Challenges in Assessing Probabilistic Classifiers: ROC Curves and Beyond

Term paper for the
Data Analytics Seminar
Winter Term 2023/2024
Chair of Econometrics and Statistics
University of Hohenheim

First examiner: M.Sc. Marius Puke
Second examiner: Prof. Dr. Thomas Dimpfl

Submitted by: Noemi Avello
Student Number: 998749

Date of Submission: March 1, 2024

Contents

1 Introduction

Evaluating the prediction power of real-valued markers or features for binary outcomes is crucial in every field. Having diagnostic tools for evaluating and comparing predictive capacity is a prerequisite for creating and refining probability projections.

Receiver Operating Characteristic (ROC) curves are essential tools for assessing the prediction power of variables, markers, or features in different contexts. When evaluating the effectiveness of probabilistic classifiers derived from statistical learning methods like random forest, ROC curves are frequently utilized. However, for a variety of reasons, this approach's interpretability is restricted. The monotonicity constraint on the conditional event probability, which is the foundation of the methodology and must be invoked to justify the building of any raw ROC diagnostic or ROC curve, is what gives concavity its crucial role in the understanding and modeling of ROC curves.

This seminar paper describes how to apply a technique that subjects the marker or feature in question to pool-adjacent violators to morph ROC curves into their concave hull. In addition, it was evaluated some alternative techniques.

The structure of the document is as follows. Some theory foundations for evaluating probability forecasts are established in Section 2, formally expressing the relationship between the probability forecast and the Conditional Event Probability (CEP) with calibration. Furthermore, as depicted visually in reliability diagrams, a probability forecast or probabilistic classifier is deemed reliable or calibrated if the anticipated probabilities correspond with ex-post observed frequencies. The traditional method of creating reliability diagrams by binning and counting has been hindered by its inability to maintain stability in the face of unforeseen, unsystematic implementation decisions. As a result, the CORP approach which produces automatically generated, appropriately binned, reproducible, and statistically consistent reliability diagrams is included in this section. In Section 2, the differences between ROC curves and raw diagnostics are discussed, along with the particular importance of concavity in ROC curve modeling and interpretation. Two ROC alternatives, precision-recall curves, and MCB-DSC plots are covered in Section 3 to examine alternative evaluation techniques with different advantages despite the same objective.

The seminar paper closes in Section 4, which discusses an empirical application using real-world data of vehicle loan default used in a Hackathon competition in 2019. Data on forecasts and seminar application have been deposited at GitHub (`https://github.com/TimoDimi/replicationDGJ20`).

2 Basics in probability forecast evaluation

Probabilistic forecasts provide a predicted probability distribution for a forthcoming item or event, taking into account forecast uncertainty. A future binary or dichotomous occurrence, such as a credit default or not credit default, or the likelihood of a precipitation occurrence in the future is the most straightforward scenario (Murphy and Winkler (1992)). Predictive probability distribution in the binary situation is only the ex-ante likelihood, or the event to occur.

Assigning a positive result as $Y = 1$ and a negative result as $Y = 0$. A probability forecasting for Y may have a probability $p \in [0.1]$ for a positive outcome.

2.1 The CEP

The Conditional Event Probability (CEP) function addresses the task of predicting an outcome not yet observed, denoted by the random variable Y, ranging between 0 and 1 based on a set of observables X. The likelihood of Y being 1 given X equal to x is analogous to both the conditional mean and the distributional forecast for the binary outcome Y.

$$CEP(x) = \mathbb{Q}(Y = 1|X = x) \tag{2.1}$$

In probability forecasting, the objective is to maximize the prediction distributions' sharpness, subject to calibration (Roopesh and Tilmann (2010)).

2.2 Calibration

The degree to which conditional event frequencies agree with prediction probabilities is known as calibration or reliability, typically assessed via a graphical analysis.

2.2.1 Probability forecast and CEP relation for calibration

A forecast quality or performance can present different attributes with their respective measures. Calibration or reliability is a fundamental necessity for any probability forecast or probabilistic classifier. Essentially, a probabilistic classifier assigns a predictive probability to an occurrence of binary events.

Defining, the expected value of Y given the forecast X:

$$\mathbb{E}[Y|X = x] \tag{2.2}$$

A probability forecast achieves calibration or is reliable when, under the condition of any forecast value p, the event materializes in 100 times p percent of the instances under consideration. This means,

$$CEP(x) = \mathbb{Q}(Y = 1|X = x) = \mathbb{E}[Y|X = x] \tag{2.3}$$

Thus, the probability forecast X is calibrated if the function is equal to the forecast value x for all relevant $x \in [0, 1]$. This criterion serves as a cohesive representation of calibration for binary outcomes. For instance, if we consider all cases with a predictive probability of about 0.60, the observed event frequency ought to be about 0.60 as well.

2.2.2 How to assess calibration

(i) Reliability diagrams

Conditional empirical event frequency versus forecast probability is plotted in one of the most well-known empirical calibration curve displays, demonstrating the possibility of an uncalibrated forecast when there are notable departures from the diagonal.

The ”binning and counting” method, which begins by choosing a specific, usually arbitrary number of bins for the forecast values, is the foundation of the classical reliability diagram (Dimitriadis, Gneiting, and Jordan (2021)). Then, the corresponding conditional event frequency for each bin is plotted against the average prediction value or midpoint of the bin.

Despite being simple to use, the traditional method of generating reliability diagrams can be quite sensitive to the bin specifications, and even the smallest alteration can have a significant impact on the visual representation. As a result, instability is a significant problem that is usually brought on by several instances of the same prediction value at bin breaks. Additionally, the instabilities affect related numerical calibration metrics like the Brier-score reliability.

(ii) CORP reliability diagram

Based on nonparametric isotonic regression and the Pool-Adjacent-Violators (PAV) algorithm to estimate Conditional Event Probabilities (CEPs), the CORP approach produces provably statistically consistent, optimally binned, and reproducible reliability diagrams in an automated manner. This results in a fully automated choice of bins that adapts to both discrete and continuous settings, without any need for tuning parameters or user intervention.

In addition, the CORP diagram includes quantitative measurements of uncertainty (UNC), discrimination ability (DSC), and (mis)calibration (MCB), which outperform the traditional Brier-score decomposition in terms of stability (Dimitriadis, Gneiting, Jordan, and Vogel (2023)). A histogram showing the forecast values' unconditional distribution is typically added to the reliability curve.

The basic idea of CORP is to use nonparametric isotonic regression to estimate a forecast's CEPs as a monotonic, non-decreasing function of the original forecast values. To each original forecast value, the PAV algorithm assigns a (re)calibrated probability under the regularizing constraint of isotonicity, and this solution is optimal under a very broad class of loss functions. In particular, the PAV solution constitutes both the nonparametric isotonic least squares and the nonparametric isotonic maximum-likelihood estimate of the CEPs.

For a specific form record,

$$(x_1, y_1), ..., (x_n, y_n) \tag{2.4}$$

Assume, without losing generality, $x_1 \leq ... \leq x_n$, and let

$$\hat{x_1} \leq ... \leq \hat{x_n} \tag{2.5}$$

represent the values that the PAV algorithm recalibrated. The piecewise linear curve that connects the points $(x_1, \hat{x_1}), ..., (x_n, \hat{x_n})$ is shown in the CORP reliability diagram. When the original forecast is calibrated, the reliability curve is on the diagonal, and $x_1 = \hat{x_1}, ..., x_n = \hat{x_n}$. In every other case, consistent departures from the diagonal point to a lack of calibration. Curves below the diagonal overestimate the predictions and above the diagonal underestimate.

CORP reliability diagrams are provably statistically consistent, in contrast to the binning and counting method, which has not been the subject of asymptotic analysis. Moreover, CORP is asymptotically efficient, meaning that an estimate as accurate as feasible in the large sample limit is produced by its automatic binning decision.

(ii.1) CORP Score Decomposition

The CORP decomposition decomposes a mean score

$$\bar{S} = MCB - DSC + UNC \tag{2.6}$$

the miscalibration element MCB represents the variation in mean scores between the first and (re)calibrated forecasts showing the degree to which the conditional event frequencies and predicted probabilities diverge. In the same direction, the DSC component evaluates a prediction's capacity to discriminate between events and non-events by calculating the difference between the mean score for the reference and the (re)calibrated forecast. Large values of the DSC component are also desirable, in addition to small values of the MCB. The traditional measure of uncertainty (UNC), which is independent of the forecast being examined is just the reference forecast's mean score used to evaluate the intrinsic complexity of the prediction task.

As before, let $\hat{x_1} \leq ... \leq \hat{x_n}$ represent the PAV re-calibrated values from equation (2.5), as depicted in the CORP reliability diagrams, and for a given record (2.4), assume without loss of generality that $x_1 \leq ... \leq x_n$. Additionally, let the realized unconditional event frequency be represented by $r = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. S serving as any appropriate scoring guideline, let's

$$\bar{S}_C = \frac{1}{n}\sum_{i=1}^{n} S(\hat{x_i}, y_i) \text{ and } \bar{S}_R = \frac{1}{n}\sum_{i=1}^{n} S(r, y_i) \tag{2.7}$$

indicate, respectively, the mean score for the constant Reference prediction r and the (re)Calibrated probability. The average score $\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S(x_i, y_i)$ splits down as (Dimitriadis et al. (2023))

$$\bar{S} = \underbrace{(\bar{S} - \bar{S}_C)}_{\text{MCB}} - \underbrace{(\bar{S}_R - \bar{S}_C)}_{\text{DSC}} + \underbrace{\bar{S}_R}_{\text{UNC}} \tag{2.8}$$

The resulting decomposition of the mean score is exact and ensures that $DSC \geq 0$ with equality if the (re)calibrated forecast is constant, and $MCB \geq 0$ with equality if the initial forecast is calibrated.

Specifically, the CORP score decomposition, never produces paradoxical negative values of the components. Parts away from the diagonal suggest a lack of calibration, whereas extended horizontal segments are indicative of diminished discriminating ability. These situations of vanishing components ($MCB = 0$ or $DSC = 0$) confirm the intuitive interpretation of CORP reliability diagrams.

2.3 Discrimination analysis via ROC curve

In binary problems, receiver operating characteristic (ROC) curves are widely used to assess variables, markers, or characteristics as possible predictors. In a nutshell, ROC curves show prospective prediction capability independent of calibration factors.

The relationship between the Hit Rate (HR) and False Alarm Rate (FAR) as the decision threshold changes is plotted graphically via the ROC. Specifically, take into consideration the joint distribution $\mathbb{Q}$ of the pair (X, Y), where Y is a binary event and X is a real-valued covariate, marker, or feature. It is implicitly understood that larger values of X give stronger support for the event to materialize ($Y = 1$) the prevalence

$\pi_1 = \mathbb{Q}(Y = 1) \in (0,1)$ and the conditional Cumulative Distribution Functions (CDFs) define the joint distribution $\mathbb{Q}$ of (X, Y).

$$F_1(x) = \mathbb{Q}(X \leq x)|Y = 1) \text{ and } F_0(x) = \mathbb{Q}(X \leq x)|Y = 0) \tag{2.9}$$

A classifier with a Hit Rate (HR), can be produced by using any threshold value x to predict a positive result $(Y = 1)$ if $X > x$ and a negative outcome $(Y = 0)$ if $X \leqslant x$.

$$HR(x) = \mathbb{Q}(X > x|Y = 1) = 1 - F_1(x), \tag{2.10}$$

and False Alarm Rate (FAR),

$$FAR(x) = \mathbb{Q}(X > x|Y = 0) = 1 - F_0(x) \tag{2.11}$$

FAR is the probability of making a false positive decision, whereas HR is that of making a correct positive decision.

ROC curves consider all potential thresholds, making them helpful for comparing different classifiers. It is desirable to have high hit rates and low false alarm rates, hence it is best if the ROC curve approaches the upper left corner of the unit square. A common metric for assessing a feature's potential predictive usefulness is the Area Under the ROC Curve (AUC).

2.3.1 The Area Under the Curve (AUC)

The area under the ROC curve measure is a widely used method in scientific publications to compare the prediction abilities of probabilistic classifiers (Gneiting and Walz (2022)). The likelihood that a value randomly selected from the empirical distribution of prediction values for an event will be greater than a value selected from the distribution for a non-event is an attractive interpretation of AUC.

Perfect discriminating ability is indicated by an AUC value of 1; no discrimination is indicated by a $\frac{1}{2}$ matching to the diagonal trivial ROC curve. If the value is less than $\frac{1}{2}$, it suggests that the accuracy of the forecast could be increased by changing the forecasts for 0 and 1. AUC is invariant under strictly increasing transformations, just as the ROC curve is, and it is noted that AUC exclusively concerns discrimination ability, while ignoring (mis)calibration (Dimitriadis et al. (2023)).

2.3.2 Raw ROC diagnostics and ROC curves

In order to understand the unique function of concavity in reading and modeling ROC curves, we must distinguish between raw ROC diagnostics and ROC curves.

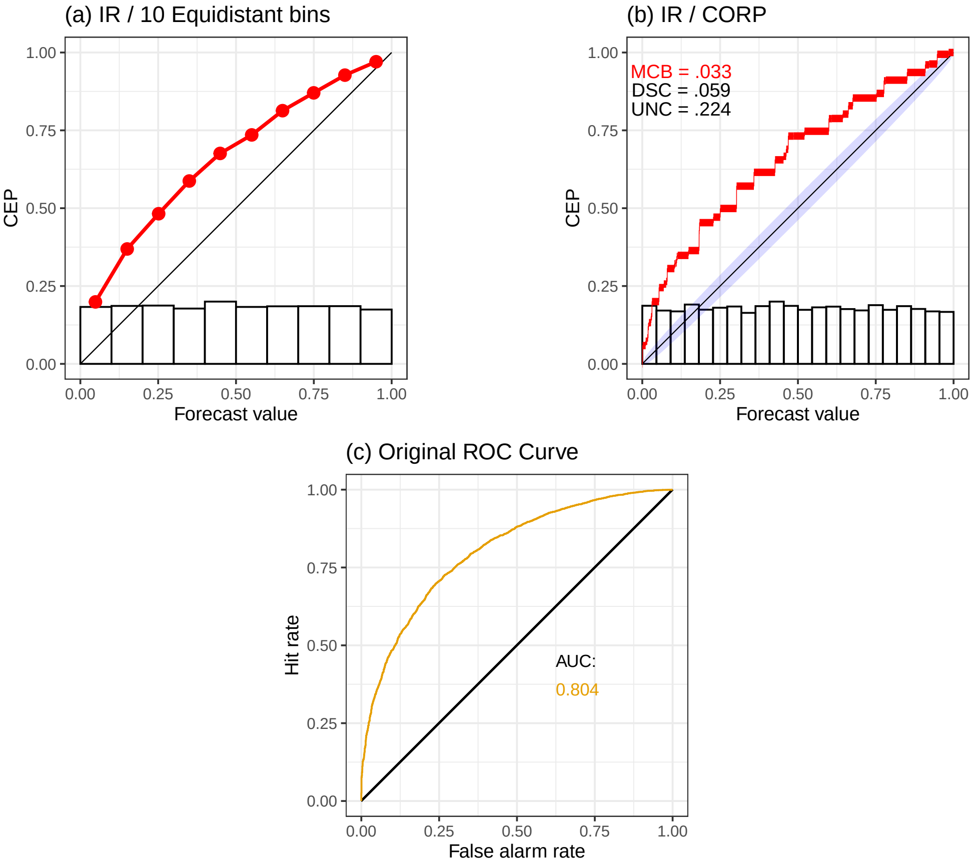

Figure 1: Reliability diagrams (a and b). Under the binning and counting approach, a reliability diagram is displayed with 10 equally spaced bins for selection (a). Furthermore, a CORP reliability diagram (b). The forecast values for n = 92 are displayed as a distribution in the histograms at the bottom. The original ROC curves for the forecasts $x_1, ..., x_n$ from (2.4) are displayed in Panel (c).

ROC diagnostics in this typical context address the form's points $(FAR(x), HR(x))'$, where $FAR(x) = 1 - F_0(x)$ is the false alarm rate and $HR(x) = 1 - F_1(x)$ the hit rate at the level of the threshold $x \in \mathbb{R}$. Formally, the point set is the raw ROC diagnostic for the bivariate distribution $\mathbb{Q}$ and the random vector (X, Y) (Gneiting and Vogel (2022)).

$$R^* = \left\{ \begin{pmatrix} 1 - F_0(x) \\ 1 - F_1(x) \end{pmatrix} : x \ \epsilon \ \mathbb{R} \right\} \tag{2.12}$$

inside the square of units. F_0, F_1, and either of the two marginal distributions clearly characterize the bivariate distribution $\mathbb{Q}$ of (X, Y). However, because of the well-known invariance of ROC diagnostics under strictly increasing transformations of X and variations in the predominance of the binary outcome, the raw ROC diagnostic combined with a single marginal does not represent $\mathbb{Q}$. Nonetheless, $\mathbb{Q}$ is determined by combining the two marginal distributions and the raw ROC diagnosis.

Linear interpolation is used to create a ROC curve from the raw ROC diagnostic. Specifically, the utilization of linear interpolation enables an equitable and straightforward comparison of continuous, discrete, and ordinal data.

For the illustration of reliability diagrams and raw ROC diagnosis, a data set was simulated, where x has a uniform distribution $x \sim U(0,1)$, the calibration curve of $p(x)$ is $p(x)^{0,5}$, y is binomial with $(n, 1, p(x))$, and choosing an n equal to 10000. Figure 1 illustrates reliability diagrams based on the binning and counting approach with a choice of m = 10, the CORP reliability diagrams also displaying measures of (most importantly, and hence highlighted) (mis)calibration (MCB), discrimination (DSC), and uncertainty (UNC), and the raw ROC curve using isotonic regression (IR).

In real-world settings, this nearly perfect situation shown in the simulation will not hold, as illustrated Gneiting and Vogel (2022), Gneiting and Walz (2022), and Dimitriadis et al. (2023), reports ROC curves that are not always increasing. In addition, Fig 4.8 in James., Witten., Hastie., and Tibshirani. (2021) on page 151 provides ROC curves that fail to be concave. The concavity of ROC curves is a factor that is frequently ignored yet is crucial. However, there can be a different approach technique to obtain concave hull ROC curves to get sensible plots to have better decision-making.

2.3.3 Concave ROC curves

The concavity of ROC curves is a valuable but frequently overlooked factor. It is nearly always the case that the original ROC curves created using empirical data are not concave. When the conditional event probability is nondecreasing with the forecast value x, which is rarely the case for empirical data, a ROC curve is concave. ROC curves measure prospective predictive ability or discrimination capacity. Although potential predictive ability does not depend on calibration, it can only be evaluated in the case of bigger prediction values, which are indicative of higher occurrence probabilities.

The ROC curve's concavity suggests right away that the area under the curve is a closed convex set, which includes the curve itself. The area under the curve can be readily demonstrated to be strictly convex if the ROC curve is strictly concave. Any convex combination of points in the area under and on the curve in this situation will be under the curve.

Assuming that F_1 and F_0 have continuous, strictly positive Lebesgue densities f_1 and f_0 in the interior of an interval, which is their common support, we obtain the regular setup (Gneiting and Vogel (2022)).

The likelihood ratio can be defined for each x in the interior of the support,

$$LR(x) = \frac{f_1(x)}{f_0(x)} \tag{2.13}$$

and the conditional event probability,

$$CEP(x) = \mathbb{Q}(Y = 1|X = x) = \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \tag{2.14}$$

It can be shown that in the regular and discrete settings statements (a), (b), and (c) are equivalent (Gneiting and Vogel (2022)):

(a) The ROC curve is concave.

(b) The likelihood ratio is nondecreasing.

(c) The conditional event probability is nondecreasing.

A function $R : [0, 1] \to [0, 1]$ can be used to identify the ROC curve in the standard setup, where $R(p)$ is defined as $R(p) = 1 - F_1(F_0^{-1}(1 - p))$ for $p \in (0, 1)$. The function R is evidently concave if the ROC curve is concave, and as a result, its derivative $R^{'}(p)$ is nonincreasing in $p \in (0, 1)$. The equivalency of (a) and (b), however, is established by the slope $R^{'}(p)$, which equals the likelihood ratio $LR(x)$ for a specific value x that declines with p. Additionally,

$$LR(x) = \frac{\pi_0}{\pi_1} \frac{CEP(x)}{1 - CEP(x)} \tag{2.15}$$

and in $c \in (0, 1)$, the function $c \to c/1 - c$ is nondecreasing, producing the equivalency of (b) and (c).

The discrete situation, where the support of feature X is either a countably infinite or finite set, is the next setting we look at. The case of empirical ROC curves is one example of this scenario, but it's not the only one. We can determine the likelihood ratio for each x in the discrete support of X.

$$LR(x) = \begin{Bmatrix} \mathbb{Q}(X = x|Y = 1)/\mathbb{Q}(X = x|Y = 0) & \text{if } \mathbb{Q}(X = x|Y = 0) > 0 \\ \infty, & \text{if } \mathbb{Q}(X = x|Y = 0) = 0 \end{Bmatrix} \tag{2.16}$$

and the conditional event probability,

$$CEP(x) = \mathbb{Q}(Y = 1|X = x) \tag{2.17}$$

Concavity plays a crucial role in the interpretation and modeling of ROC curves because of the monotonicity condition (c) on the conditional event probability. This condition is fundamental to the methodology and must be used to support the creation of any raw ROC diagnostic or ROC curve.

However, the solution is quite straightforward. Instead of using the original forecasts, the pool-adjacent violators (PAV) can be used to compute the ROC curve, which then transforms into its concave hull the smallest concave curve which is located to its upper left. Thus, the PAV algorithm converts into an isotonic, calibrated probabilistic classifier. Specifically, the PAV solution consists of the nonparametric isotonic maximum-likelihood estimate of the CEPs as well as the nonparametric isotonic least squares.

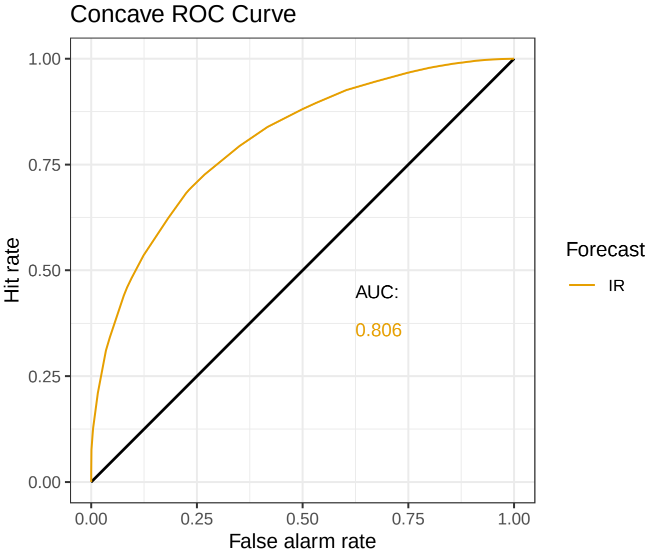

Figure 2: The original ROC curve transforms into its concave hull, as seen by concave ROC curves from the PAV re-calibrated forecasts.

Recalibrating removes variations in calibration, so the corrected, concave ROC curve is the only one used to measure discriminating ability, as illustrated in Figure 2. On the other hand, conditional event frequencies that are not monotone have confusing effects, whereas the original ROC curves concentrate on discrimination abilities. A change in the ROC curve does not contradict the previously stated invariance under strictly increasing transformations because, in general, the transformation from the original probabilities $x_1 \leq ... \leq x_n$ to the PAV transformed, recalibrated probabilities $\hat{x_1} \leq ... \leq \hat{x_n}$ is monotonic, but not strictly monotonic.

In conclusion, it is highly advised to employ concave ROC curves for empirical data, which are obtained from PAV-transformed forecast values.

3 ROC alternatives

3.1 Precision-recall curves

For situations where there is a significant skew in the class distribution, like credit card fraud, the Precision-Recall (PR) curve might be used instead of ROC curves (Fayzrakhmanov, Kulikov, and Repp (2018)). A common scalar performance metric for contrasting various classifiers is the area under prediction-recall curves, which provide valuable insights into the effectiveness of binary classifiers.

Therefore, to assess the performance of classifiers for imbalanced data sets, it is requires estimate recall and precision,

$$Recall = \frac{TP}{TP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

where TP is true positives, FN false negatives, and FP false positives.

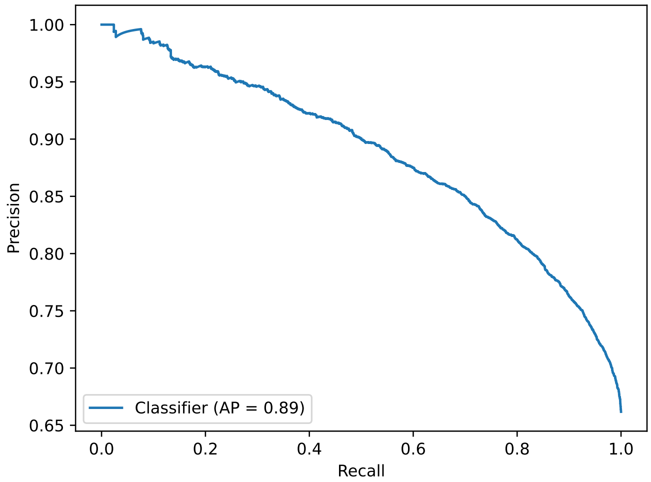

Figure 3: Precision-recall curve for simulated application.

Class imbalance has a direct impact on precision since it affects false positives, but the hit rate is only dependent on positives. For this reason, these impacts are not captured by ROC curves. PR curves therefore have a bigger benefit over ROC curves in skewed data instances where both precision and recall are critical.

The area under the curve (AUC) measure, which is typically used to determine which classifier is superior, summarizes the performance of the classifier into a single quantitative measure. AUCs of superior classifiers are typically higher than those of weaker ones. Clear visual comparisons between two or more classifiers over a wide range of operational points are made possible by ROC and PR curves.

3.2 MCB-DSC plots

Plots of the calibration metric MCB against the DSC measure of discrimination ability show the performance of the classifier against several rivals. MCB-DSC plots help identify methods of interest by visualizing the strengths and limitations of forecasting methods through their simplicity and joint assessment of overall predictive ability, calibration, and discrimination. Therefore, the approach is recommended when a large amount of forecast methods need to be compared. The MCB-DSC plot utilizing brier and logarithmic score is displayed in Figure 4, despite a single forecast.

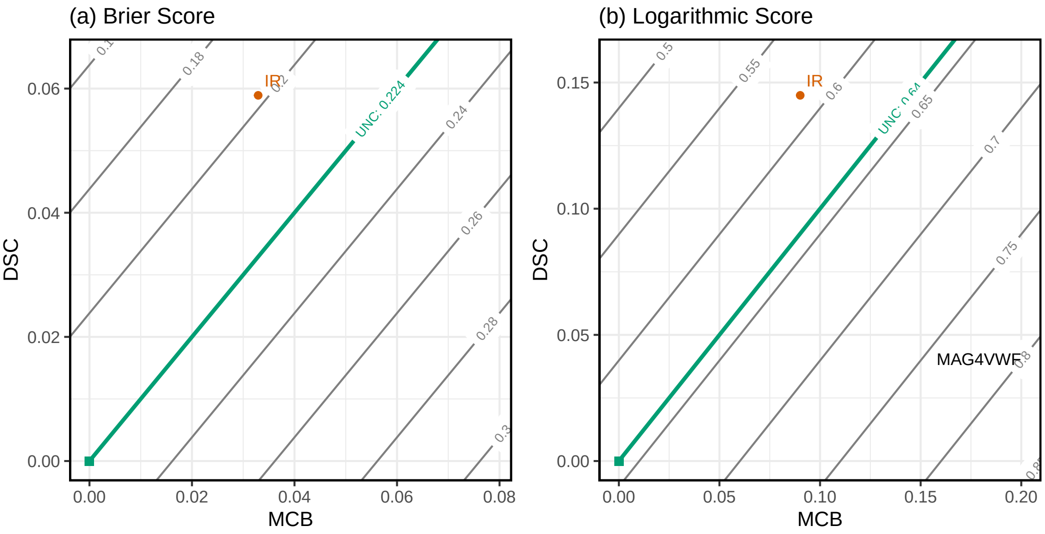

Figure 4: The Brier score and the logarithmic score are the two MCB-DSC graphs used to predict the likelihood of simulated data using isotonic regression. Forecasts that are better (above the line) and worse (below the line) than this baseline are distinguished by the thick green line. The green square at the origin represents the ex post best constant forecast, or the unconditional event frequency.

The difference between the mean score of the original and the (re)calibrated forecast is the miscalibration term $MCB = \bar{S} - \bar{S}_C$. It expresses variations in the score under evaluation from the diagonal of the CORP reliability curve. The discriminating component $DSC = \bar{S}_R - \bar{S}_C$ measures the extent to which the (re)calibrated prediction outperforms the reference score $\bar{S}_R$, which is derived from a calibrated but constant forecast (Dimitriadis et al. (2023)). It is noteworthy that DSC remains invariant under strictly increasing transformations of the forecast values under construction.

4 Empirical application

A real-world dataset of loan default predictions collected in India is used to support and empirically illustrate the theoretical explanation of ROC curve challenges and beyond.[1]

4.1 Description of the data set

The vehicle loan default dataset provides information to predict the probability of borrowers defaulting on a vehicle loan on the due date of their first EMI (equated monthly installment). Furthermore, this dataset comprises three categories of borrower data: bureau data and history; loan information, including loan-to-value ratio, amount, and disbursal details; and borrower personal data, including age and type of job. This dataset contains 41 unique features in total. In terms of payback scenarios, $y = 0$ denotes no loan default and $y = 1$ denotes a loan default.

The training set and test set are the two components of the loan default dataset. The test set is devoid of labels and just contains features because it was submitted as part of a Hackathon competition in 2019. 233154 observations make up the training set, which is the only one useful to be used. Of these findings, loan default occurs in 50611 cases (21.71%).

4.2 Application

The data set explained in the previous section was manipulated to have data in the right format, consistent, and ready for model building. The main techniques applied to the vehicle loan default dataset were missing values and outliers treatment by applying a binning technique. In addition, feature selection drops useless features such as ID variables, data standardization, and dummy insertion to include categorical variables.

To illustrate the seminar application through a real data set three models were applied, logistic model, Naive bayes (NB), and Random Forest (RF).

[1] Data source: https://www.kaggle.com/datasets/avikpaul4u/vehicle-loan-default-prediction

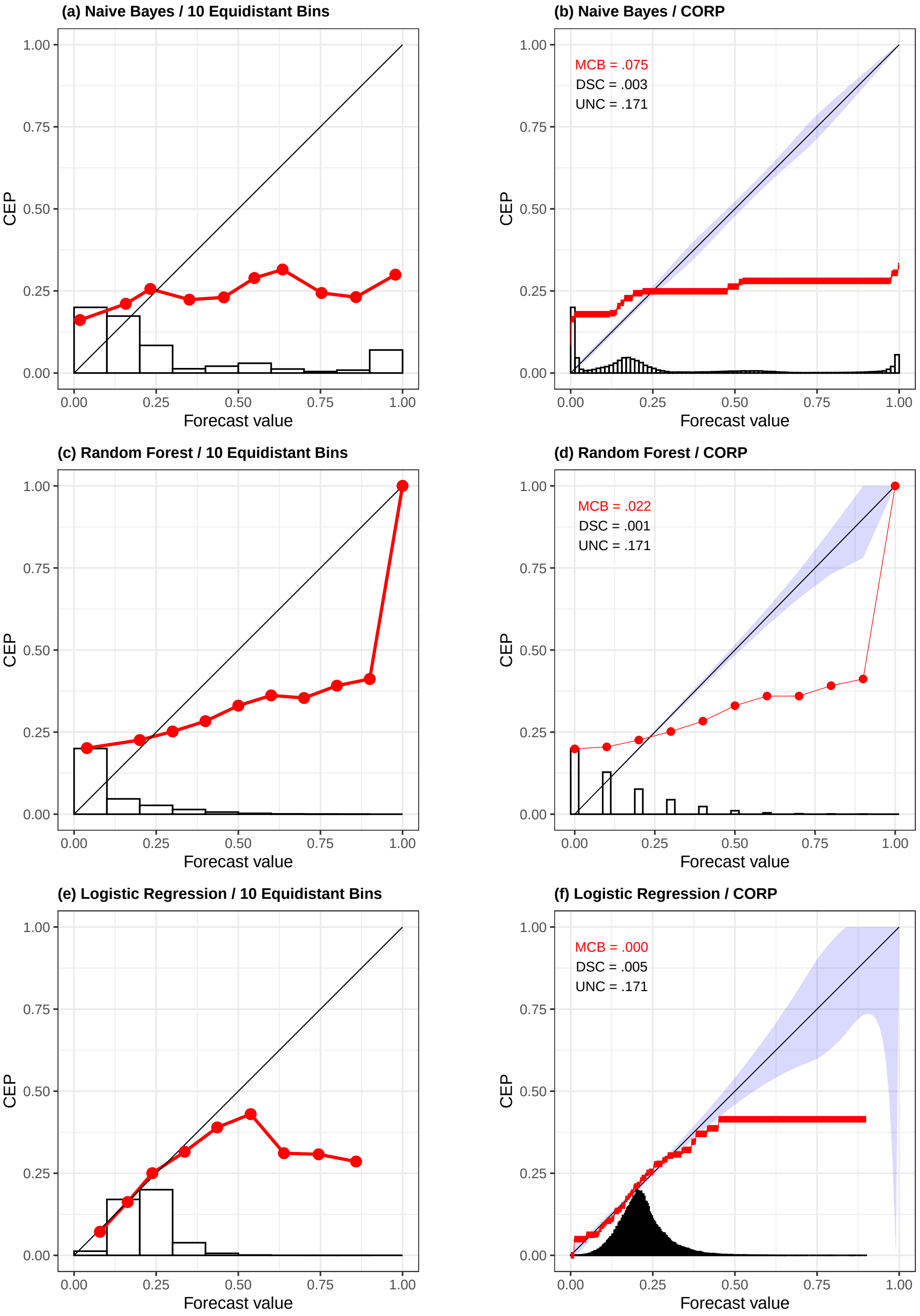

Figure 5: Reliability diagrams based on the binning and counting approach were displayed in plots (a), (c), and (e). For each of the three models use m = 10 equally spaced bins. In addition, the CORP reliability diagrams in (b), (d), and (f) illustrate how the PAV-calibrated probability is plotted against the original prediction value.

The reliability diagram, which compares the predicted probability to the observed event frequency, is the primary diagnostic tool for assessing calibration mostly used in environments with limited prediction probabilities, such as discrete environments. However, the CORP reliability diagram which uses nonparametric isotonic regression and the Pool-Adjacent-Violators (PAV) algorithm to estimate Conditional Event Probabilities (CEPs), yields a fully automated choice of bins that adapts to both discrete and continuous settings, without any need for tuning parameters or implementation decisions, see Figure 5. Furthermore, the plot provides the main variables in the CORP Brier-score decomposition for each model, being the logistic forecast the model with the best (smallest) MCB term and the best (highest) DSC component.

The only ROC curve utilized to assess discriminating ability is the corrected, concave one since recalibrating eliminates calibration discrepancies. Conversely, non-monotonic conditional event frequencies have a confounding effect, while the original ROC curves focus on discriminatory power. Figure 6 shows the concave ROC curve on the right side of the panel and the original version of the ROC curve on the left side for logistic, naive bayes, and random forest predictions.

Since the transformation from the original probability to the PAV transformed, recalibrated probabilities are often monotonic but not strictly monotonic, a change in the ROC curve does not contradict the previously established invariance under strictly rising transformations. Concave ROC curves, which are produced from forecast values that have been PAV-transformed, are highly recommended in conclusion.

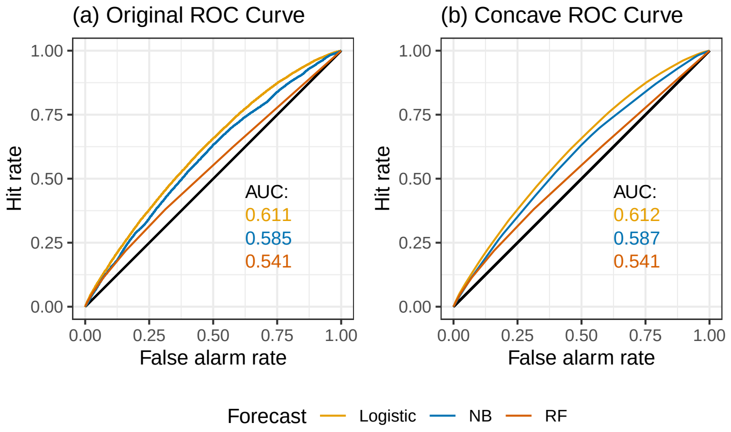

Figure 6: Concave ROC curves from the PAV re-calibrated forecasts demonstrate that the original ROC curve morphs into its concave hull

The Area Under the Precision-Recall Curve (AUC-PR) is the primary metric for assessing model performance due to the imbalanced nature of the dataset. The model with the highest AUC-PR was considered the best performer because higher PR-AUC values indicate better performance in identifying positive instances of loan defaulters. For the application, the logistic model is the best performer, as is shown in Figure 7.

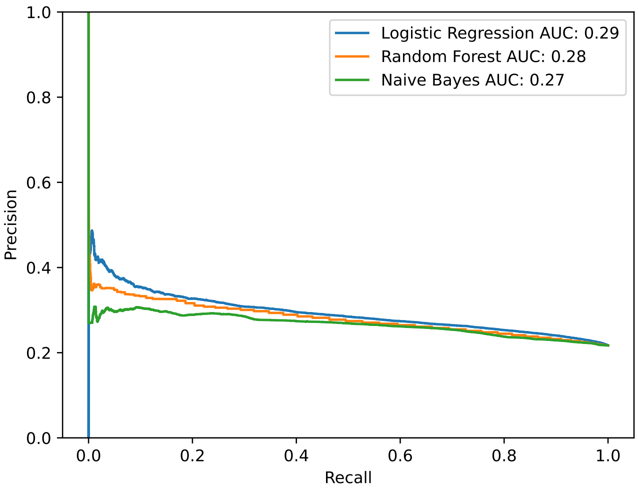

Figure 7: Precision-recall curves for logistic, random forest, and naive bayes models.

The MCB-DSC plots, which display the DSC measure plotted against the MCB component for each competitor involved and are supplemented by parallel contour lines that denote an identical mean score, are useful for comparing a variety of competing forecasting techniques. Because of their ease of use and the way they jointly evaluate overall predictive ability, calibration, and discrimination, MCB-DSC plots help to visualize the advantages and disadvantages of forecasting techniques.

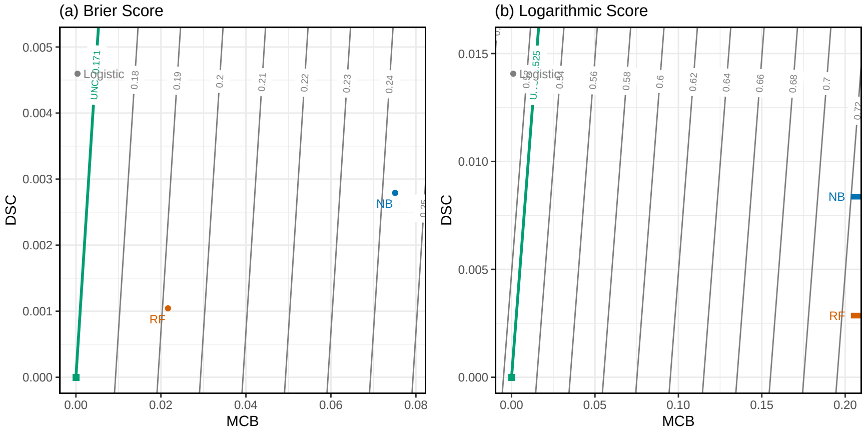

Figure 8: MCB-DSC plots for the logistic, naive bayes and random forest probability forecasts under (a) the Brier score and (b) the logarithmic score.

Brier score MCB-DSC plots for probability forecasts from logistic, random forest, and naive bayes forecasts are displayed in 8. It can be used the thick green line to separate forecasts that are better (above the line) and that are worse (below the line). For instance, only the logistic model is above the line, which is consistent with the previous analysis and the measures of MCB and DSC components. High DSC value but a low MCB metric.

5 Conclusion

ROC curves have been widely used in many scientific fields to evaluate the potential predictive value of variables, features, or markers in binary scenarios. One of the desirable and appealing characteristics of ROC curves in this context is their ease of interpretation in terms of feasible operational circumstances; however, ROC curves are related to a concavity problem, and concavity is crucial to the interpretation and modeling of ROC curves.

Nevertheless, the empirical ROC curve is produced by the PAV method from the corresponding concave hull. PAV-calibrated probabilities, given to the regularizing constraint of isotonicity, are optimal concerning any correct scoring technique. Furthermore, ROC curves offer several alternatives, such as MCB-DSC plots and Precision recall curves, which are mostly utilized for data sets that are unbalanced.

The ROC curve concavity problem can be seen in the empirical application to highlight the limitations of ROC curves, and how concave fits are desirable, if not necessary, as they have nondecreasing conditional event probabilities and likelihood ratios for the predictor variables.

The fitted logistic, naive bayes, and random forest ROC curves fail to be concave initially, but those change markedly towards morphing into its concave hull, when concavity is enforced using re-calibrated probabilities generated by the PAV algorithm.

Examining the alternative evaluation techniques MCB-DSC plots, and precision-recall curves over the dataset, the results were consistent in the evaluation of discrimination ability between the models. MCB-DSC plots generate advantages when comparing multiple forecasts, however, in this context, its value is not as advantageous since it is only three models. In addition, precision-recall curves are mostly utilized for unbalanced data sets which is the case in this empirical application.

However, it needs to be considered that the vehicle loan default dataset's imbalanced data set condition has an impact on this outcome. This gives room for improvement for further iterations.

6 References

References

T. Dimitriadis, T. Gneiting, and A. I. Jordan. Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

T. Dimitriadis, T. Gneiting, A. I. Jordan, and P. Vogel. Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 2023.

R. Fayzrakhmanov, A. Kulikov, and P. Repp. The difference between precision-recall and roc curves for evaluating the performance of credit card fraud detection models, 2018.

T. Gneiting and P. Vogel. Receiver operating characteristic (roc) curves: equivalences, beta model, and minimum distance estimation. *Machine Learning*, 111, 2022.

T. Gneiting and E.-M. Walz. Receiver operating characteristic (roc) movies, universal roc (uroc) curves, and coefficient of predictive ability (cpa). *Machine Learning*, 111: 1–29, 2022.

G. James., D. Witten., T. Hastie., and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer US, 2021. ISBN 9781071614181.

A. H. Murphy and R. L. Winkler. Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455, 1992.

R. Roopesh and G. Tilmann. Combining Probability Forecasts. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):71–91, 2010.

7 Appendix

Appendix “Declaration on the Use of Generative AI"

In the creation of this academic paper, I have used the following system(s)[1] based on artificial intelligence (AI):

1. Grammarly
2.
3.

I also declare that I

☒ have actively informed myself about the capabilities and limitations of the AI systems listed above,

☒ have marked the passages which I have incorporated into the academic paper directly from the AI systems listed above,

☒ have verified that the content generated by the above-mentioned AI systems and adopted by me is factually correct,

☒ am aware that as the author of this academic paper, I bear the responsibility for the statements and assertions made in it.

I have used the aforementioned AI systems as outlined in the Table below.

Design Step	**AI System(s) Used**	**Description of Usage**
Generation of ideas and conception of academic paper		
Literature search		
Literature analysis		
Literature management and citation management		

[1] If you are unsure whether you need to list a given AI system, please contact your examiner.

Selection of methods and models		
Data collection and analysis		
Creation of visualizations		
Interpretation and validation		
Structuring the text		
Formulating the text		
Translating the text		
Editing of the text		
Preparing a presentation of the text	Grammarly	Review of spelling, grammar, and punctuation mistakes.
Other		

Stuttgart, 01.03.2024

Signature

Place, Date