On the Use of an Automated, Reproducible Binning Approach to Bring Consistency in Calibration of Predictive Models built on Electronic Health Records

Madhusree Chowdhury and Richard Jordan, Health Data Analytics Institute

Which Model is Better?

Model A: 80% accurate with 0.81 confident on the prediction it makes

Model B: 80% accurate with 0.98 confident on the prediction it makes

Why is Calibration Important?

In risk prediction models, calibration is important:

- It eliminates risk of misleading clinical decisions
- It improves our models by reducing mistakes with high probabilities

A model is perfectly calibrated when $p * 100\%$ of patients with predicted probability p of experiencing the adverse event in question actually do experience the event.

Problem Statement

Can we use an automated, reproducible binning approach to bring statistical consistency in calibration and the assessment of risk prediction models in a clinical setting?

Agenda

- Brier Score Definition
- Pain points in the historical method of assessing calibration
- Solution to the pain points
- Assess the effectiveness of CORP and compare the calibration and discrimination of several machine learning methods for predicting three health outcomes of interest: sepsis, mortality and respiratory failure

Brier Score

$$B = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

N is the sample size, o_i is the binary outcome and $p_i = Prob(o_i = 1)$ is the predicted probability.

Brier Score Decomposition

$$B = \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k (p_k - o_k)^2}_{\text{Reliability}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} n_k (o_k - \bar{o})^2}_{\text{Resolution}} + \underbrace{\bar{o}(1-\bar{o})}_{\text{Uncertainty}}$$

- The predicted probabilities have been discretized into a number K of bins
- n_k is the number of data points in bin k
- p_k is the average predicted probability in the bin
- o_k the average outcome in the bin
- $\bar{o}$ is the average outcome over the entire population (incidence)

Another useful metric is the **Skill Score**

$$Skill = 1 - \frac{B_{model}}{B_{random}} = \frac{Resolution - Reliability}{Uncertainty}$$

Description of the Components

- **Reliability** (measure of **Miscalibration**) : For a perfectly reliable model ($p_k = o_k$ for all k) it is 0. The smaller the Reliability, the better

- **Resolution** (measure of **Discrimination**) : Measures the distance between incidence and model predictions. It tells how well a model can separate classes, so the larger it is, the better

- **Uncertainty**: The variance in the observations/outcomes. It is a characteristic of the data and is independent of the model being used to predict outcomes

Pain Points in Existing Method:

- How many bins to choose for plotting the reliability diagrams?
- Upon changing the width and the population of bins, the appearance of the calibration plots change along with the metrics of miscalibration and discrimination
- The classical counting and binning approach relies on a manual, ad-hoc way of choosing the bin size/number of bins. This leads to lack of stability in the Brier Score decomposition metrics, particularly in the miscalibration/reliability and reproducibility of the reliability diagrams
- This instability can reduce a clinician's confidence in a model and impede the adoption of the model in a clinical setting

CORP Approach – Way To Address the Pain Points

RESEARCH ARTICLE | PHYSICAL SCIENCES

Stable reliability diagrams for probabilistic classifiers

Timo Dimitriadis, Tilmann Gneiting, and Alexander I. Jordan Authors Info & Affiliations

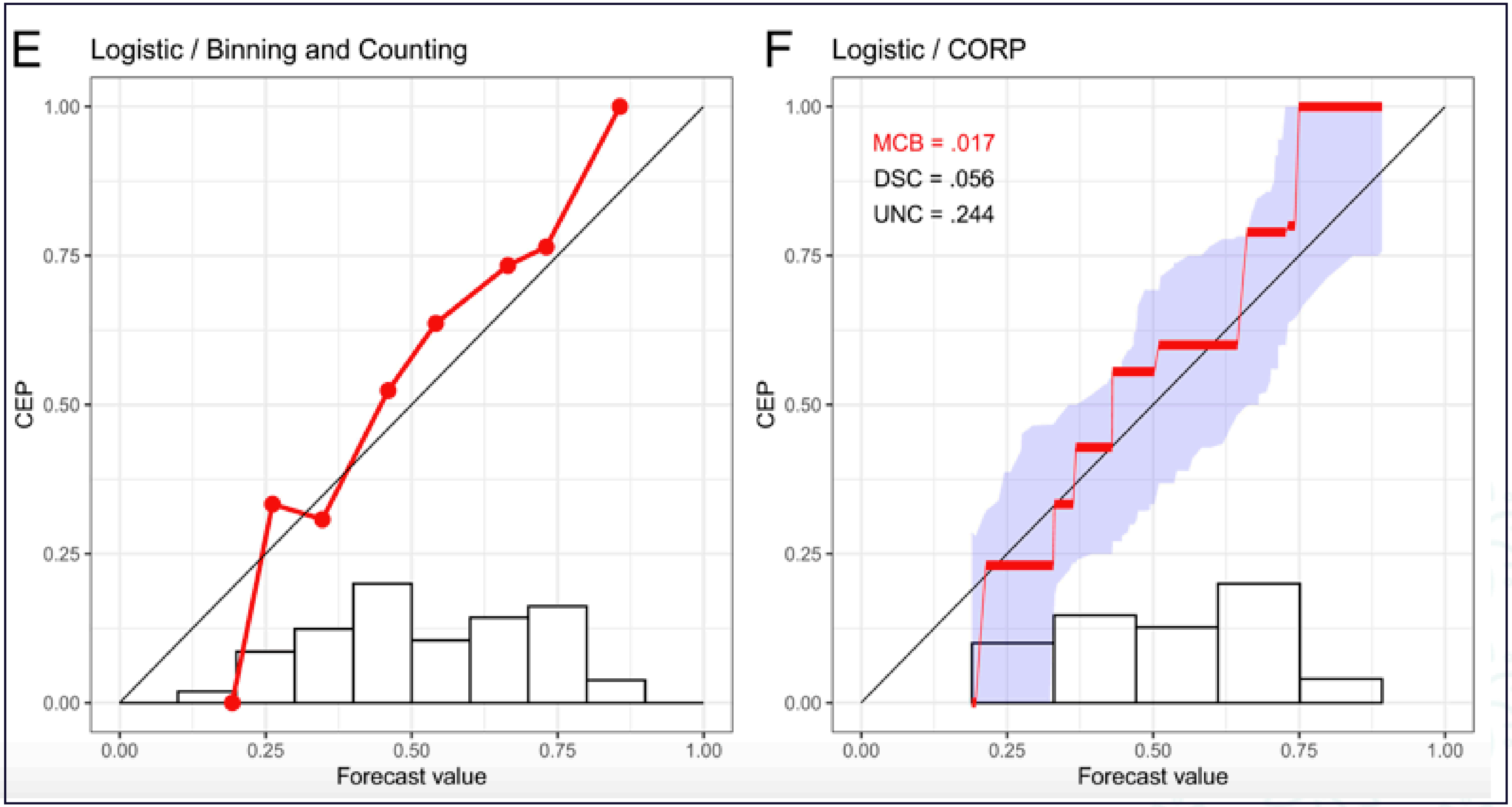
Edited by Bin Yu, University of California, Berkeley, CA, and approved January 13, 2021 (received for review August 5, 2020)

February 17, 2021 | 118 (8) e2016191118 | https://doi.org/10.1073/pnas.2016191118

- This approach provides an automatic way of selecting the optimal bins which is reproducible and produces statistically consistent reliability diagrams -- without the requirement of implementation of decisions or parameter tuning

- It is constructed via nonparametric isotonic regression and implemented using pool-adjacent-violators algorithm, which assigns a (re)calibrated probability under the regularizing constraint of isotonicity
  - C - Consistency
  - O - Optimality
  - R – Reproducibility
  - P – PAV Algorithm Based

Graphical Illustration of PAV Algorithm



Binning and Counting Approach with 10 equally spaced bins

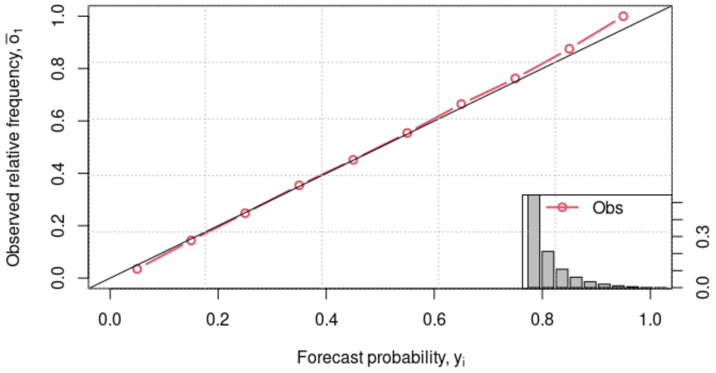
CORP with uncertainty quantification through 90% consistency bands

(Snippet from the referred paper)

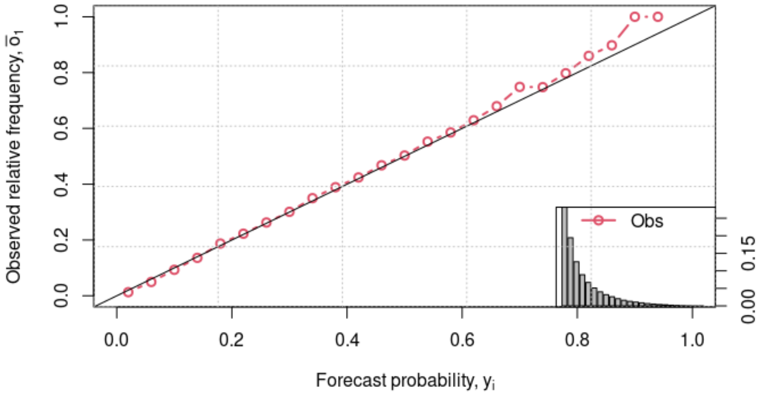
Comparison between Classical Approach and CORP Approach

We compare the Reliability Diagrams under the binning and counting approach with various choices of bin width and CORP approach for LightBGM Model for Mortality:

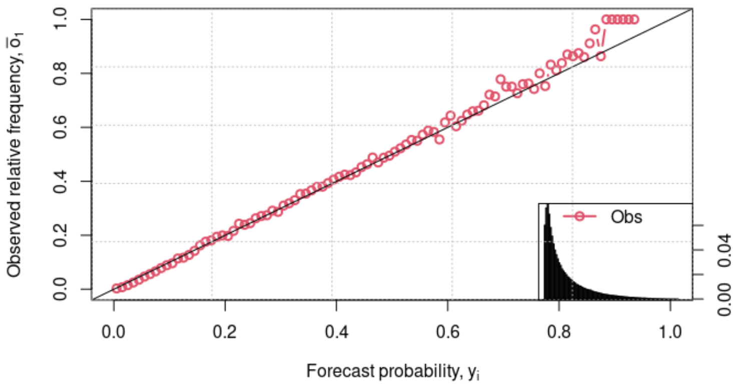
Bin Width = 0.1



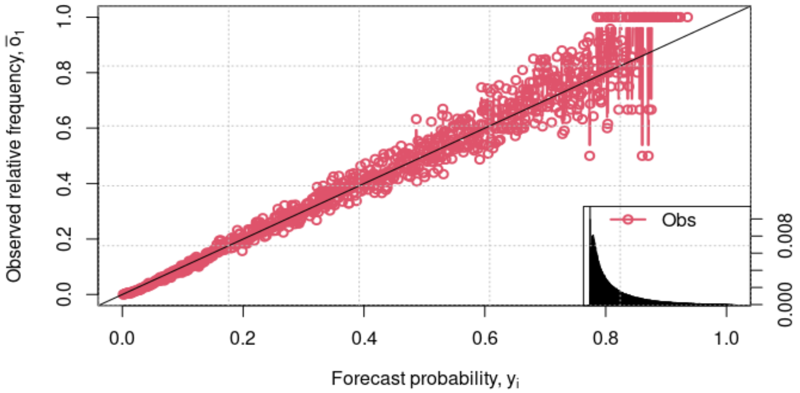
Bin Width = 0.04



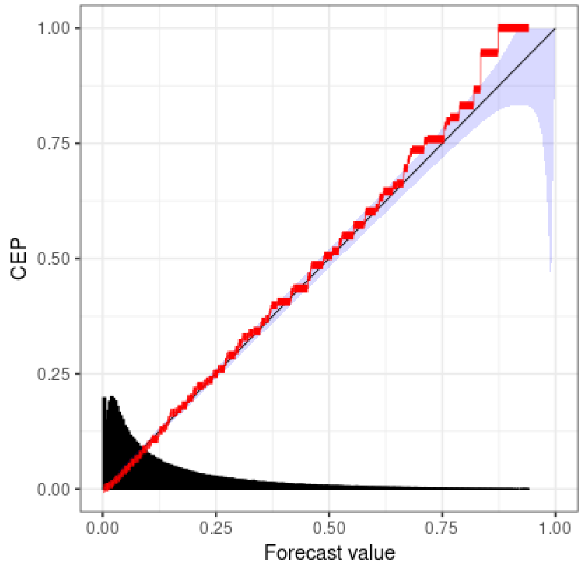
Bin Width = 0.01



Bin Width = 0.001



95% Consistency bands



Classical Binning and Counting Approach

CORP Approach

Comparison between Classical Approach and CORP Approach

We compare the scores under the binning and counting approach with various choices of bin width and CORP approach for LightBGM Model for Mortality:

Bin size = 0.1, No. of bins: 10

Brier Score	0.09620379
Miscalibration	1.39E-04
Discrimination	2.27E-02
Skill	0.18991
Uncertainty	0.1187569
AUC	82.5%

Bin size = 0.04, No. of bins: 25

Brier Score	0.09541323
Miscalibration	6.64E-05
Discrimination	2.34E-02
Skill	0.196567
Uncertainty	0.1187569
AUC	82.5%

Bin size = 0.01, No. of bins: 100

Brier Score	0.09528684
Miscalibration	8.93E-05
Discrimination	2.36E-02
Skill	0.1976312
Uncertainty	0.1187569
AUC	82.5%

Bin size = 0.001, No. of bins: 1000

Brier Score	0.09528246
Miscalibration	3.82E-04
Discrimination	2.39E-02
Skill	0.1976681
Uncertainty	0.1187569
AUC	82.5%

Classical Binning and Counting Approach

Optimally binned by CORP approach

Brier Score	0.09528252
Miscalibration	0.0001289288
Discrimination	0.02360333
Skill	0.1976677
Uncertainty	0.1187569
AUC	82.5% (82.3% - 82.6%)

CORP Approach

Knowing the Data

- The dataset is a 5% nation-wide sample of the Medicare patients while admitted into hospitals
- The train dataset has information of the year 2018 and patients with 12 months of part A and part B coverage, without any part C and age >= 18 are considered
- The test dataset has information of the year 2019

	No. of Patients	No. of Features
Train Set	476593	15341
Test Set	465064	15341

Feature Details

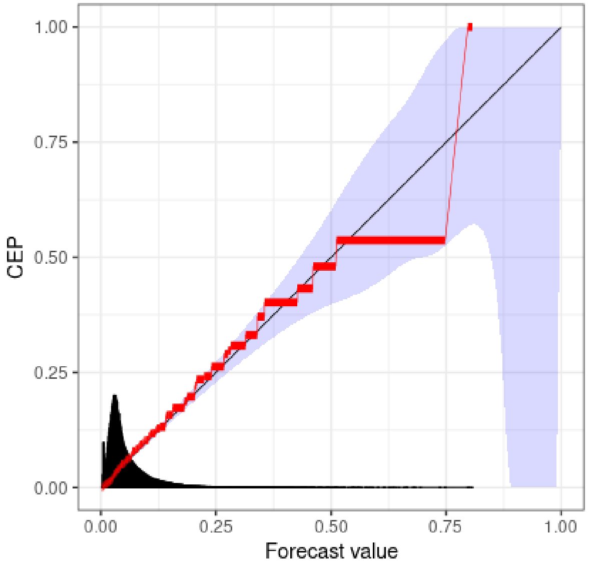
The features consisted of :

- ICD-10 codes of the diseases the patients are affected within 90 and 365 days prior to the admission date in the hospital
- CCSR codes the patient had within 90 days and 365 days before his admission
- CPT codes which are the codes telling us whether the patient had a surgery within 90 days and 365 days prior to admission
- ICD-10 and CCSR codes of the diseases the patient had at the time of admission
- Age and Sex of the patient

Reliability Diagrams for Sepsis

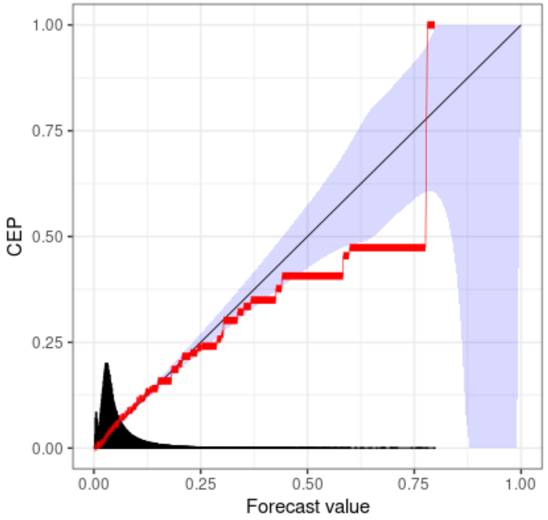
LightGBM gives the best AUC, least Brier Score, maximum skill.

LightGBM



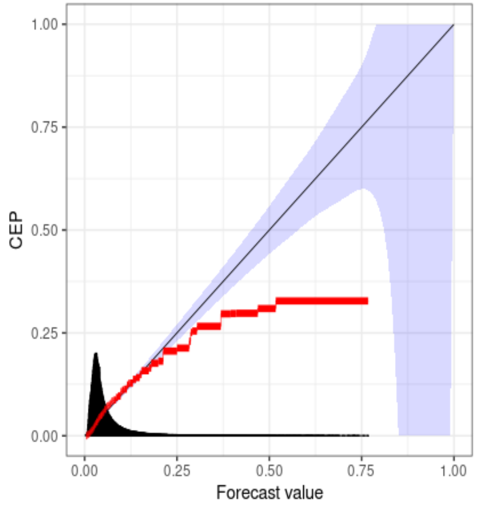
Brier Score	0.05097948
Miscalibration	4.765206e-05
Discrimination	0.002492027
Skill	0.04575438
Uncertainty	0.05342385
AUC	72.8% (72.5% - 73.1%)

XGBoost



Brier Score	0.05106985
Miscalibration	6.373789e-05
Discrimination	0.00241774
Skill	0.04406276
Uncertainty	0.05342385
AUC	72.6% (72.3%-72.9%)

Regularized-logistic-regression

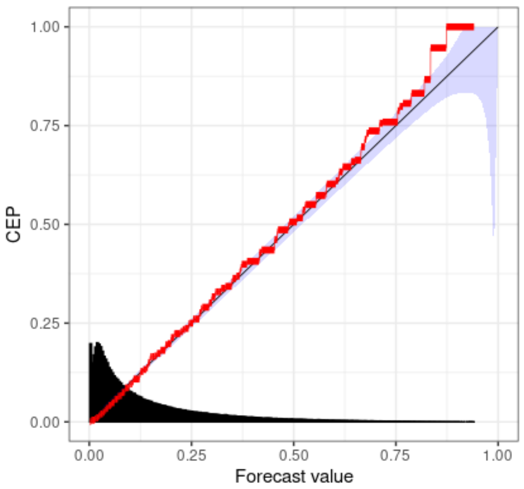


Brier Score	0.05142141
Miscalibration	0.0002054804
Discrimination	0.002207926
Skill	0.03748224
Uncertainty	0.05342385
AUC	71.9% (71.6% - 72.2%)

Reliability Diagrams for Mortality

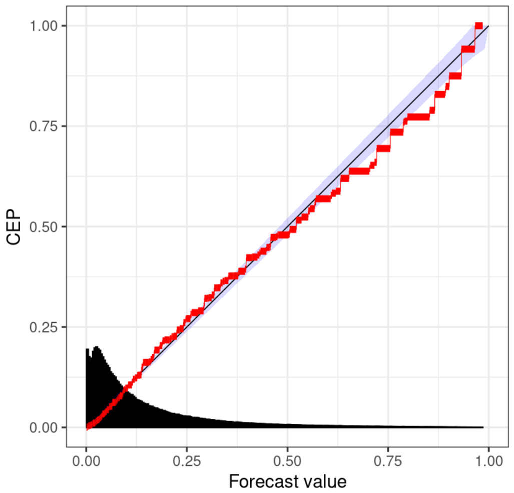
LightGBM gives the best AUC, least Brier Score, maximum skill.

LightGBM



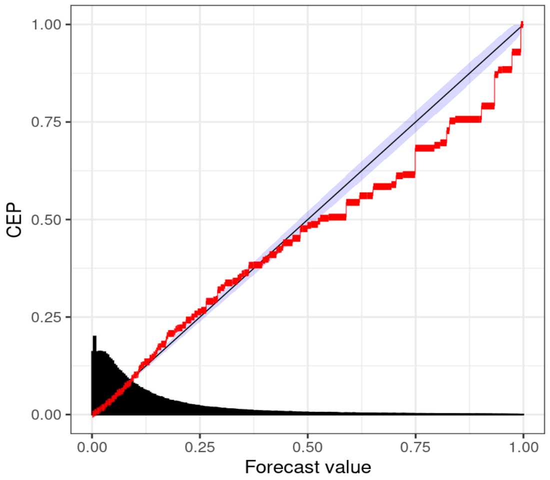
Brier Score	0.09528252
Miscalibration	0.0001289288
Discrimination	0.02360333
Skill	0.1976677
Uncertainty	0.1187569
AUC	82.5% (82.3% - 82.6%)

XGBoost



Brier Score	0.0957091
Miscalibration	0.0001970022
Discrimination	0.02324482
Skill	0.1940756
Uncertainty	0.1187569
AUC	82.2% (82.1% - 82.4%)

Regularized-logistic-regression

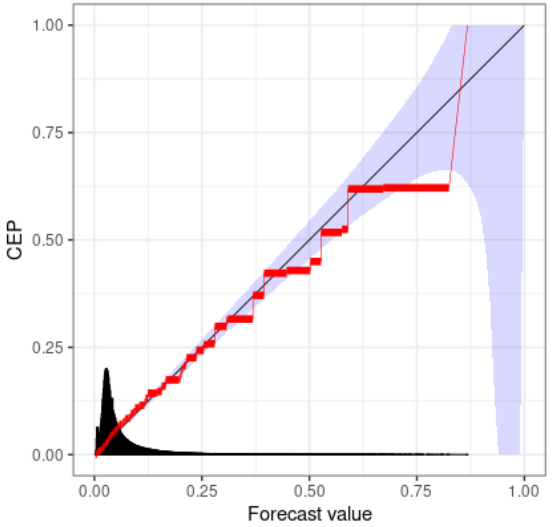


Brier Score	0.09635019
Miscalibration	0.0004582208
Discrimination	0.02286495
Skill	0.1886772
Uncertainty	0.1187569
AUC	82.0% (81.9% to 82.2%)

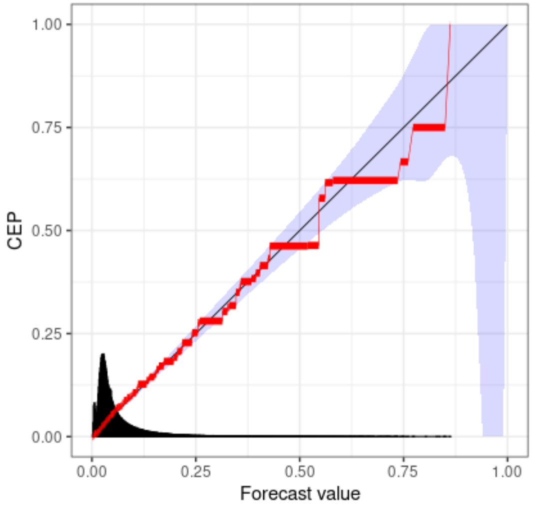
Reliability Diagrams for Respiratory Failure

LightGBM gives the best AUC, least Brier Score, maximum skill.

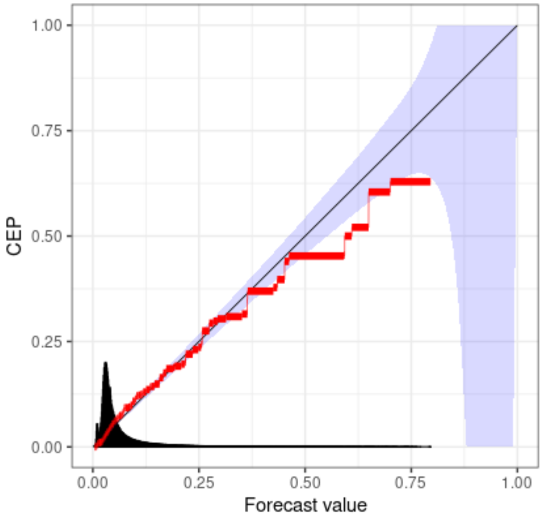
LightGBM



XGBoost



Regularized-logistic-regression



Brier Score	0.05559781
Miscalibration	8.451616e-05
Discrimination	0.004112924
Skill	0.06756102
Uncertainty	0.05962622
AUC	74.30% (74.1% - 74.6%)

Brier Score	0.05547933
Miscalibration	5.834745e-05
Discrimination	0.004205233
Skill	0.06954802
Uncertainty	0.05962622
AUC	74.6% (74.3% - 74.9%)

Brier Score	0.05575578
Miscalibration	0.0001112746
Discrimination	0.003981713
Skill	0.06491168
Uncertainty	0.05962622
AUC	74.1% (73.8% - 74.4%)

Conclusion

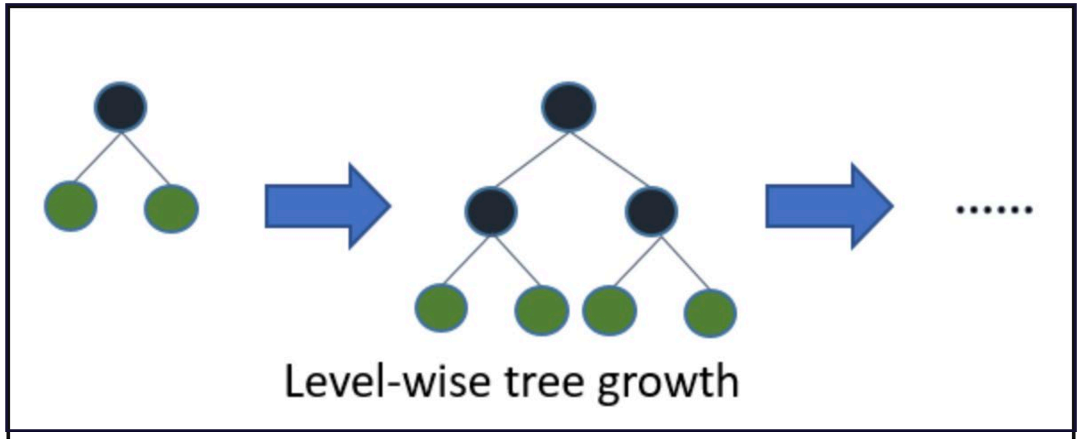
- The CORP approach allows us to compare both calibration and discrimination across different models
- It is a mathematically rigorous and justifiable method for automatically choosing bin sizes/number of bins in calibration analyses
- Eliminates instabilities associated with ad hoc choices of bin widths/bin counts
- At least for our 5% sample LDS data, boosting models generally outperform logistic models on both calibration and discrimination
- LightGBM appears to provide better calibration and discrimination than XGBoost
- Logistic models, however, are much more interpretable, and with the 100% VRDC data, differences between boosting and logistic may disappear
- From a business point of view and a clinical point of view, is the extra performance of boosting or other fancy ML models worth the loss of interpretability?
- There are, of course, methods (RuleFit, for example), that allow us to incorporate simple rules from boosted tree models

HEALTH
DATA ANALYTICS
INSTITUTE

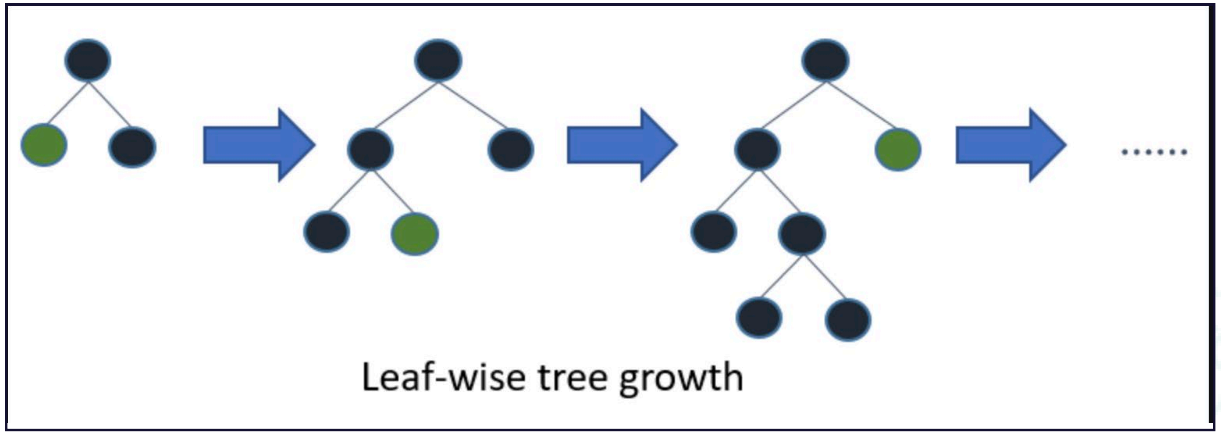
Thank You

madhusree.chowdhury@hda-institute.com

Appendix : Differences Between XGBoost and LightGBM



XGBoost



LightGBM

Appendix : Differences Between XGBoost and LightGBM

XGBoost	LightGBM
Uses a pre-sorted and histogram-based algorithm for computing the best split.	Faster due to utilization of Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).
Treats categorical variables as numerical variables with order.	Accepts a parameter to check which column is a categorical column and handles this issue with ease by splitting on equality.
Gain is available in feature importance methods.	Gain is available in feature importance methods.
Split/ Frequency/ Weight is available in feature importance methods.	Split/ Frequency/ Weight is available in feature importance methods.
Coverage is available.	Coverage is not available.