Forecast Intervals for the Area Under the ROC Curve with a Time-varying Population

Lealand Morin[1]
Department of Economics
College of Business Administration
University of Central Florida
Lealand.Morin@business.ucf.edu
https://business.ucf.edu/person/lealand-morin/

December 26, 2018

Abstract

There are several approaches to quantify the variability of the area under the receiver operating characteristic curve (AUROC) to measure the predictive performance of a binary classification model. These methods typically involve calculating the variance of the estimate and calculating a z-statistic, under specified distributional assumptions to characterize sampling variability. In contrast, this paper proposes an approach to characterize variability in the AUROC by allowing for additional variation in the sampling distribution. This additional variation corresponds to changes in the population over time. Then, the variation is quantified by measuring the distance between joint distributions of subsamples from that of the full sample. Under this approach, distance is measured as the Kullback-Leibler divergence between the full sample and other potential samples. The extreme values of the AUROC statistics from distributions with the specified value of divergence provide an interval for forecasting the AUROC to be expected in future applications of the model.

KEY WORDS: ROC curve, AUC, AUROC, classification models, predictive modeling, Mann-Whitney statistic, Wilcoxon statistic, Kullback-Leibler divergence.

[1]I am grateful for helpful feedback from seminar participants at the University of Central Florida. All errors are my own.

1 Introduction

The area under the receiver operating characteristic curve (AUROC) is a popular statistic for measuring the performance of a classification model. There are several approaches to calculate confidence intervals for the AUROC, each one with a set of distributional assumptions. However, practitioners are often concerned with the variability in model performance while the model is used in the future. After a model is built, the performance could be monitored until either the performance degrades or the sample becomes materially different from that on which the model was built. With this possibility, the AUROC statistic will demonstrate variation that is beyond that expected from sampling variability alone. This paper provides a solution to this problem, by proposing a method to calculate forecasting intervals for the area under the ROC curve.

The variation in the predicted AUROC is characterized by allowing for some distance between the empirical distribution, on which the model is built, and future distributions that may hold while the classification model is used out of sample. The prediction interval is defined by solving for the values of the AUROC statistic that lie furthest from that of the observed sample but within a specified distance from the empirical distribution. These bounds of the interval are calculated by first solving for the distribution with the closest data generating process that produces a particular AUROC value. Then the bounds are calculated as the upper and lower values of the AUROC that lie exactly the specified distance from the empirical distribution.

In this context, distance is defined by using nonparametric test statistics for the difference between the joint distributions of the model predictions. Distance is measured as the Kullback-Leibler divergence between candidate distributions. Under this approach, the solution is characterized by a minimization of differential entropy to find the extreme values of AUROC statistics that correspond to the alternative distributions. The underlying mechanism is a recursive algorithm for reallocation of sampling weights in the empirical distribution to the closest distribution with a specified value of AUROC. The algorithm solves for a set of fixed points that characterize the distribution, from which the extreme AUROC values are calculated, which define the bounds of the forecast interval. Simulation exercises show the prediction intervals to have higher coverage rates than competing approaches that allow only for sampling variation.

Many examples of classification models are employed in the physical sciences, in economics, and in business decision-making. The AUROC statistic is often used to compare the quality of estimates from different competing models. In predictive modeling competitions, the competitors are often evaluated on the AUROC obtained on a holdout sample. In many such competitions, the winners are routinely decided on the second or third decimal. This is a true reflection of the reality in business, in which it is often the case that competing models are close competitors. It is in the interest of business decision-makers to determine the optimal investment in model-building efforts, such that the return to a more complex model is justified by the added predictive performance and the resulting value so generated. Quite often, small improvements in model performance can generate high values.

However, there is often a drop in performance when the model is used out of sample, pos-

sibly indicating a degree of overfitting. Even when measures are taken to mitigate the effects of overfitting, model performance can degrade substantially. Moreover, the performance of a model, at first use, can result in an AUROC well outside the calculated standard error bounds when used out of sample. This erodes the interpretation of a confidence interval as the set that is likely to contain the true AUROC value, when calculated this way in repeated samples. When the repeated samples are drawn from a shifting distribution, the observed AUROC will lie outside the stated interval more often than is suggested by the significance level specified in traditional methods. The approach taken here allows for some degree of variation in the distribution of variables, in addition to sampling variation arising from a fixed distribution. In order to measure this variation, an allowable distance is specified along with the classification model.

A common approach in business applications is to compare the distribution of explanatory variables until the distribution is statistically different from that in the original model build. When such a difference is found, it is often an indication that the model should be rebuilt. Such a difference can be detected by observing a statistically significant value of a statistic that follows the chi-squared distribution. In the case presented in this paper, the Kullback-Leibler information criterion (Kullback and Leibler (1951)) is used, which has advantages over the most common alternatives. In particular, it places greater weight on deviations in regions with low probability density. It also affords computational advantages, in the sense that it automatically enforces boundaries for probability weights between the unit interval. Furthermore, it also naturally offers an interpretation in terms of the maximum likelihood estimation of the distribution in question. Finally, the allowable distance between distributions is the maximum relative entropy allowable for the life of the model. This statistic is discussed in Section 3.

Even without a significant change in the distribution of classification variables, such small changes can correspond to large changes in the measured AUROC. The difference in AUROC that can be observed for a particular change in distance is shown in Section 4, along with some competing techniques. A shift in AUROC may be reduced to a statistically significant deviation from a pre-specified distribution. However, such a shift could possibly correspond to a distribution that is quite a small distance from the reference distribution. In this case, one might conclude that such a change in AUROC is quite likely under a policy of replacing the model once the distribution crosses a farther threshold. The approach pursued in this article takes this variation into account.

A good deal of literature has involved the estimation of the variance of the AUROC. A comparison of the performance of a variety of methods is presented in Section 4. The early approaches among these were based on the assumption of a particular distribution. Generally, the classification variable can be divided into two samples, corresponding to the positive and negative binary outcomes, with two separate distributions. In particular, these samples could be drawn from normal distributions, together forming a bi-normal classification model. This formulation allows for a calculation of the AUROC that implies confidence bounds that are functions of the normal CDF. These are augmented in Demidenko (2012) to account for variation in the estimates of the parameters. Another possibility is that the samples of

classification variables are drawn from exponential distributions. As in the bi-normal case, the variance of the AUROC is calculated as functions of the parameters of the exponential distributions. Hanley and McNeil (1982) presents a formulation of the variance for this model. This formulation allows for an expression that provides a more general form, in which statistics are calculated from the empirical distributions. This is further investigated in DeLong and Clarke-Pearson (1988), and is made computationally efficient in Sun and Xu (2014). Each of these approaches involve making distributional assumptions that would be unreasonable to the practitioner.

On another extreme, one could also consider an upper bound for the sampling variation in AUROC. This is calculated as the variance calculated over all possible distributions, as described in Birnbaum and Klose (1957) and Van Dantzig (1945). As the literature progresses, recent approaches emerge that impose certain assumptions about the future distribution, without imposing a rigid specification of the distribution itself. In particular, Cortes and Mohri (2004) employed an approach that restricts the distribution to the set of all distributions that imply a fixed error rate of the classification model. This affords the user a tuning parameter that can be used to consider a wider array of situations in which the model may be used in practice. This approach taken in the current paper follows this line of research, in that it allows for a particular kind of variation that drives the variability of the AUROC. To put these calculations in proper context, a deeper explanation of the AUROC statistic itself is in order.

2 Environment

2.1 Area Under the ROC Curve

The ROC curve characterizes the relationship between variables in a classification model. One one side is the categorical outcome, which, in this paper, is a binary indicator variable. The outcomes are referred to as positive or negative, according to whether the event occurs or does not. The other component is a classification variable. It is, ideally, monotonically related to the probability that the binary event will occur. The classification variable can be interpreted as a model prediction, which, for example, could include a credit score for the prediction of default. In practice, it is often the output of a classification model, which could range from standard models, such as logistic regression, to a variety of nonparametric classification algorithms and machine learning methods.

Insert brief paragraph outlining main points

The classification model can be thought about in two ways. From the modeler's perspective, there is a data set of two, variables. One is the score s or classification variable, which is often the prediction from a classification model, which could be a probability that an observation should belong to a certain class or a variable from outside, such as a credit score for predicting default probability. The other variable is a binary outcome z, which indicates the positive or negative outcome of the realization.

This dataset can be divided into two distributions of the score s, depending on the corresponding binary outcome z. The scores corresponding to the positive outcomes is represented by the series y_j for $j = 1, \ldots, n$. Similarly, the scores corresponding to the negative outcomes is represented by the series x_i for $i = 1, \ldots, m$.

It is the difference between these distributions that determines the ability to build an effective predictive model. Conversely, the act of building a classification model is akin to testing whether the distributions of x and y are different. This is the situation discussed in the literature on ranking statistics used to discriminate between distributions.

The AUROC can be interpreted as an evaluation of a pairwise comparison of the correct ordering of predictions for all pairs of predictions that could be made in the sample. Specifically, if one were to pick a pair of predictions, drawn randomly from predictions corresponding to pairs of the positive (y) and the negative (x) outcomes, the AUROC is the probability that these predictions are correctly ordered. The sample analogue is calculated as follows

$$\hat{A} = \hat{\Pr}\{y > x\} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{y_j > x_i\}}. \tag{1}$$

In this version of the statistic, the calculation is the sample analogue of the probability mass under the joint density for the independent variable x and y within the region where the model orders correctly, i.e. in above the line $y = x$.

This calculation is provided in Hanley and McNeil (1982), while an implementation of the calculation is made available in Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller (2011). This is an implementation of the approach taken in DeLong and Clarke-Pearson (1988). A more efficient method of computation is proposed by Sun and Xu (2014), which is a faster approach to calculating these quantities.

Thinking of the AUROC statistic in this way... the problem boils down to an exercise in distinguishing between distributions of scores corresponding the positive and negative outcomes. This way of framing the statistic is more appealing to practitioners. Parametric solutions are not preferred in the literature, since the classification model is likely to be much more unstructured than a parametric model would allow. A nonparametric solution is not only possible but preferred by users in industry. Regardless of the particular model employed in practice, the parametric examples can still be used to explain an important correspondence between modeling methods.

The connection to the ranking statistics in the nonparametric literature is appealing to practitioners, since it precludes the need for parametric modeling of distributions that may be highly unstructured. While the ranking statistics above are intended to compare differences in sampling distributions, the mathematical equivalence to the classification problem opens up the possibility of using these approaches to evaluate models, without imposing parametric restrictions on the distributions.

Taken together, with the equivalence of the ranking problem to the classification problem, one concludes that the variation of distributions of classification variables is of paramount importance. This is especially so, since the AUROC statistic is the sum of the volume under the joint distribution defined by the classification distributions themselves. Any variation in

these distributions can cause material variation in the value of the AUROC.

These ranking statistics also highlight the fact that the AUROC is inherently a nonparametric measure of performance. It is often used in combination with modeling techniques that are also nonparametric. In some circles in industry, there is a general distaste for parametric assumptions, particularly when not supported by the data, especially when there is access to large datasets, which can be leveraged to reveal the structure of the data with more precision. In this context, there is little justification for imposing a parametric specification for variation in distributions, when parametric distributions are not used to model the distributions themselves. In particular, the change in classification distributions can be summarized by a nonparametric distance measurement and what follows is a discussion of the relevance of such a distance.

In practice, a business user would track the performance of a model while it is in use for making business decisions. If the model is run continuously, it is prudent to take periodic measurements in AUROC. It is also worth monitoring the variability in distributions of classification variables passing through the model. Once any model moves into uncharted territory, there arises the possibility that the predictions are no longer valid, in the case that the functional form cannot be extrapolated beyond the support of the build sample. It would not be a problem, so long as the specification were correct throughout the domain of the model.

However, in the classification problem, the distributions themselves are of first order importance, since this defines the measurement of the statistic itself. A careful use of a classification model would involve a specification of the terms under which the model will be discontinued or scheduled for rebuild with new data. If it is the case that the model is planned to be used for situations in which the variables have distributions sufficiently similar to the original build distribution, then the AUROC statistic can be expected to vary accordingly, over the foreseeable lifespan of the model. It is for this reason that forecast intervals are proposed that take this variation into account. In order to quantify this variation, a distance measure is required for the calculation of the bounds of these intervals. This will be presented in the following section, after the formal specification of the modeling problem.

3 The Optimization Problem

A distribution could shift in a number of directions, a given distance from the reference distribution. For many shifts likely to occur, the AUROC may change little, while for some deviations in particular directions, the AUROC could show an extreme change. The most conservative solution, from the forecasters perspective, is the most extreme values that can be reached by shifting the distribution a specified distance.

3.1 Optimization Problem: Extreme Values of AUROC

The solution of the desired bounds of the prediction interval is the distribution that satisfies the following optimization problem. Suppose that classification variables for positive

observations have marginal distribution f and those for negative observations have marginal distribution g. Together, these have a joint distribution with weights defined by the product $\mathbf{f} \otimes \mathbf{g}$, when discretized and expressed in vector form. To find the upper bound of the prediction interval, $A^{(U)}$, this optimization problem is

$$\max_{\mathbf{u},\mathbf{v}} \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} u_i v_j I_{\{y_j > x_i\}} \tag{2}$$

subject to

$$D(\mathbf{u} \otimes \mathbf{v}, \mathbf{f} \otimes \mathbf{g}) \leq \bar{D} \tag{3}$$

$$\sum_{i=1}^{m} u_i = 1, \quad \sum_{j=1}^{n} v_j = 1 \tag{4}$$

$$\{u_i \geq 0\}_{i=1}^{m}, \quad \{v_j \geq 0\}_{j=1}^{n}. \tag{5}$$

The objective function is the expression for the area under the ROC curve as it would be calculated if the sample had classification variables of y, with distribution u, and x, with distribution v. Together, these have a joint distribution with weights defined by the product $\mathbf{u} \otimes \mathbf{v}$. The first condition constrains the distribution of the joint densities $\mathbf{u} \otimes \mathbf{v}$ and $\mathbf{f} \otimes \mathbf{g}$ to differ by the specified distance $\bar{D}$. The remaining conditions specify that the distributions are well-defined. The corresponding minimization problem results in the lower bound of $A^{(L)}$. Where the distance function $D(\cdot)$ is Kullback-Leibler divergence, this problem is equivalent to finding the highest value $A^{(U)}$ and lowest value $A^{(L)}$ such that the relative entropy is within an allowable level $\bar{D}$.

3.2 Dual Problem: Distance to a Distribution with AUROC A_0

There is, however, an indirect method of solving this problem. This is because an increase in distance allows for a larger choice set, so there is the potential for finding more extreme values of $A^{(L)}$ and $A^{(U)}$ with an increase in $\bar{D}$. There is a weakly monotonic relationship between optimized distance in A_0-space and distance in distribution space. While the original problem is to maximize distance in A_0-space, subject to a constraint on the distance allowed, the dual problem is to minimize distance in the space of distributions, subject to a constraint on the candidate AUROC.

Find the *minimum* distance from the observed distribution and a distribution that has a particular value of the AUROC.

$$\min_{\mathbf{u},\mathbf{v}} D(\mathbf{u} \otimes \mathbf{v}, \mathbf{f} \otimes \mathbf{g}) \tag{6}$$

subject to

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} u_i v_j I_{\{y_j > x_i\}} \geq A_0^{(U)} \tag{7}$$

$$\sum_{i=1}^{m} u_i = 1, \quad \sum_{j=1}^{n} v_j = 1 \tag{8}$$

$$\{u_i \geq 0\}_{i=1}^{m}, \quad \{v_j \geq 0\}_{j=1}^{n}. \tag{9}$$

Where the distance function $D(\cdot)$ is Kullback-Leibler divergence, this problem is equivalent to finding the distribution with minimum differential entropy among those with AUROC statistics no smaller than the candidate value of $A_0^{(U)}$.

The solution can be stated in terms of the derivatives of the distance functions with respect to the optimized distribution weights. This solution results in first order conditions

$$\frac{dD(\mathbf{u}, \mathbf{f})}{du_i} = \lambda \sum_{j=1}^{n} v_j I_{\{y_j > x_i\}} + \gamma_x + \delta_{x,i}, i = 1, \ldots, m \tag{10}$$

$$\frac{dD(\mathbf{v}, \mathbf{g})}{dv_j} = \lambda \sum_{i=1}^{m} u_i I_{\{y_j > x_i\}} + \gamma_y + \delta_{y,i}, j = 1, \ldots, n. \tag{11}$$

That is, the change in distance is equal to a partial AUROC term plus some extra terms for the Lagrange multipliers on the remaining constraints. Exactly how this is satisfied depends on the particular choice of distance function $D(\mathbf{u} \otimes \mathbf{v}, \mathbf{f} \otimes \mathbf{g})$.

3.3 Specification of Distance

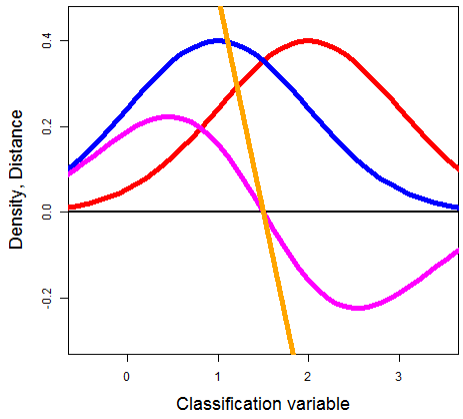
A distance function $D(\cdot)$ must be specified for calculating the distance to the distributions for calculation of the bounds on AUROC. The recommended metric is Kullback-Leibler divergence (KLD), first presented in Kullback and Leibler (1951). The original specification is not technically a distance metric, since it is not symmetric, but it still specifies a useful criterion for discriminating between distributions:

$$KLD_1(h_1, h_2) = \sum_{k=1}^{K} \left\{ h_1(t_k) \log\left(\frac{h_1(t_k)}{h_2(t_k)}\right) \right\}. \tag{12}$$

where h_1 and h_2 are two density functions summed over the index $k = 1, \ldots, K = mn$. In the objective functions of the optimization problems above, these functions are the joint densities $\mathbf{f} \otimes \mathbf{g}$ and $\mathbf{u} \otimes \mathbf{v}$, before and after optimization. In the first order conditions, the derivatives of this function have the pairs of distributions set at either $\mathbf{g}$ and $\mathbf{v}$ or $\mathbf{f}$ and $\mathbf{u}$, for the positive and negative classification variables, respectively.

With a single density term in the equation, it is implied that the distribution f_1 is a benchmark distribution, or null hypothesized distribution, under which the true probabilities

Figure 1: Calculation of Kullback-Leibler Distance



Terms in the Calculation of Kullback-Leibler Distance: Differences (magenta) and log-differences (orange) for two normal densities (red and blue). Notice that the difference is damped where density is low, while the log-difference grows linearly. This places greater emphasis on areas of support with low density.

are calculated. The same logic applies to the differential weighting on low density areas of the support, so it is also a good candidate for the specification of distance in the optimization problem.

Bigi (2003) highlights a symmetric version of this criterion, which is useful for explaining the suitability of such a measure of distance.

$$KLD_2(h_1, h_2) = \sum_{k=1}^{K} \left\{ (h_1(t_k) - h_2(t_k)) \log \left(\frac{h_1(t_k)}{h_2(t_k)} \right) \right\}. \tag{13}$$

As seen in Figure 1 above, the log-distance term takes on much more extreme values, particularly in areas where one of the distributions has low density. This is because $KLD_2(\cdot)$ is designed as a means of discriminating between two distributions. When there is an observation that occurs in a region of the support where there is very low density in one distribution and very high density in the other, there is very strong support for the hypothesis that it came from the second distribution. In the extreme case when the supports of the distributions do not overlap, there is perfect ability to discriminate between the distributions.

This feature is also useful in the sense that it will place more weight on the changes in density that are more closely related to changes in AUROC. This is concentrated in the region between the two distributions of classification variables, particularly when only the tails overlap. In this case, comparatively large changes can be made in terms of AUROC with marginal changes in densities.

There are several good reasons to choose either of these functions as the measure of distance. First of all, there is an information-theoretic justification. It adequately measures the quality of information for discriminating between pairs of distributions. Secondly, it also plays a role in the derivation of the distribution theory for the maximum likelihood estimator

(see Amemiya (1985), for example). In some versions of this derivation, the distribution is made up of two important terms. The first of which measures the information from the true distribution from which the sample is drawn. This is defined as the expected value of the log of the density, a quantity that is also defined in Kullback and Leibler (1951). This is fixed relative to the fitted distribution and any parameters that may appear in a maximum likelihood estimation procedure. The second term is the KLD_1 function evaluated at the fitted distribution and the true distribution. Thus, the KLD_1 metric is the first order term in the distribution of the fitted model from the true distribution in a maximum likelihood framework. This extends the applicability to any situation in which the researcher could specify a distribution and estimate its variability by maximum likelihood.

Still, one may put forth the suggestion that the modeling could be simplified by the use of a χ^2 calculation instead, such as the sum of the squared differences between the two joint density functions on the grid $k = 1, \ldots, K$. This specification has the drawback of assigning equal weight on equal deviations at all points in the distribution. For discriminating between distributions, this is not ideal, since it places undue weight on dense regions where a small difference is less likely to be informative. Aside from the conceptual advantage, this option suffers from a computational drawback. Computationally, this specification of distance requires additional constraints to impose non-negativity of densities when shifting distributions toward the limits of AUROC. The KLD metric guarantees an interior solution, which is the subject of the next section.

3.4 Solution

While a more complete listing of the calculations is offered in Appendix B, the solution to the distance minimization problem, with KLD_1 as the chosen distance metric, is represented by

$$u_i = \exp\left\{1 + \ln f_i + u_i \ln f_i + \lambda \sum_{j=1}^{n} v_j I_{\{y_j > x_i\}} + \gamma_x\right\} \tag{14}$$

$$v_j = \exp\left\{1 + \ln g_j + v_j \ln g_j + \lambda \sum_{i=1}^{m} u_i I_{\{y_j > x_i\}} + \gamma_y\right\} \tag{15}$$

where k_x and k_y are normalizing constants and Lagrange multiplier λ is the step size.

This nonlinear system of equations has a large number of parameters and would be difficult to solve explicitly. Instead, this system of equations is solved via a switching algorithm with the recurrence relations

$$u_i^{(t+1)} = k_x f_i^{1+u_i^{(t)}} \exp\left\{\lambda \sum_{j=1}^{n} v_j^{(t)} I_{\{y_j > x_i\}}\right\} \tag{16}$$

$$v_j^{(t+1)} = k_y g_j^{1+v_j^{(t)}} \exp\left\{\lambda \sum_{i=1}^{m} u_i^{(t)} I_{\{y_j > x_i\}}\right\} \tag{17}$$

A switching solution is effective, since the main terms in the recurrence relations are the cross-partial-AUROC expressions represented by, for example, $\sum_{j=1}^{n} v_j^{(t)} I_{\{y_j > x_i\}}$. These terms specify the direction toward the optimal distribution and the Lagrange multiplier λ specifies the step size. The log operators in the distance function enforce an interior solution, so the non-negativity constraints are automatically satisfied. At each step, the constants k_x and k_y serve to normalize to unit probability mass.

3.5 Forecast Intervals

The solution of this problem allows for the calculation of the bounds of the forecast interval. First, use the classification variables from the entire [2] build sample and measure the AUROC, $\hat{A}$. Next, measure the distance $\bar{D}$ between the distributions through some modeling approach. This is facilitated if the researcher has some knowledge of how to model the variation in the distribution. In particular, a maximum likelihood approach could be used to model this variation. This can be achieved if the modeler has a specific model in mind. However, it could also be achieved as long as the sample can be divided into several segments with known differences, such as seasonality. Then, the variation in distributions could be calculated in a series of subsamples from the empirical distribution defined by the full sample. Of course, in order to calculate $\bar{D}$ using this approach, it is necessary that the modeler is aware of a segmentation of the sample into sections that will indicate the appropriate distance. Another approach would be to re-sample from the full distribution and calculate a bootstrap version of the distance between subsamples and the full empirical distribution. Take the average of the distances between the distributions and record it as distance $\hat{D}$. Finally, search over the values of A_0 both above and below $\hat{A}$ until the distance $\hat{D}$ is achieved, searching in the direction of the partial-AUROC term of the first order conditions above. This is achieved by solving for the distance minimization problem for each candidate value of A_0 for bounds A_L and A_U, above and below $\hat{A}$, until $\bar{D} = \hat{D}$ in each direction.

It is important to consider the conditions under which the solution to the above problem exists. The required regularity condition is that the classification variable is not a perfect classifier. For example, it is sufficient that the support of the positive and negative distributions are not disjoint. If so, then one can find a distribution with any AUROC value: 1 if all weight of y is on the higher value and all weight of x is on the lower value, 0 if all weight of x is on the higher value and all weight of y is on the lower value, and anything in between for the fractions of weight in between. If there are more values in the support, there are more permutations possible.

[2] In practice, a modeler may choose a specific sample deemed appropriate for estimation of a model. This will often not be the full sample of all available observations, as the modeler would take into account the features of a sample to best represent the anticipated scenarios in future uses of the model. This sample will often be combined with several validation samples, representing a variety of situations that could be encountered through the use of the model but these would typically be used as a form of stress test and not as a benchmark for model performance. Thus, these would be left out of the sample for estimating the reference AUROC value from which the forecast interval is constructed.

4 Simulation Evidence

4.1 Data Generating Process

This next section presents the results of some simulation exercises to demonstrate the effectiveness of the statistical methodology, with $1,000$ replications in each. Each exercise will show two types of coverage rates for the AUROC. The first type, following the standard definition, are defined as the proportion of confidence intervals that contain the true values of the AUROC.

Confidence intervals represent the range of values such that, when calculated in such a way, would be expected to contain the true value of the parameter in question. The width of such an interval is calculated in a way that only accounts for the sampling variation.

The statistics that are labeled as correct forecast rates are the proportion of forecast intervals that contain the estimated values of the AUROC that are realized from the data generating process. These rates account for both the variation in the sample from a fixed population and the variation in the data generating process, which is described next.

The data generating process is a two-state regime-switching model with a bi-normal classification model in each state. That is, with equal probability, a drawing is made from the high (H) and low (L) regimes. Then a sample of data is drawn from the selected distribution, with true AUROC values of A_L in the low state and A_H in the high state. In each state, the drawing is made from a bi-normal classification model. This model is made up of a drawing from a pair of normal distributions. The observations corresponding to positive events are drawn from one normal distribution and those corresponding to the negative distribution are drawn from another. There are $1,000$ negative observations and 100 negative observations. Both normal distributions have a standard deviation of $1/\sqrt{2}$, which implies a pooled variance of 1. The classification variables corresponding to negative events in the bi-normal model have mean zero, while those linked to positive events have mean in accordance with the specified AUROC values.

The simulation is divided into four sections. In the first, the AUROC parameters are set at $A_L = 0.68$ and $A_H = 0.72$. In the second, the parameters are farther apart at $A_L = 0.65$ and $A_H = 0.75$. In the third, the AUROC parameters take on higher values at $A_L = 0.75$ and $A_H = 0.80$. The fourth simulation has no regime-switching behavior, with AUROC in both states set to 0.70.

4.2 Alternative Methods of Inference

The forecast intervals produced in this analysis are compared to several other intervals produced in the literature. The first case takes advantage of the knowledge that the model is bi-normal. In this special case, the AUROC reduces to the normal CDF evaluated at the statistic for testing the hypothesis of equality of the means of the normal distributions. The confidence interval is calculated by evaluating the CDF at two quantiles of the test statistic, with an adjustment for the nonlinear transformation, by way of the delta method. The expression is described in Demidenko (2012).

The second example is a more flexible method, in which the confidence interval is constructed using an estimate of the variance of AUROC. The expression for this estimate is calculated in Hanley and McNeil (1982), using the methods in Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller (2011), following the computational strategy in Sun and Xu (2014).

$$\sigma_A^2 = \frac{1}{mn}\{A(1-A) + (n-1)(P_{yyx} - A^2) + (m-1)(P_{yxx} - A^2)\} \tag{18}$$

The terms P_{yyx} and P_{yxx} in the above are defined as

$$P_{yyx} = \frac{1}{mnn}\sum_i\sum_j\sum_k I_{\{y_j > x_i \cap y_k > x_i\}} \qquad P_{yxx} = \frac{1}{mmn}\sum_i\sum_j\sum_k I_{\{y_j > x_i \cap y_j > x_k\}}. \tag{19}$$

The expressions are higher-order terms of the calculation of the AUROC itself. They are evaluated by drawing three observations by drawing two from either the positive or negative classification variables and one from the other. The values of P_{yyx} and P_{yxx} are the probabilities that both such pairs are correctly ordered.

The third method is calculated with a bootstrap technique, with 399 bootstrap replications for each confidence interval. These are calculated using the approach in Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller (2011), and is computed by drawing realizations of the observations, with replacement, and calculating the AUROC. The confidence intervals are then computed by using the upper and lower quantiles.

Taking an upper bound on the variance of the AUROC produces a more conservative estimate of a range that should contain the true values. The approach is described in Birnbaum and Klose (1957) and Van Dantzig (1945) and has the following form.

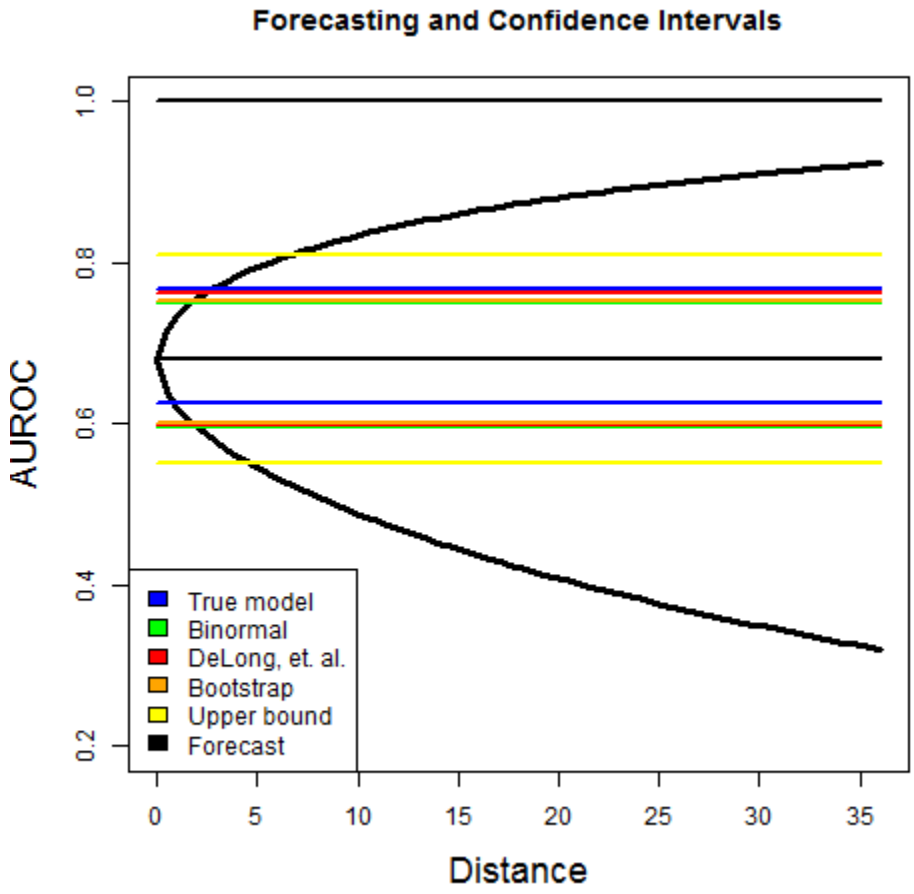
$$\sigma_{max}^2 = \frac{A(1-A)}{\min\{m,n\}}\left(\leq \frac{1}{4\min\{m,n\}}\right) \tag{20}$$

Figure 2 shows confidence intervals for the AUROC for the above list of methods, at the 95% confidence level. It is contrasted against the forecast intervals that are produced as a function of distance from the sample distribution, represented by the black curves. As the forecast intervals are scaled by the estimate of distance, it is possible that the intervals will expand to account for added variation, while others remain fixed. This allows for the potential for variation that erodes the coverage rates of the fixed intervals, while the forecast intervals are flexible to accommodate such variation.

Finally, one more approach is presented in Cortes and Mohri (2004), in which the confidence interval is parameterized as a function of the anticipated error rate. In this way, the variance can be expanded beyond the distribution represented in the sample. It is also reeled in from the upper bound drawn up by all possible distributions that could occur. This novel approach is excluded from the simulation but is mentioned since it allows for the intervals to be scaled by a parameter chosen by the modeler, as is done in the simulations that follow.

Under each of these simulations, the sample is drawn from the regime-switching model, with labels to denote the subsamples by regime. This can be interpreted as a form of seasonality that is known to the researcher, except that the features of the sample are unknown.

Figure 2: Forecast and Confidence Intervals



Forecast and Confidence Intervals: Forecast intervals are produced as a function of distance from the sample distribution (black curves) measured by Kullback-Leibler divergence. Confidence intervals are shown for the AUROC for a variety of methods, at the 95% confidence level. As the forecast intervals are scaled by the estimate of distance, it is possible that the intervals will expand to account for added variation, while others remain fixed. The sample is drawn from a bi-normal distribution with a true AUROC of 0.70, with positive classification variable distribution with a mean of 0 and standard deviation of $1/\sqrt{2}$, and negative positive classification variable distribution with a mean of 0.5244 and standard deviation of $1/\sqrt{2}$. Sample size is $1,100$ in total, with $1,000$ negative observations and 100 positive observations.

Another interpretation is that the modeler can propose a model for the distributions in each regime. This approach was not followed here, as it is intended to apply without parametric model specifications. This sort of simulation employed here is designed as a simple approach to modeling the time-varying nature of the data generating process in a way that can be easily understood.

4.3 Coverage Rates

Across the four models, the performance of the methods follows a particular pattern. In the first case, with $A_L = 0.68$ and $A_H = 0.72$, there is a modest amount of variability in the model, so the competing methods still produce reasonable coverage rates greater than 80%. Performance drops for forecast accuracy, with the specified intervals containing future estimates roughly two thirds of the time. The distance-based forecast method produces coverage greater than 90% in all cases.

In the next model, with $A_L = 0.65$ and $A_H = 0.75$, the true AUROC has much more variability. The other techniques are unable to capture this variation and coverage rates are much lower. In the third case, with $A_L = 0.75$ and $A_H = 0.80$, the true variability is lower, but the variances within each of the first three standard methods is based on a leading term of the form $A(1 - A)$, which is lower for populations with AUROC in this higher range. For this reason, these techniques underestimate the variance of the estimates and produce lower coverage rates. The exception is the method that uses an upper bound on the variance, which produces higher coverage rates with the wider intervals. In the fourth case, the AUROC is a fixed 0.70 and the competing methods fare much better. Under the Bi-normal, DeLong and Bootstrap approaches, the model is correctly specified and the statistics show results close to the nominal 95% coverage rates.

Consider the results along the list of modeling approaches considered. The AUROC is estimated from the full sample and a confidence interval is constructed around this value under each of the approaches listed above. To this list, the forecast intervals defined in this paper are proposed as an alternative. These are constructed by calculating the average distance $\bar{D}$ between the subsamples and the full samples. These distances are used to shift the distributions of classification variables to generate the farthest values of AUROC that can be generated with a distribution of distance $\bar{D}$ from the sample observed. The extreme values of AUROC define the bounds of the forecast interval.

Overall, the distance-based forecasting method produces acceptable performance in all cases, as a method for constructing forecast intervals. The closest competitor is the technique using an upper bound on the variance, which is enough to capture the variability in estimated AUROC, at least for the cases with less variability. The distance-based forecast method is the only one that scales the width of the interval to account for additional variation.

Table 1 shows the coverage rates for the four models under five approaches for calculating confidence intervals. In the third column of coverage rates, the 'Forecast' technique presented here shows coverage rates from 99% to 100%. That is, in nearly every realization, the forecast interval was shown to contain the true value of the realized parameter in the regime switching model. This is expected, since the method is designed to produce forecast intervals, which likely contain the realized estimate. In the next column, the forecast interval for this method contains the realized estimate of the AUROC from 92 to 97 percent of the realizations.

Compare this to the other methods currently available. The bi-normal method is designed to impose true restrictions on the model, which will provide more accurate confidence intervals when those restrictions are true. In three of the four models considered, these assumptions are violated and the coverage rates are far below the nominal 95% level. The

Table 1: Rates of Coverage and Correct Forecasts

True Values	Method	Coverage Rate	Correct Forecast Rate
$A_L = 0.68$,	Bi-normal	0.8265	0.6612
$A_H = 0.72$	DeLong et. al.	0.8395	0.6689
	Bootstrap	0.8310	0.6616
	Upper Bound	0.9885	0.9074
	Forecast	0.9955	0.9600
$A_L = 0.65$,	Bi-normal	0.2360	0.3427
$A_H = 0.75$	DeLong et. al.	0.2470	0.3515
	Bootstrap	0.2455	0.3453
	Upper Bound	0.7725	0.6615
	Forecast	0.9795	0.9259
$A_L = 0.75$,	Bi-normal	0.6805	0.5912
$A_H = 0.80$	DeLong et. al.	0.6845	0.5979
	Bootstrap	0.6730	0.5889
	Upper Bound	0.9665	0.8654
	Forecast	0.9940	0.9451
$A_L = 0.70$,	Bi-normal	0.951	0.7402
$A_H = 0.70$	DeLong et. al.	0.944	0.7477
	Bootstrap	0.941	0.7386
	Upper Bound	1.000	0.9464
	Forecast	0.999	0.9702

Rates of Coverage and Correct Forecasts: Coverage rates are the proportion of confidence intervals that contain the true values of the AUROC. Correct forecast rates are the proportion of confidence intervals that contain the estimated values of the AUROC. Number of replications is $1,000$ for all models, with 399 bootstrap replications for bootstrap confidence intervals. The data generating process is a two-state regime-switching model with a bi-normal classification model in each state, with true AUROC values of A_L in the low state and A_H in the high state.

exception is in the fourth model, with no regime switching, with coverage rates as expected, since the model is correctly specified. Realized estimates of AUROC are only more variable, so the coverage rates are lower, as the estimate will often stray beyond the range expected for the true parameter.

The confidence intervals labeled 'DeLong et.al.' and those defined using the bootstrap techniques produce similar results. They are both built using a process that draws from the empirical distribution of classification variables. The first approach does so by enumerating all the combinations, while the bootstrap technique draws samples of observations randomly. In both cases, the coverage rates range from 25 to 85 percent, aside from the fourth model, in which the model is correctly specified. These methods achieve the nominal coverage rates of 95% for coverage of the true parameters. This rate drops when considering the outcome of the estimation of the AUROC statistic, which is reasonable as this variation is outside of that assumed under the construction of the intervals.

5 Conclusion

The technique presented in this paper has answered a question of interest to the applied researcher. Practitioners have long been aware of the statistical techniques for estimating the variance of an AUROC statistic from a fixed population. The applicability in practice has, however, been less fruitful. In the practical application of such empirical techniques, one often finds a substantial increase in variability, compared to that expected from the existing approaches. When one takes into account the variation induced by changes in the population, the existing techniques fall short and the distance-based forecast interval method fills this requirement.

Practitioners now have a method that will give a reasonable indication of the drop in AUROC that may be experienced from a shift in the population. This technique can also be used to provide a more realistic assessment of the performance expected from competing models. A difference between competing models may be statistically significant for the fixed population under study but one may find that a performance difference between models is much less important than a change in the population. The practitioner might be surprised by a performance change in the application of a model, but this work should serve to mitigate the element of surprise. The key question that is answered refers to the performance of *future* applications of the model.

This piece of work opens up several avenues of future research. The first of which employs the use of the bootstrap. The existing bootstrap technique considered in this paper, presented in Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller (2011), is a simple approach to simulating the estimated AUROC. By resampling from the observed sample, the variation is not only limited to sampling variation but it is also limited to the data observed from the data generating process. It does not, however, impose a null hypothesis that may be the focus of a test.

Using the distance-based approach presented here, the researcher can solve for the closest distribution that satisfies a null hypothesis of the form $H_0 : A = A_0$. Resampling methods can then be applied to this re-weighted distribution, which will produce results that correspond to a data generating process under the null hypothesis. This can then be used to calculate test statistics of the form presented in Davidson (2008), for example. Davidson (2008) can easilty be extended to the AUROC, since it is designed for the Gini index, which is mathematically similar to one calculation method for the AUROC. The test statistic can be simulated under the null hypothesis and the researcher would reject the null if the estimated values from the observed sample appear unlikely under the null hypothesis.

This approach can also be extended to multivariate comparisons between competing classification models. A null hypothesis of equal model performance can be imposed and one can solve for the closest pair of distributions to those observed. With resampling done in this way, the variability in the difference in performance in each test can be similarly used to evaluate the performance in the observed sample. This sort of analysis would serve as the distance-based version of the tests discussed in Hanley and McNeil (1983), in which pairs of classification models are considered. Such comparisons can now be used to anticipate relative performance levels in future applications of the model.

While the use of the AUROC began as a tool for analyzing radio transmissions for military applications, the competition is no less intense in business. Classification models are used in a variety of business applications and their performance is closely scrutinized. The forecasting approach presented here will bring about improvements in the way classification methods are used in business decisions. The ability to use these methods effectively will determine the winners on the industrial battlefield.

References

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, MA.

Bamber, D. (1947): "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, 12(12), 387–415.

Bigi, B. (2003): *Advances in Information Retrieval: 25th European Conference on IR Research, ECIR 2003*chap. Using Kullback-Leibler Distance for Text Categorization, pp. 305–319. Springer-Verlag, New York.

Birnbaum, Z., and O. Klose (1957): "Bounds for the Variance of the Mann-Whitney Statistic," *Annals of Mathemtical Statistics*, 38.

Cortes, C., and M. Mohri (2004): "Confidence Intervals for the Area under the ROC Curve," *Advances in Neural Information Processing Systems*, 17, 305–312.

Davidson, R. (2008): "Reliable Inference for the Gini Index," Discussion paper, Department of Economics and CIREQ, McGill University.

DeLong, E., D. D., and D. Clarke-Pearson (1988): "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, pp. 837–845.

Demidenko, E. (2012): "Confidence intervals and bands for the binormal ROC curve revisited," *Journal of Applied Statistics*, 39, 67–79.

Hanley, J. (1989): "Receiver operating characheristic (ROC) methodology: the state of the art," *Critical Reviews in Diagnostic Imaging*, 29(3), 307.

Hanley, J., and B. McNeil (1982): "The Meaning and Use of the Area under a Receiver Operating (ROC) Curve," *Radiology*, 143, 29–36.

——— (1983): "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases," *Radiology*, 148, 839–843.

Kendall, M., and J. Gibbons (1990): *Rank Correlation Methods*. Arnold, London.

Kullback, S., and R. A. Leibler (1951): "On Information and Sufficiency," *Ann. Math. Statist.*, 22(1), 79–86.

Lehmann, E., and H. D'Abrera (2006): *Nonparametrics: statistical methods based on ranks*. Springer, New York, NY.

Mann, H., and D. Whitney (1947): "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, 18(1), 50–60.

Newson, R. (2006): "Confidence intervals for rank statistics: Somers' D and extensions," *The Stata Journal*, 6(3), 309–334.

Peterson, W, B. T., and W. Fox (1954): "The theory of signal detectability," *IRE Professional Group on Information Theory*, 4(4), 171–212.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011): "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, 12, 77.

Somers, R. H. (1962): "A new asymmetric measure of association for ordinal variables," *American Sociological Review*, 27, 799–811.

Sun, X., and W. Xu (2014): "Fast Implementation of Delong's Algorithm for Comparing the Areas under Correlated Receiver Operating Characteristic Curves," *IEEE Signal Processing Letters*.

Van Dantzig, D. (1945): "On the Consistency and Power of Wilcoxon's Two Sample Test," *Koninklijke Nederlandse Akademie van Wetershappen, Series A*, 54.

Van Meter, D., and D. Middleton (1954): "Modern statistical approaches to reception in communication theory," *IRE Professional Group on Information Theory*, 4(4), 119–145.

Wald, A. (1949): "Statistical Decision Functions," *The Annals of Mathematical Statistics*, pp. 165–205.

Wilcoxon, F. (1945): "Individual comparisons by ranking methods," *Biometrics Bulletin*, 1(6), 80–83.

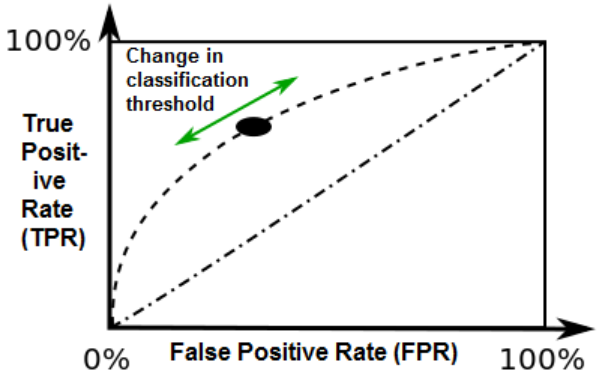
Appendix A Area Under the ROC Curve

The ROC curve is a parametric curve, plotting the true positive rate against the false positive rate. The curve is parameterized by the discrimination threshold, the value of the classification variable, above which observations are classified as positive. A particular point on this curve shows the performance of the classification, plotting the observations correctly classified as positive, against those that are classified as positive but are truly negative.

The result can be used as a measurement of the frequency of correct interpretation of a signal for predicting a binary element of a message. In fact, the use of this curve for a measure of signal quality dates back to the measurement of the quality of radio transmissions for military use. Early academic references include Peterson and Fox (1954), Van Meter and Middleton (1954) and Wald (1949) for uses in this context. A later review can be found in Hanley (1989), in which the statistic is used to assess the quality of medical diagnoses. Recently, this curve has been used by the business community to assess the discriminatory power of variables used in classification models for predictive modeling.

Figure 3 shows an example of the Receiver Operating Characteristic curve. It plots the True Positive Rate (TPR) on the vertical axis against the False Positive Rate (FPR) on the horizontal axis. These proportions are calculated for a particular value of the classification threshold, which the user can vary to make a classification more selective or more inclusive. The observations with classification variable above the classification threshold are classified as positive and the rates are then calculated according to whether that classification is correct in each case. In the bottom left corner, only the strongest signals are used to identify positive outcomes and, for an effective model, the ratio of correct to incorrect positive identifications should be greater than equal, implying a steeper slope. As the threshold is increased, progressively weaker signals are used to identify positive outcomes and the performance should decline, resulting in a declining slope. At the other extreme, the classification threshold will include nearly the entire sample, for which the accuracy will have declined enough to coincide with the sample-wide event rates.

Figure 3: Receiver Operating Characteristic Curve



Receiver Operating Characteristic Curve: True Positive Rate vs. False Positive Rate, for a particular value of the classification threshold. Observations with classification variables above the classification threshold are classified as positive. The true positive rate is the proportion of positive values of the classification that are correctly classified as negative, i.e. that lie above the threshold. Conversely, the true negative rate is the proportion of negative values of the classification that are incorrectly classified as positive, also above the threshold. The ROC curve is the parametric curve that is followed as the classification threshold varies from the highest to the lowest values in the sample.

The better the model, the steeper will be the initial slope and the closer the curve will be to the upper boundaries. The area under this curve is the statistic that indicates the

performance value of the classification variable. The diagonal line indicates the classifications that are no better than random allocation to positives and negatives. In this locus, the event rate is simply the average event rate for the entire sample. A model with a curve below this line is not effective in classifying, unless used in reverse order, when the curve will then lie within the upper triangular region.

The area under this curve can be computed directly by integration:

$$A = \int_{-\infty}^{\infty} TPR(t)[-FPR'(t)]dt, \tag{21}$$

in which the true and false positive rates are defined as

$$TPR(t) = \int_{t}^{\infty} g_y(t)dt, \quad FPR(t) = \int_{t}^{\infty} f_x(t)dt. \tag{22}$$

The observations of the classification variable corresponding to positive outcomes is denoted by y and that for negative outcomes is denoted by x. The functions $g_y(\cdot)$ and $f_x(\cdot)$ are the corresponding density functions. A brief series of calculations will reveal this to be equivalent to the following probability

$$A = Prob\{y > x\}. \tag{23}$$

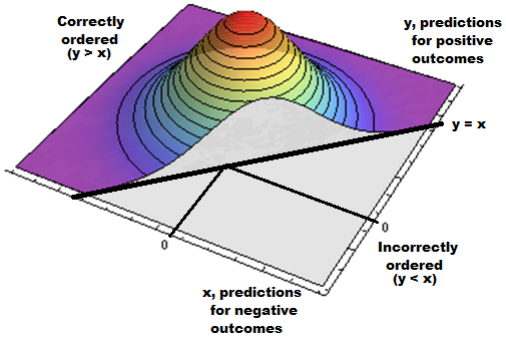
This can be interpreted as an evaluation of a pairwise comparison of the correct ordering of predictions for all pairs of predictions that could be made in the sample. Specifically, if one were to pick a pair of predictions, drawn randomly from predictions corresponding to pairs of the positive (y) and the negative (x) outcomes, the AUROC is the probability that these predictions are correctly ordered. The sample analogue is calculated as follows

$$\hat{A} = \hat{\Pr}\{y > x\} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{y_j > x_i\}}. \tag{24}$$

This calculation is provided in Hanley and McNeil (1982), while an implementation of the calculation is made available in Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller (2011). This is an implementation of the approach taken in DeLong and Clarke-Pearson (1988). A more efficient method of computation is proposed by Sun and Xu (2014), which is a faster approach to calculating these quantities.

A more intuitive approach to the calculation of the AUROC is to visualize it as the probability mass that satisfies a certain condition. Consider the following joint distribution formed by the two groups of classification variables. Combine the marginal distributions of the classification variables corresponding to the positive and negative outcomes by simply multiplying the marginal distributions. The resulting joint density will represent the probability density of all possible (x, y) pairs from the sample. Now restrict this space to the region defined by the indicator in equation (24), which is the region above the $y = x$ line. The area under the receiver operating curve is the volume under this "receiver operating surface" shown in Figure 4.

Figure 4: Volume Under the "ROC Surface"



Volume Under the "ROC Surface": With the AUROC defined in the form $A = Prob\{y > x\}$, it is clear that it can be visualized as the volume of the probability mass under the surface of the joint distribution of classification variables from positive and negative outcomes.

A.1 Examples

Figure 4 also illustrates the calculation of the AUROC for the specific case of the bi-normal model. The calculation is known to be equivalent to the normal CDF evaluated at the value of a test statistic for testing the difference between the two normal distributions. Similarly, the variability of the estimate results in confidence bounds that are also calculated by the normal CDF, which is specified in Demidenko (2012). However, it is noteworthy that the AUROC statistic is a function of the test statistic for a test of the difference between the pair of distributions of classification variables.

Another possibility is the bi-exponential model, in which both sets of classification variables are drawn from exponential distributions. This alternative is featured in Hanley and McNeil (1982) for a calculation of the variance of the AUROC. This approach allows for a more generic formulation that is calculated directly from the empirical distributions of the classification variables from both the positive and negative outcomes.

Features of these two examples can be extended to the study of generic models for classification. However, parametric solutions are not preferred in the literature, since the classification model is likely to be much more unstructured than a parametric model would allow. A nonparametric solution is not only possible but preferred by users in industry. Regardless of the particular model employed in practice, the parametric examples can still be used to explain an important correspondence between modeling methods. The mapping presented next is generalized to all classification models.

A.2 Equivalence to Ranking Statistics

There is an important similarity between the act of creating a classification model and that of comparing two distributions. A difference between the distributions is a necessary condition for a classification variable to have predictive power for the binary outcome. Conversely, if the classification variable produces predictions with an AUROC above 0.5, then the distri-

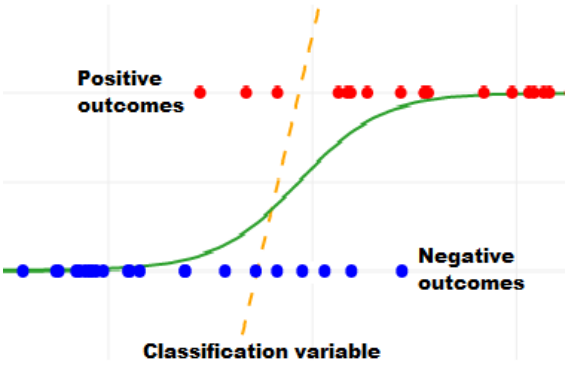


Figure 5: Predictive value of classification variables

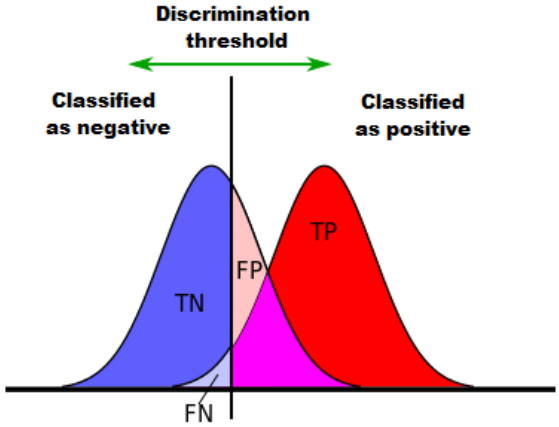


Figure 6: Difference in the distributions of variables

butions must be different. For this reason, an important connection can be made between nonparametric statistics and the AUROC, and this connection is useful for finding ways to further characterize the AUROC.

Because of this close connection between statistical methodologies, a number of methods in the nonparametric literature are useful for analyzing the AUROC. Nonparametric ranking statistics are described in Lehmann and D'Abrera (2006) and Kendall and Gibbons (1990), in reference to the literature on the comparison of distributions. One method that stands out as especially relevant is the Mann-Whitney U-statistic, which is presented in Mann and Whitney (1947), for comparing two distributions nonparametrically,

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{y_j > x_i\}}. \tag{25}$$

This is the same as AUROC, without the normalization by the count of pairs mn.

This equivalence is leveraged for computational advantage, since the U-statistic is calculated as follows. First, sort the observations in increasing order. Then, assign ranks to the observations, with 1 for the smallest score and the sample size for the observation with the largest score. Finally, take the sum of the ranks corresponding to the positive outcomes and subtract half of the quantity $n_1(n_1 + 1)$, where n_1 is the number of positive observations in the sample. A similar statistic was proposed in Wilcoxon (1945) for similar purposes. Bamber (1947) made the connection between the AUROC and the Mann-Whitney U-statistic. Other approaches for evaluating classification models are similarly inspired by tools such as Somers' d (Somers (1962)) and Kendall's τ (see Kendall and Gibbons (1990)) A simulation-based approach to modeling the variability of these statistics is found in Newson (2006), which involved jackknifing the numerator and denominator of the ratio in Kendall's τ and taking a Taylor expansion.

Appendix B Distance to Distribution with AUROC A_0

The objective is to minimize distance $D(\mathbf{u}\otimes\mathbf{v}, \mathbf{f}\otimes\mathbf{g})$, which could be either of the functions $CHI(h_1, h_2)$ or $KLD(h_1, h_2)$ or any other measure of distance between distributions, such that the chosen distribution corresponds to a specified AUROC statistic. The optimization problem is formalized as follows.

$$\min_{\mathbf{u},\mathbf{v}} D(\mathbf{u}\otimes\mathbf{f}, \mathbf{v}\otimes\mathbf{g}) \tag{26}$$

subject to

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} u_i v_j I_{\{y_j > x_i\}} = A_0, \tag{27}$$

and the conditions on the probabilities

$$\sum_{i=1}^{m} u_i = 1, \quad \sum_{j=1}^{n} v_j = 1, \quad \{u_i \geq 0\}_{i=1}^{m}, \quad \{v_j \geq 0\}_{j=1}^{n}. \tag{28}$$

Stated as a Lagrange optimization problem, this becomes

$$\min_{\mathbf{u},\mathbf{v}} Q(\mathbf{u}, \mathbf{v}, \mathbf{f}, \mathbf{g}, \gamma_x, \gamma_y, \delta_x, \delta_y) \tag{29}$$

$$= \min_{\mathbf{u},\mathbf{v}} D(\mathbf{u}\otimes\mathbf{f}, \mathbf{v}\otimes\mathbf{g}) \tag{30}$$

$$- \lambda\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} u_i v_j I_{\{y_j > x_i\}} - A_0\right] \tag{31}$$

$$- \gamma_x\left[\sum_{i=1}^{m} u_i - 1\right] - \gamma_y\left[\sum_{j=1}^{n} v_j - 1\right] - \sum_{i=1}^{m}\delta_{x,i}u_i - \sum_{j=1}^{n}\delta_{y,j}v_j \tag{32}$$

The first order conditions for this problem are

$$\frac{dD(\mathbf{u},\mathbf{f})}{du_i} = \lambda\sum_{j=1}^{n} v_j I_{\{y_j > x_i\}} + \gamma_x + \delta_{x,i}, i = 1, \ldots, m, \tag{33}$$

and

$$\frac{dD(\mathbf{v},\mathbf{g})}{dv_j} = \lambda\sum_{i=1}^{m} u_i I_{\{y_j > x_i\}} + \gamma_y + \delta_{y,i}, j = 1, \ldots, n. \tag{34}$$

For $D(f_1, f_2) = KLD_1(f_1, f_2)$ the distance function has terms of the form

$$D(\mathbf{u}_i, \mathbf{f}_i) = u_i \ln\left\{\frac{u_i}{f_i}\right\} \tag{35}$$

so the derivative term is

$$\frac{dD(\mathbf{u},\mathbf{f})}{du_i} = \ln\left\{\frac{u_i}{f_i}\right\} - u_i \ln f_i - 1. \tag{36}$$

Note that the nonnegativity constraints are not binding in the case of the Kullback-Leibler metric, since the probability weights in $\mathbf{u}$ and $\mathbf{v}$ appear within the natural log function. Setting these to zero would result in an infinite distance, meaning that they provide absolute information for discriminating between distributions. As a result, the Lagrange multipliers $\{\delta_{x,i}\}_{i=1}^{m}$ and $\{\delta_{y,j}\}_{j=1}^{n}$ are all equal to zero.

The first order conditions can be solved to isolate u_i and v_j so that they satisfy the following fixed points.

$$u_i = \exp\left\{1 + \ln f_i + u_i \ln f_i + \lambda \sum_{j=1}^{n} v_j I_{\{y_j > x_i\}} + \gamma_x\right\}, \tag{37}$$

and

$$v_j = \exp\left\{1 + \ln g_j + v_j \ln g_j + \lambda \sum_{i=1}^{m} u_i I_{\{y_j > x_i\}} + \gamma_y\right\}. \tag{38}$$

Note that this system of equations has dimension equal to the respective sample sizes, which could make the solution prohibitively costly to compute. To this end, one can iterate on the fixed point implied by the above:

$$u_i^{(t+1)} = k_x f_i^{1+u_i^{(t)}} \exp\left\{\lambda \sum_{j=1}^{n} v_j^{(t)} I_{\{y_j > x_i\}}\right\}, \tag{39}$$

and

$$v_j^{(t+1)} = k_y g_j^{1+v_j^{(t)}} \exp\left\{\lambda \sum_{i=1}^{m} u_i^{(t)} I_{\{y_j > x_i\}}\right\}, \tag{40}$$

where the constants k_x and k_y serve to normalize to unit probability mass.

Finally, the λ could, in principle, be solved for numerically, on each iteration. However, it is sufficient to choose λ so that it is small enough not to overstep the root of the fixed point. In addition, λ determines the direction of the iterations, toward either higher or lower A_0. To this end, it is sufficient to set the step size as

$$\lambda = \eta(\hat{A}^{(t)} - A_0), \tag{41}$$

for a small η, so that the step size is declining at each iteration, as the AUROC $\hat{A}^{(t)}$ moves in the direction of the specified value of A_0. The iterations continue until the first order conditions are satisfied and the distance $|\hat{A}^{(t)} - A_0|$ is near zero, such that both conditions are met, up to a chosen tolerance.