, and MELD were removed. Date and Patient No. were removed as they are identifiers not useful for prediction. Child-Pugh score and MELD were dropped to avoid redundancy, since they are composite scores already calculated from other features like INR, bilirubin, and creatinine. Lastly, columns with missing values were dropped entirely to avoid inconsistencies during training, and entries containing "." were treated as missing values and thus removed from the dataset.

7.2. Discretisation and One-Hot Encoding of Continuous Features

A selection of continuous clinical variables—such as age, platelet count, WBC, hemoglobin, and bilirubin—were discretised into two bins using quantile-based binning. I’ve limited the number of bins to 2 to simplify the input space and make the features compatible with interpretable models like RiskSLIM. After binning, the features were

one-hot encoded to convert them into binary variables. This makes it easier for risk scoring models to assign clear, interpretable weights to each feature group.

7.3. Train-Test Splitting and Target Label Conversion

For each prediction task—infection, rebleeding, and mortality—a separate binary classification was performed. A stratified 70/30 train-test split was used to preserve the class balance across splits. The original target values ("yes"/"no") were mapped to binary numeric labels: 1 for positive cases and -1 for negative cases. All input features were also converted to floating-point numbers to ensure numerical consistency.

8 Experiment results

The three tables in this section summarise the performance of three interpretable risk scoring models (RiskSLIM, FasterRisk, AutoScore) and one PyCaret-selected machine learning model (which varies by task) across three prediction targets: infection within 14 days, rebleeding within 14 days, and mortality within 42 days. The models are evaluated using a comprehensive set of performance metrics derived from the confusion matrix—such as sensitivity, specificity, precision, negative predictive value (NPV), F1 score—and threshold-independent metrics such as AUC. Thresholds for the risk scoring models were selected by maximising the Youden Index on the training set, as explained in Section 6. AUC was computed using sklearn.metrics.roc_auc_score based on each model's raw prediction scores.

Across all tasks, the PyCaret models display very high specificity and accuracy but fail to identify any true positive cases (sensitivity = 0.000). As a result, their precision is undefined (NA) because precision is calculated as $\frac{TP}{TP+FP}$, and both the numerator and denominator are zero when the model predicts no positive cases at all. In practical terms, this indicates that the PyCaret models default to predicting only the negative class, likely due to class imbalance, offering no clinical utility for identifying at-risk patients.

In contrast, all three risk scoring models—RiskSLIM, FasterRisk, and AutoScore—demonstrate the ability to make meaningful trade-offs between sensitivity and specificity. This enables them to detect at least some true positives, which is critical in clinical applications where false negatives may result in missed diagnoses and worse outcomes.

8.1 Predicting Infection Within 14 Days

Model	**Threshold**	**Youden Index**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**NPV**	**F1**	**AUC**
RiskSlim	-3	0.112	0.779	0.308	0.804	0.078	0.955	0.499	0.669
FasterRisk	-2	0.214	0.806	0.385	0.829	0.109	0.961	0.530	0.684
AutoScore	55	0.097	0.834	0.231	0.867	0.086	0.954	0.517	0.592
PyCaret (K Neighbors Classifier)	NA	0	0.949	0.000	1.000	NA	0.949	0.487	0.345

Table 2. Experiment Results for Predicting Infection Within 14 Days

In this prediction task, the FasterRisk model outperforms all others in terms of overall discrimination, achieving the highest Youden Index (0.214), F1 score (0.530), and AUC (0.684). These results indicate that FasterRisk is better able to balance sensitivity (0.385) and specificity (0.829), effectively identifying patients at risk of developing infection without generating excessive false positives.

RiskSLIM achieves a slightly lower AUC (0.669) and F1 (0.499), with a slightly worse sensitivity (0.308). Its threshold of -3 provides a reasonable balance, though its

precision (0.078) is the lowest among the risk scoring models. This suggests RiskSLIM identifies more false positives compared to FasterRisk.

AutoScore shows the lowest Youden Index (0.097) and sensitivity (0.231) among the three, although it still maintains high specificity (0.867). The relatively low AUC (0.592) reflects weaker ranking performance in differentiating between positive and negative cases.

The PyCaret model, which selected a K-Nearest Neighbors classifier for this task, demonstrates zero sensitivity—failing to predict any positive infection cases. Despite the high accuracy (0.949) and specificity (1.000), the Youden Index is 0, and the AUC is only 0.345, indicating poor discriminatory ability. This shows that the model overfits to the majority class due to class imbalance, making it clinically non-actionable despite high nominal accuracy.

8.2 Predicting Rebleeding Within 14 Days

Model	**Threshold**	**Youden Index**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**NPV**	**F1**	**AUC**
RiskSlim	-2	0.033	0.680	0.318	0.714	0.096	0.917	0.475	0.500
FasterRisk	-2	0.433	0.697	0.740	0.693	0.183	0.966	0.551	0.766
AutoScore	59	0.039	0.648	0.364	0.675	0.096	0.918	0.465	0.519
PyCaret (Random Forest Classifier)	NA	0	0.913	0.000	1.000	NA	0.913	0.477	0.454

Table 3. Experiment Results for Predicting Rebleeding Within 14 Days

Next, in the second prediction task, rebleeding within 14 days, the results again highlight the strengths and limitations of each model when faced with imbalanced data. While PyCaret's Random Forest classifier achieved the highest accuracy (0.913) and perfect specificity (1.000), it failed to identify any true positive cases, as indicated by a sensitivity of 0.000 and Youden Index of 0.0. This suggests the model defaulted to always predicting the majority class (no rebleeding), resulting in no practical utility for this clinical task.

In contrast, the FasterRisk score-based model showed the most balanced performance, with a Youden Index of 0.433, sensitivity of 0.740, and specificity of 0.693. Despite a relatively low precision of 0.183 due to the imbalance, its AUC (0.766) and F1 score (0.551) were the highest among all models, indicating that it was the most effective at distinguishing rebleeding from non-rebleeding cases.

RiskSLIM and AutoScore both struggled with sensitivity, achieving 0.318 and 0.364 respectively, which means many positive cases were missed. Although their specificity values were decent (above 0.67), their Youden Index values remained low (~0.03–0.04), and their F1 scores were below 0.48. These scores suggest that while they correctly classify the majority class well, they do not effectively detect rebleeding cases.

Overall, this experiment shows that while the ML model appears accurate at first glance, it is essentially non-functional due to zero sensitivity. The scoring models—especially FasterRisk—offer more clinically useful trade-offs by identifying a non-trivial portion of positive cases.

8.3 Predicting Mortality Within 42 Days

Model	Threshold	Youden Index	Accuracy	Sensitivity	Specificity	Precision	NPV	F1	AUC
RiskSlim	-3	0.375	0.810	0.556	0.820	0.102	0.980	0.533	0.803
FasterRisk	-4	0.646	0.752	0.900	0.746	0.111	0.995	0.525	0.865
AutoScore	55	0.066	0.822	0.222	0.844	0.050	0.967	0.492	0.729
PyCaret (Logistic Regression)	NA	0	0.964	0.000	1.000	NA	0.964	0.491	0.226

Table 4. Experiment Results for Predicting Mortality Within 42 Days

Lastly, in the mortality prediction task, FasterRisk also achieved the highest overall performance among the risk scoring models, with a Youden Index of 0.646, sensitivity of 0.900, specificity of 0.746, and AUC of 0.865. This strong trade-off between sensitivity and specificity indicates that FasterRisk effectively distinguishes between patients at risk of 42-day mortality and those who are not. Its relatively high sensitivity is especially valuable in a clinical setting, where identifying high-risk patients is critical for early intervention.

RiskSLIM also showed promising performance, with an AUC of 0.803 and a Youden Index of 0.375. It achieved moderate sensitivity (0.556) and specificity (0.820), indicating a more conservative model that slightly favors false negatives over false positives compared to FasterRisk. This could be suitable in settings where false alarms are particularly costly, although it sacrifices some recall in detecting true positive mortality cases.

AutoScore performed worse than both RiskSLIM and FasterRisk. It produced a low Youden Index (0.066) and sensitivity (0.222), while still maintaining high specificity (0.844). This behavior is consistent with earlier tables, where AutoScore models tend to lean toward conservative predictions, possibly due to coarser score bins or simpler feature interactions. Its AUC of 0.729 further confirms a relatively weaker ability to separate the classes compared to the other two interpretable models.

PyCaret’s selected model for this task, Logistic Regression, again defaulted to a trivial classifier that predicts all cases as negative. This resulted in a sensitivity of 0.000, a specificity of 1.000, and a Youden Index of 0, as seen in the previous tasks. Precision is marked as NA due to the lack of predicted positive cases, which makes the precision calculation undefined. Although the accuracy is high (0.964), it is misleading due to severe class imbalance and the model's failure to identify any true positive cases. Its AUC of 0.226—well below chance—further highlights the model's ineffectiveness for this particular task, possibly due to poor calibration or overfitting during automatic model selection.

Overall, the mortality task illustrates a clearer performance gap between interpretable scoring models and standard ML pipelines. FasterRisk stands out as the most effective method, combining high recall and strong discriminative performance, while PyCaret's model underperforms significantly. This supports the value of tailored risk scores in clinical prediction, especially when they are optimised with domain-specific considerations in mind.

9 Conclusion and Future work

In conclusion, this study demonstrates that interpretable scoring models can be highly effective tools for predicting key clinical events in patients with liver cirrhosis. Among the models evaluated, FasterRisk consistently offered the best balance between sensitivity and specificity, along with the highest AUC across all tasks, suggesting its robustness in identifying high-risk patients. RiskSLIM also performed well, particularly in settings favoring higher specificity, while AutoScore offered simpler yet moderately effective risk stratification. On the other hand, the PyCaret-selected models, despite achieving high accuracy and specificity, systematically failed to detect any positive cases due to extreme class imbalance. These findings reinforce the necessity of both interpretability and careful threshold calibration in clinical risk prediction. Future work may explore hybrid models that integrate domain knowledge into interpretable architectures, or leverage interpretable machine learning models such as KAN (Liu et al., 2023), while validating performance across larger and more heterogeneous patient populations.

11 Appendix

A1. Score Tables using RiskSlim

Blood_transfused_in_48_hours_u__bin_1	5 points	+ .....
Hospitalization_day__bin_1	5 points	+ .....
Hospitalization_day__bin_0	5 points	+ .....
Albumin_g_dL__bin_0	5 points	+ .....
Etiology_of_bleeding_portal_hypertension	5 points	+ .....
Etiology_of_bleeding_peptic_ulcer	4 points	+ .....
Age_bin_0	4 points	+ .....
Age_bin_1	4 points	+ .....
Albumin_g_dL__bin_1	3 points	+ .....
Blood_transfused_in_48_hours_u__bin_0	3 points	+ .....
ICU_admission	1 points	+ .....
Creatinine_mg_L__bin_1	1 points	+ .....
Creatinine_mg_L__bin_0	1 points	+ .....
Etiology_of_cirrhosis_BC	1 points	+ .....
Platelet_count_x10_3_uL__bin_1	1 points	+ .....
Hemoglobin_g_L__bin_0	-1 points	+ .....
Treatment_EVL	-1 points	+ .....
Treatment_APC	-1 points	+ .....
Prior_SBP	-5 points	+ .....
ADD POINTS FROM ROWS 1 to 19	SCORE	= .....

Na_mEq_L__bin_1	5 points	+ .....
Na_mEq_L__bin_0	5 points	+ .....
INR_bin_1	5 points	+ .....
INR_bin_0	5 points	+ .....
Systolic_blood_pressure_mmHg__bin_0	5 points	+ .....
Albumin_g_dL__bin_0	5 points	+ .....
ALT_IU_L__bin_1	5 points	+ .....
ALT_IU_L__bin_0	5 points	+ .....
Heart_rate_beats_min__bin_1	5 points	+ .....
Heart_rate_beats_min__bin_0	5 points	+ .....
Etiology_of_bleeding_peptic_ulcer	5 points	+ .....
Albumin_g_dL__bin_1	4 points	+ .....
Etiology_of_bleeding_portal_hypertension	4 points	+ .....
Systolic_blood_pressure_mmHg__bin_1	4 points	+ .....
HCC	2 points	+ .....
ICU_admission	1 points	+ .....
Hospitalization_day__bin_1	1 points	+ .....
Hemoglobin_g_L__bin_0	1 points	+ .....
Platelet_count_x10_3_uL__bin_0	-1 points	+ .....
Treatment_APC	-2 points	+ .....
ADD POINTS FROM ROWS 1 to 20	SCORE	= .....

A1. Scoring Tables for Predicting Rebleeding Within 14 Days and Mortality Within 42 Days Respectively

A2. Score Tables using AutoScore

	Variable_Category	Score
0	Ascitesnot_0	4
1	HCCnot_0	3
2	Albumin_g_dL__bin_1not_1	13
3	Platelet_count_x10_3_uL__bin_0not_1	6
4	Etiology_of_cirrhosis_HCVnot_0	1
5	Creatinine_mg_L__bin_0not_0	4
6	Albumin_g_dL__bin_0not_0	9
7	ALT_IU_L__bin_10	1
8	ALT_IU_L__bin_0not_1	9
9	Treatment_ElSnot_0	7
10	Treatment_no_treatmentnot_0	5
11	Na_mEq_L__bin_01	1
12	Platelet_count_x10_3_uL__bin_1not_0	9
13	Creatinine_mg_L__bin_1not_1	9
14	Heart_rate_beats_min__bin_01	2
15	Sex1	6
16	Bilirubin_mg_dL__bin_0not_0	9
17	INR_bin_1not_1	2

	Variable_Category	Score
0	HCCnot_0	11
1	Ascitesnot_0	3
2	Heart_rate_beats_min__bin_1not_0	5
3	Platelet_count_x10_3_uL__bin_00	7
4	Creatinine_mg_L__bin_1not_1	1
5	Systolic_blood_pressure_mmHg__bin_0not_0	5
6	Albumin_g_dL__bin_0not_0	10
7	Etiology_of_cirrhosis_BCnot_0	4
8	Age_bin_00	2
9	Creatinine_mg_L__bin_0not_0	7
10	Bilirubin_mg_dL__bin_00	2
11	ALT_IU_L__bin_00	1
12	ALT_IU_L__bin_1not_1	7
13	WBC_x10_3_uL__bin_1not_1	5
14	Age_bin_1not_1	7
15	Antibiotic_prophylaxisnot_0	6
16	Platelet_count_x10_3_uL__bin_1not_1	7
17	WBC_x10_3_uL__bin_0not_0	7
18	Na_mEq_L__bin_01	2

A2. Scoring Tables for Predicting Rebleeding Within 14 Days and Mortality Within 42 Days Respectively

A3. Score Tables using FasterRisk

	Feature	Score Contribution
0	Antibiotic_prophylaxis	-1.0
1	Prior_SBP	-5.0
2	ICU_admission	1.0
3	Blood_transfused_in_48_hours_u__bin_1	2.0
4	Platelet_count_x10_3_uL__bin_1	1.0
5	Hemoglobin_g_L__bin_1	1.0
6	Albumin_g_dL__bin_0	1.0
7	Etiology_of_bleeding_portal_hypertension	1.0
8	Treatment_APC	-1.0
9	Treatment_EVL	-1.0

	Feature	Score Contribution
0	Antibiotic_prophylaxis	1.0
1	HCC	2.0
2	Hepatic_encephalopathy	1.0
3	Prior_SBP	-5.0
4	ICU_admission	1.0
5	Platelet_count_x10_3_uL__bin_0	-1.0
6	WBC_x10_3_uL__bin_1	-1.0
7	Hemoglobin_g_L__bin_1	-1.0
8	Na_mEq_L__bin_0	1.0
9	Albumin_g_dL__bin_0	2.0
10	Systolic_blood_pressure_mmHg__bin_0	1.0
11	Heart_rate_beats_min__bin_0	-1.0
12	Hospitalization_day__bin_1	2.0
13	Etiology_of_cirrhosis_BC	1.0
14	Etiology_of_cirrhosis_NBNC	1.0
15	Etiology_of_bleeding_portal_hypertension	-2.0
16	Treatment_APC	-2.0
17	Treatment_EVL	1.0

A3. Scoring Tables for Predicting Rebleeding Within 14 Days and Mortality Within 42 Days Respectively

12 References

Ali, M. (2020). PyCaret: An open-source, low-code machine learning library in Python [Computer software]. Retrieved from https://www.pycaret.org/

Lee, J., & Rudin, C. (2022). FasterRisk: Scalable risk score learning via approximate partitioning. arXiv preprint arXiv:2210.05846. https://arxiv.org/pdf/2210.05846

Liu, S., Shen, Z., Dai, B., & Tan, M. (2023). KAN: Kolmogorov–Arnold Networks. arXiv preprint arXiv:2305.15034. https://arxiv.org/abs/2305.15034

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. arXiv preprint arXiv:1610.00168. https://arxiv.org/pdf/1610.00168

Xie, F., Chakraborty, B., Ong, M. E. H., Goldstein, B. A., & Liu, N. (2020). AutoScore: A machine learning–based automatic clinical score generator and its application to mortality prediction using electronic health records. JMIR Medical Informatics, 8(10), e21798. https://medinform.jmir.org/2020/10/e21798/PDF