measuring performance in classification models

classification models usually generate two types of predictions.

classification models produce a continuous valued prediction, which is usually in the form of a probability (i.e., the predicted values of class membership for any individual sample are between 0 and 1 and sum to 1).

in addition to a continuous prediction, classification models generate a predicted class, which comes in the form of a discrete category.

for most practical applications, a discrete category prediction is required in order to make a decision

although classification models produce both of these types of predictions, often the focus is on the discrete prediction rather than the continuous prediction.

the probability estimates for each class can be very useful for gauging the model's confidence about the predicted classification

> returning to the spam e-mail filter example, an e-mail message with a predicted probability of being spam of 0.51 would be classified the same as a message with a predicted probability of being spam of 0.99
>
> while both messages would be treated the same by the filter, we would have more confidence that the second message was truly spam

in some applications, the desired outcome is the predicted class probabilities which are then used as inputs for other calculations

most classification models generate predicted class probabilities

when some models are used for classification, like neural networks and partial least squares, they produce continuous predictions that do not follow the definition of a probability-the predicted values are not necessarily between 0 and 1 and do not sum to 1

for example, a partial least squares classification model would create 0/1 dummy variables for each class and simultaneously model these values as a function of the predictors.

when samples are predicted, the model predictions are not guaranteed to be within 0 and 1.

for classification models like these, a transformation must be used to coerce the predictions into probability-like” values so that they can be interpreted and used for classification.

one transformation method is the softmax transformation which is defined as

$$\hat{p}_\ell^* = \frac{e^{\hat{y}_\ell}}{\sum_{l=1}^{C} e^{\hat{y}_l}}$$

where $\hat{y}_\ell$ is the numeric model prediction for the ℓ^{th} class and $\hat{p}_\ell^*$ is the transformed value between 0 and 1

suppose that an outcome has three classes and that a PLS model predicts values of $\hat{y}_1 = 0.25$, $\hat{y}_2 = 0.76$, and $\hat{y}_3 = -0.1$. The softmax function would transform these values to $\hat{p}_1 = 0.30$, $\hat{p}_2 = 0.49$, and $\hat{p}_3 = 0.21$. To be clear, no probability statement is being created by this transformation; it merely ensures that the predictions have the same mathematical qualities as probabilities.

well-calibrated probabilities

whether a classification model is used to predict spam e-mail or customer lifetime value calculations, we desire that the estimated class probabilities are reflective of the true underlying probability of the sample

the predicted class probability (or probability-like value) needs to be well-calibrated

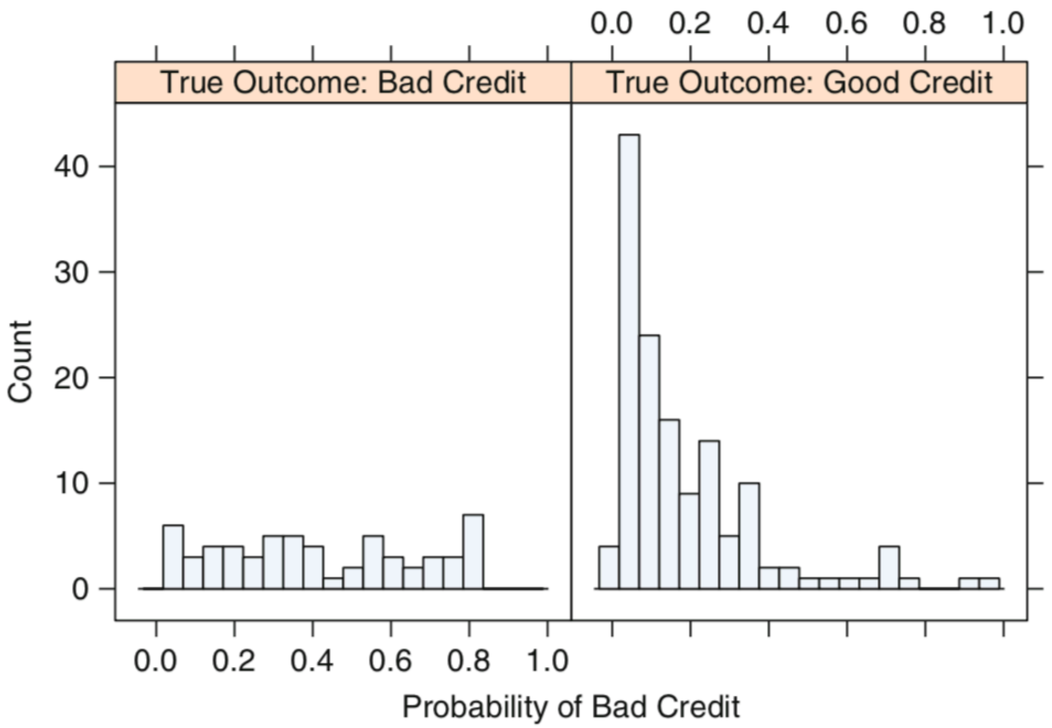
to be well-calibrated, the probabilities must effectively reflect the true likelihood of the event of interest. Returning to the spam filter illustration, if a model produces a probability or probability-like value of 20% for the likelihood of a particular e-mail to be spam, then this value would be well-calibrated if similar types of messages would truly be from that class on average in 1 of 5 samples.

presenting class probabilities

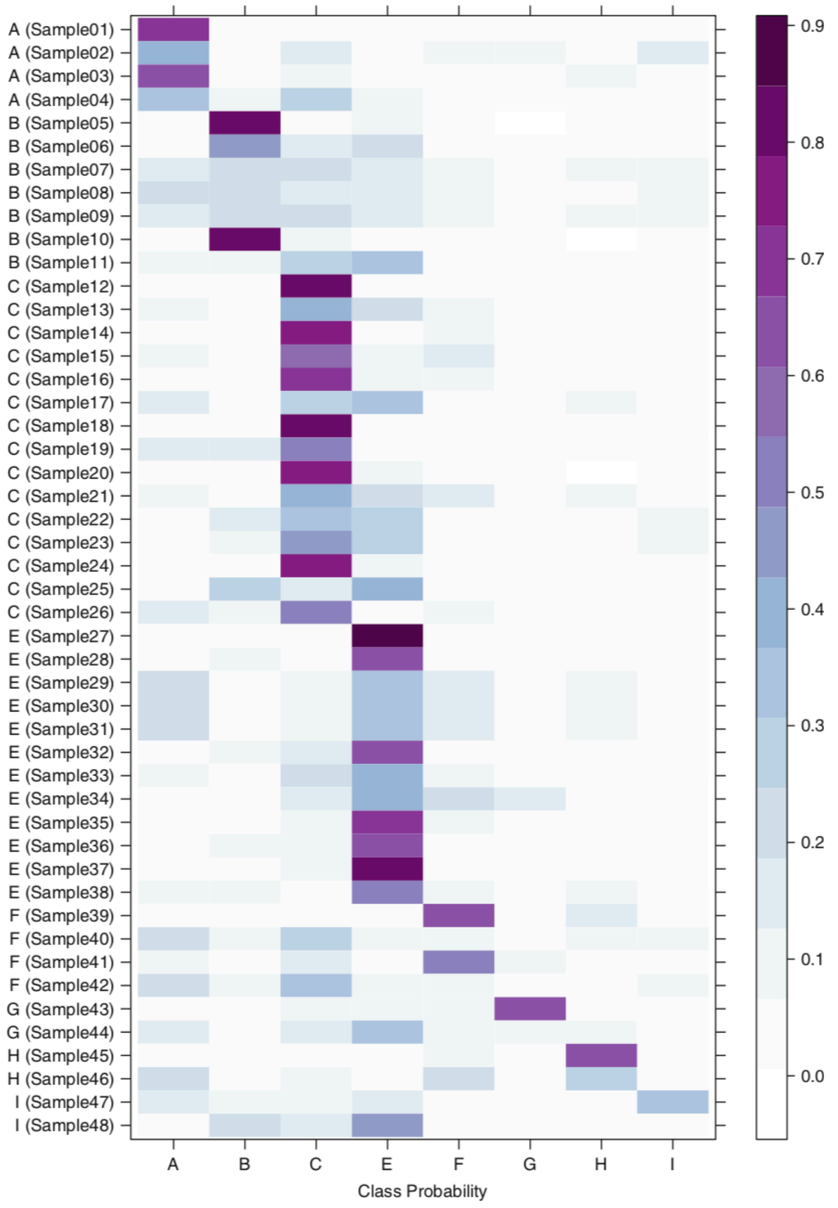
visualizations of the class probabilities are an effective method of communicating model results.

for two classes, histograms of the predicted classes for each of the true outcomes illustrate the strengths and weaknesses of a model

histograms of the test set probabilities for the logistic regression model (the panels indicate the true credit status). The probability of bad credit for the customers with good credit shows a skewed distribution where most customers' probabilities are quite low. In contrast, the probabilities for the customers with bad credit are flat (or uniformly distributed), reflecting the model's inability to distinguish bad credit cases.



when there are three or more classes, a heat map of the class probabilities can help gauge the confidence in the predictions



equivocal zones

an approach to improving classification performance is to create an equivocal or indeterminate zone where the class is not formally predicted when the confidence is not high.

model performance would be calculated excluding the samples in the indeterminate zone

the equivocal rate should also be reported with the performance so that the rate of unpredicted results is well understood

equivocal zones

for a two-class problem that is nearly balanced in the response, the equivocal zone could be defined as 0.50±z.

if z were 0.10, then samples with prediction probabilities between 0.40 and 0.60 would be called “equivocal.”

for data sets with more than 2 classes ($C > 2$), similar thresholds can be applied where the largest class probability must be larger than $(1/C)+ z$ to make a definitive prediction.

evaluating predicted classes

a common method for describing the performance of a classification model is the confusion matrix.

this is a simple cross-tabulation of the observed and predicted classes for the data. the confusion matrix for the two-class problem. the table cells indicate number of the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)

Predicted	Observed	
	Event	Nonevent
Event	TP	FP
Nonevent	FN	TN

evaluating predicted classes

the simplest metric is the overall accuracy rate (or, for pessimists, the error rate)

this reflects the agreement between the observed and predicted classes and has the most straightforward interpretation

there are a few disadvantages to using this statistic.

- overall accuracy counts make no distinction about the type of errors being made

- one must consider the natural frequencies of each class

what benchmark accuracy rate should be used to determine whether a model is performing adequately?

the no-information rate is the accuracy rate that can be achieved without a model. There are various ways to define this rate

for a data set with C classes, the simplest definition, based on pure randomness, is 1/C. However, this does not take into account the relative frequencies of the classes in the training set

an alternate definition of the no-information rate is the percentage of the largest class in the training set

models with accuracy greater than this rate might be considered reasonable

rather than calculate the overall accuracy and compare it to the no-information rate, other metrics can be used that take into account the class distributions of the training set samples.

the Kappa statistic (also known as Cohen's Kappa) was originally designed to assess the agreement between two raters

Kappa takes into account the accuracy that would be generated simply by chance. the form of the statistic is

$$Kappa = \frac{O - E}{1 - E}$$

where O is the observed accuracy and E is the expected accuracy based on the marginal totals of the confusion matrix

kappa can take on values between –1 and 1

a value of 0 means there is no agreement between the observed and predicted classes, while a value of 1 indicates perfect concordance of the model prediction and the observed classes

negative values indicate that the prediction is in the opposite direction of the truth, but large negative values seldom occur

when the class distributions are equivalent, overall accuracy and Kappa are proportional

depending on the context, Kappa values within 0.30 to 0.50 indicate reasonable agreement

suppose the accuracy for a model is high (90%) but the expected accuracy is also high (85%), the Kappa statistic would show moderate agreement (Kappa = 1/3) between the observed and predicted classes

the Kappa statistic can also be extended to evaluate concordance in problems with more than two classes

when there is a natural ordering to the classes (e.g., “low,” “medium,” and “high”), an alternate form of the statistic called weighted Kappa can be used to enact more substantial penalties on errors that are further away from the true result

a “low” sample erroneously predicted as “high” would reduce the Kappa statistic more than an error were “low” was predicted to be “medium.”

two-class problems

for two classes, there are additional statistics that may be relevant when one class is interpreted as the event of interest

the sensitivity of the model is the rate that the event of interest is predicted correctly for all samples having the event, or

$$Sensitivity = \frac{\#\text{ samples with the event } \textit{and} \text{ predicted to have the event}}{\#\text{ samples having the event}}$$

the sensitivity is sometimes considered the true positive rate since it measures the accuracy in the event population.

two-class problems

conversely, the specificity is defined as the rate that nonevent samples are predicted as nonevents, or

$$Specificity = \frac{\#\text{ samples without the event } and \text{ predicted as nonevents}}{\#\text{ samples without the event}}$$

the false-positive rate is defined as one minus the specificity.

assuming a fixed level of accuracy for the model, there is typically a trade-off to be made between the sensitivity and specificity.

intuitively, increasing the sensitivity of a model is likely to incur a loss of specificity, since more samples are being predicted as events.

potential trade-offs between sensitivity and specificity may be appropriate when there are different penalties associated with each type of error

the most common method for combining sensitivity and specificity into a single value uses the receiver operating characteristic (ROC) curve

receiving operating (ROC) curves

ROC curves were designed as a general method that, given a collection of continuous data points, determine an effective threshold such that values above the threshold are indicative of a specific event.

we describe how the ROC curve can be used for determining alternate cutoffs for class probabilities

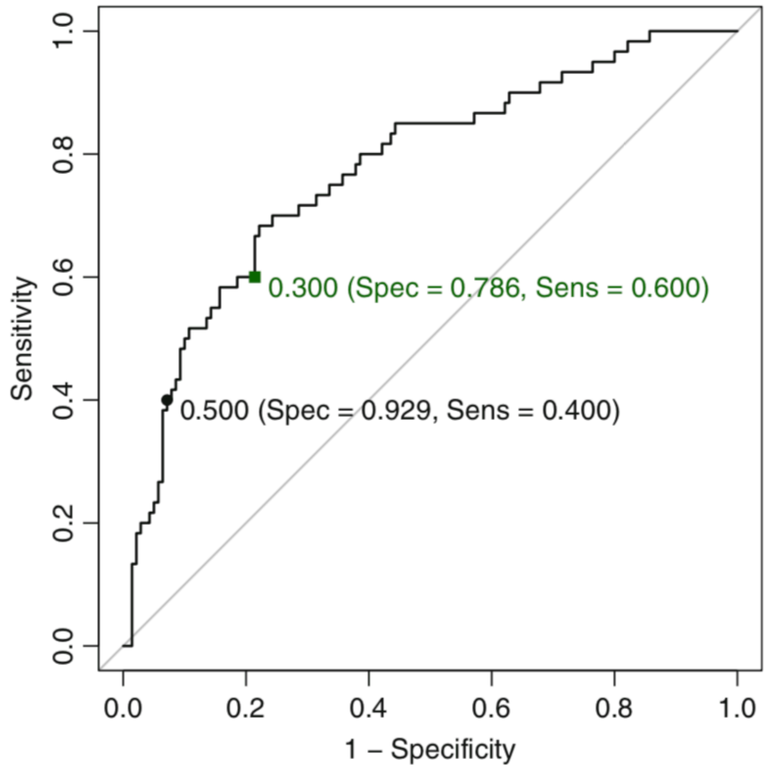
for the credit model test set previously discussed, the sensitivity was poor for the logistic regression model (40%), while the specificity was fairly high (92.9%).

these values were calculated from classes that were determined with the default 50% probability threshold

can we improve the sensitivity by lowering the threshold to capture more true positives?

the ROC curve is created by evaluating the class probabilities for the model across a continuum of thresholds.

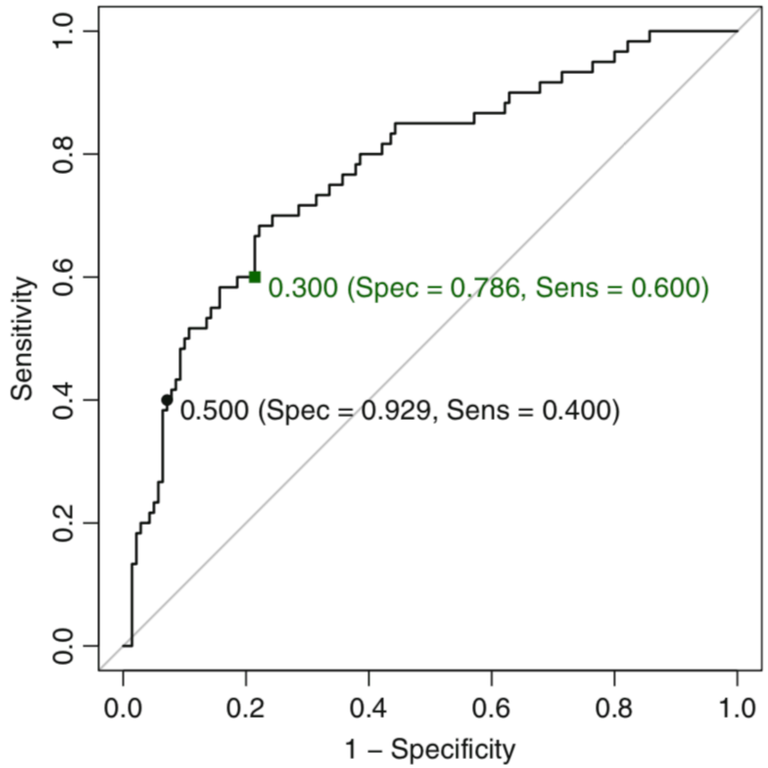
for each candidate threshold, the resulting true-positive rate (i.e., the sensitivity) and the false-positive rate (one minus the specificity) are plotted against each other.



can we improve the sensitivity by lowering the threshold to capture more true positives?

lowering the threshold for classifying bad credit to 30% results in a model with improved sensitivity (60%) but decrease specificity (79.3%)

in the plot we see that decreasing the threshold begins to capture more of the customers with bad credit but also begins to encroach on the bulk of the customers with good credit



ROC Curves

this plot is a helpful tool for choosing a threshold that appropriately maximizes the trade-off between sensitivity and specificity

altering the threshold only has the effect of making samples more positive or negative

in the confusion matrix, it cannot move samples out of both off-diagonal table cells.

there is almost always a decrease in either sensitivity or specificity as one is increased

the ROC curve can also be used for a quantitative assessment of the model

a perfect model that completely separates the two classes would have 100% sensitivity and specificity.

visually, the ROC curve would be a single step between (0, 0) and (0, 1) and remain constant from (0, 1) to (1, 1). the area under the ROC curve for such a model would be one

a completely ineffective model would result in an ROC curve that closely follows the 45° diagonal line and would have an area under the ROC curve of approximately 0.50

to visually compare different models, their ROC curves can be superimposed on the same graph

Comparing ROC curves can be useful in contrasting two or more models with

different predictor sets (for the same model),

different tuning parameters (i.e., within model comparisons),

or complete different classifiers (i.e., between models)

the optimal line should be shifted towards the upper left corner of the plot

alternatively, the model with the largest area under the ROC curve would be the most effective.

for the credit data, the logistic model had an estimated area under the ROC curve of 0.78 with a 95% confidence interval of (0.7, 0.85) determined using a bootstrap confidence interval method

ROC Curves

one advantage of using ROC curves to characterize models is that, since it is a function of sensitivity and specificity, the curve is insensitive to disparities in the class proportions

a disadvantage of using the area under the curve to evaluate models is that it obscures information

> when comparing models, it is common that no individual ROC curve is uniformly better than another (i.e., the curves cross). By summarizing these curves, there is a loss of information, especially if one particular area of the curve is of interest

ROC Curves

for example, one model may produce a steep ROC curve slope on the left but have a lower AUC than another model

if the lower end of the ROC curve was of primary interest, then AUC would not identify the best model

the partial area under the ROC curve is an alternative that focuses on specific parts of the curve

one often overlooked aspect of sensitivity and specificity is that they are conditional measures

using the sensitivity and specificity, an obstetrician can make statements such as “assuming that the fetus does not have Down syndrome, the test has an accuracy of 95 %.”

however, these statements might not be helpful to a patient since, for new samples, all that is known is the prediction

The person using the model prediction is typically interested in unconditional queries such as “what are the chances that the fetus has the genetic disorder?”

This depends on three values:

the sensitivity and

the specificity of the diagnostic test and

the prevalence of the event in the population

intuitively, if the event is rare, this should be reflected in the answer

taking the prevalence into account, the analog to sensitivity is the positive predicted value, and the analog to specificity is the negative predicted value.

these values make unconditional evaluations of the data.The positive predicted value answers the question “what is the probability that this sample is an event?”

the formulas are

$$PPV = \frac{Sensitivity \times Prevalence}{(Sensitivity \times Prevalence) + ((1 - Specificity) \times (1 - Prevalence))}$$

$$NPV = \frac{Specificity \times (1 - Prevalence)}{(Prevalence \times (1 - Sensitivity)) + (Specificity \times (1 - Prevalence))}$$

the predictive values are nontrivial combinations of performance and the rate of events.

large negative predictive values can be achieved when the prevalence is low

as the event rate becomes high, the negative predictive value becomes very small. the opposite is true for the positive predictive values.

predictive values are not often used to characterize the model.

there are several reasons why, most of which are related to prevalence.

- prevalence is hard to quantify. few people, even experts, are willing to propose an estimate of this quantity based on prior knowledge

- prevalence is dynamic. For example, the rate of spam emails increases when new schemes are invented but later fall off to baseline levels

non-accuracy based criteria

for many commercial applications of predictive models, accuracy is not the primary goal for the model. Often, the purpose of the model might be to:

- predict investment opportunities that maximize return
- improve customer satisfaction by market segmentation
- lower inventory costs by improving product demand forecasts or
- reduce costs associated with fraudulent transactions

while accuracy is important, it only describes how well the model predicts the data.

non-accuracy based criteria

these metrics quantify the consequences of correct and incorrect predictions (i.e., the benefits and costs).

For example, in fraud detection, a model might be used to quantify the likelihood that a transaction is fraudulent. Suppose that fraud is the event of interest. Any model predictions of fraud (correct or not) have an associated cost for a more in-depth review of the case. For true positives, there is also a quantifiable benefit to catching bad transactions. Likewise, a false negative results in a loss of income.

non-accuracy based criteria

the models could alternatively be characterized using the profit gain or lift, estimated as the model profit above and beyond the profit from a mass mailing.

with two classes, a general outline for incorporating unequal costs with performance measures is given by the probability-cost function (PCF) as

$$PCF = \frac{P \times C(+|-)}{P \times C(-|+) + (1 - P) \times C(+|-)}$$

where P is the (prior) probability of the event, C(–|+) is the cost associated with incorrectly predicting an event (+) as a nonevent, and C(+|–) is the cost of incorrectly predicting a nonevent. The PCF is the proportion of the total costs associated with a false-positive sample.

non-accuracy based criteria

then you can use the normalized expected cost (NEC) function to characterize the model

$$NEC = PCF \times (1 - TP) + (1 - PCF) \times FP$$

for a specific set of costs.

essentially, the NEC takes into account the prevalence of the event, model performance, and the costs and scales the total cost to be between 0 and 1.

Note: this approach only assigns costs to the two types of errors and might not be appropriate for problems where there are other cost or benefits

remedies for severe class imbalance

when modeling discrete classes, the relative frequencies of the classes can have a significant impact on the effectiveness of the model.

an imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes

imbalance can be present in any data set or application — should be aware of the implications of modeling this type of data

predicting caravan policy ownership

a data set generated by the computational intelligence and learning (CoIL) research network is used to illustrate methods for combatting class imbalances. The 2000 CoIL Challenge was to predict whether customers would purchase caravan insurance

The outcome, whether the costumer purchased caravan insurance, is highly unbalanced with only 6% of customers having purchased policies.

in all there were 85 predictors. Many of the categorical predictors had 10 or more levels and the count-based predictors tended to be fairly sparse. the predictors in the data set consisted of:

Customer subtype designation, such as “Traditional Families” or “Afflu-ent Young Families.” There were 39 unique values, although many of the subtypes comprise less than 5 % of the customers.

Demographic factors, such as religion, education level, social class, income, and 38 others. The values of the predictors were derived from data at the zip code level, so customers residing in the same zip code will have the same values for these attributes.[2]

Product ownership information, such as the number of (or the contribution to) policies of various types.

predicting caravan policy ownership

to demonstrate different methodologies with these data, stratified random sampling (where the strata was the response variable) was used to create three different data sets:

- A training set of customers (n = 6877) that will be used to estimate model parameters, tuning models, etc.

- A small evaluation set of customers (n = 983) that will be used for developing post-processing techniques, such as alternative probability cutoffs

- A customer test set (n = 1962) that is solely used for final evaluations of the models

the effect of class imbalance

to begin, three predictive models were used to model the data: random forest, a flexible discriminant analysis model (with MARS hinge functions), and logistic regression.

to tune the models, 10-fold cross-validation was used; each holdout sample contained roughly 687 customers, which should provide reasonable estimates of uncertainty.

to choose the optimal model, the area under the receiver operating characteristic (ROC) curve was optimized.[3]

the effect of class imbalance

the random forest model used 1500 trees in the forest and was tuned over 5 values of the m_{try} parameter. the final model had an optimal m_{try} value of 126

the FDA model used first-degree features and was tuned over 25 values for the number of retained terms. The resampling process determined that 13 model terms was appropriate

logistic regression utilized a simple additive model (i.e., no interactions or nonlinear terms) with a reduced predictor set (many near-zero variance predictors were removed so that the model resulted in a stable solution)

the effect of class imbalance

a number of different performance metrics were estimated including: overall accuracy, Kappa, area under the ROC curve, sensitivity, and specificity (where a purchased policy was defined as the “event” of interest).

all models predicted the samples in the evaluation data set and yielded very similar results

Model	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Random forest	93.5	0.091	6.78	99.0	0.757
FDA (MARS)	93.8	0.024	1.69	99.7	0.754
Logistic regression	93.9	0.027	1.69	99.8	0.727

the effect of class imbalance

any patterns that were useful for predicting the outcome were overwhelmed by the large percentage of customers with no caravan insurance.

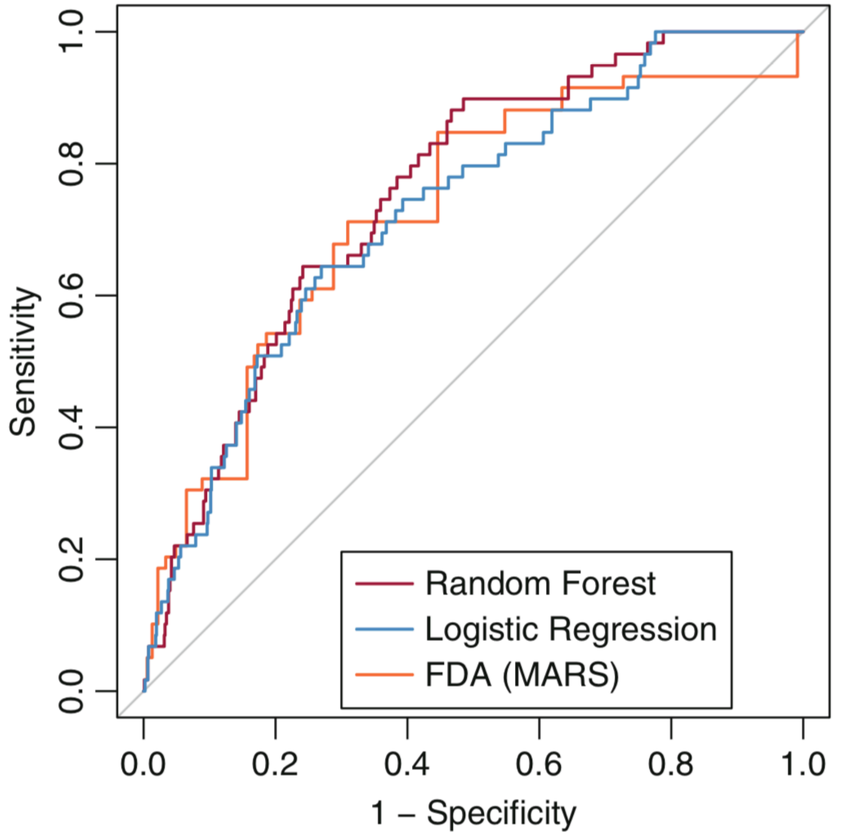
none of the models predicted more than 13 customers on the evaluation set as having insurance, despite 59 customers with insurance in the evaluation set.

Model	Accuracy	Kappa	Sensitivity	Specificity	ROC AUC
Random forest	93.5	0.091	6.78	99.0	0.757
FDA (MARS)	93.8	0.024	1.69	99.7	0.754
Logistic regression	93.9	0.027	1.69	99.8	0.727

the imbalance also had a severe effect on the predicted class probabilities. In the random forest model, for example, 82% of the customers have a predicted probability of having insurance of 10% or less.

the effect of class imbalance

the ROC curves show considerable overlap and does not differentiate the models.



model tuning

the simplest approach to counteracting the negative effects of class imbalance is to tune the model to maximize the accuracy of the minority class(es)

for insurance prediction, tuning the model to maximize the sensitivity may help desensitize the training process to the high percentage of data without caravan policies in the training set

model tuning

the random forest model that was tuned for these data did not show a meaningful trend in sensitivity across the tuning parameter

the FDA model did show a trend; as the number of model terms was increased, there was a rise in sensitivity from effectively 0% for very simple models to 5.4 % when 16 terms were retained — this minor improvement in sensitivity comes at virtually no cost to specificity

given that the increase in sensitivity is not high enough to be considered acceptable, this approach to solving the problem is not effective for this particular data set

alternate cutoffs

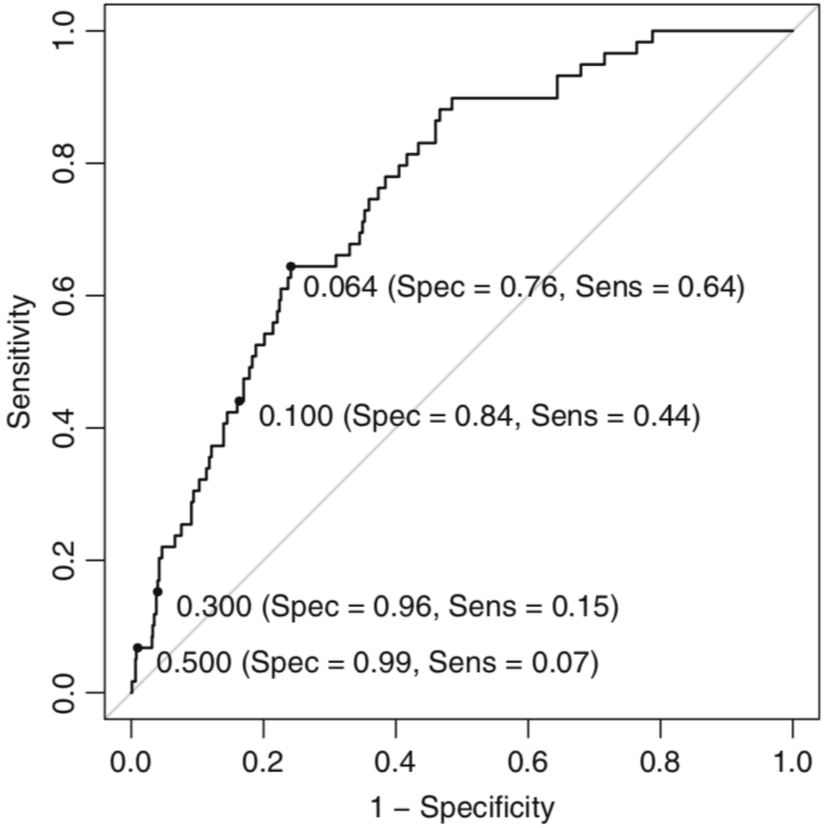
when there are two possible outcome categories, another method for increasing the prediction accuracy of the minority class samples is to determine alternative cutoffs for the predicted probabilities which effectively changes the definition of a predicted event.

the most straightforward approach is to use the ROC curve since it calculates the sensitivity and specificity across a continuum of cutoffs.

using ROC curve, an appropriate balance between sensitivity and specificity can be determined

several cutoffs are shown on the curve and it is apparent that decreasing the cutoff for the probability of responding increases the sensitivity (at the expense of the specificity).

there may be situations where the sensitivity/specificity trade-off can be accomplished without severely com-promising the accuracy of the majority class



alternate cutoffs

several techniques exist for determining a new cutoff

if there is a particular target that must be met for the sensitivity or specificity, this point can be found on the ROC curve and the corresponding cutoff can be determined

find the point on the ROC curve that is closest (i.e., the shortest distance) to the perfect model (with 100 % sensitivity and 100 % specificity), which is associated with the upper left corner of the plot

can also use Youden's J Index (won't be discussed)

alternate cutoffs

using the evaluation set, the predicted sensitivity for the new cutoff of 0.064 is 64.4%, which is a significant improvement over the value generated by the default cutoff.

the consequence of the new cutoff is that the specificity is estimated to drop from 99% to 75.9% — may or may not be acceptable based on the context of how the model will be used.

	0.50 Cutoff		0.064 Cutoff	
	Insurance	no insurance	Insurance	no insurance
Insurance	11	19	71	441
Noinsurance	105	1827	45	1,405

alternate cutoffs

in our analysis, the alternate cutoff for the model was not derived from the training or test sets

It is important, especially for small samples sizes, to use an independent data set to derive the cutoff

> If the training set predictions are used, there is likely a large optimistic bias in the class probabilities that will lead to inaccurate assessments of the sensitivity and specificity.
>
> If the test set is used, it is no longer an unbiased source to judge model performance.

alternate cutoffs

it is also worth noting that the core of the model has not changed — the same model parameters are being used.

changing the cutoff to increase the sensitivity does not increase the overall predictive effectiveness of the model.

the main impact that an alternative cutoff has is to make trade-offs between particular types of errors.

alternate cutoffs

for many classification problems, comparing models on the basis of the default sensitivity and specificity may be misleading. Since a better cutoff may be possible, an analysis of the ROC curve can lead to improvements in these metrics

consequently, performance metrics that are independent of probability cutoffs (such as the area under the ROC curve) are likely to produce more meaningful contrasts between models.

however, some predictive models only produce discrete class predictions

sampling methods

when there is *a priori* knowledge of a class imbalance, one straightforward method to reduce its impact on model training is to select a training set sample to have roughly equal event rates during the initial data collection

basically, instead of having the model deal with the imbalance, we can attempt to balance the class frequencies — taking this approach eliminates the fundamental imbalance issue that plagues model training.

if the training set is sampled to be balanced, the test set should be sampled to be more consistent with the state of nature and should reflect the imbalance so that honest estimates of future performance can be computed

sampling methods

if an a priori sampling approach is not possible, then there are post hoc sampling approaches that can help attenuate the effects of the imbalance during model training.

Two general post hoc approaches are down-sampling and up-sampling the data.

Up-sampling is any technique that simulates or imputes additional data points to improve balance across classes, while down-sampling refers to any technique that reduces the number of samples to improve the balance across classes.

sampling methods

one approach to up-sampling in which cases from the minority classes are sampled with replacement until each class has approximately the same number.

For the insurance data, the training set contained 6466 non-policy and 411 insured customers.

> if we keep the original minority class data, adding 6055 random samples (with replacement) would bring the minority class equal to the majority.
>
> some minority class samples may show up in the training set with a fairly high frequency while each sample in the majority class has a single realization in the data

sampling methods

down-sampling selects data points from the majority class so that the majority class is roughly the same size as the minority class(es).

there are several approaches to down-sampling.

a basic approach is to randomly sample the majority classes so that all classes have approximately the same size

another approach would be to take a bootstrap sample across all cases such that the classes are balanced in the bootstrap set — the advantage of this approach is that the bootstrap selection can be run many times so that the estimate of variation can be obtained about the down-sampling

sampling methods

one implementation of random forests can inherently down-sample by controlling the bootstrap sampling process within a stratification variable

if class is used as the stratification variable, then bootstrap samples will be created that are roughly the same size per class

these internally down-sampled versions of the training set are then used to construct trees in the ensemble.

the synthetic minority over-sampling technique (SMOTE) is a data sampling procedure that uses both up-sampling and down-sampling, depending on the class, and has three operational parameters: the amount of up-sampling, the amount of down-sampling, and the number of neighbors that are used to impute new cases.

synthetic minority over-sampling technique (SMOTE)

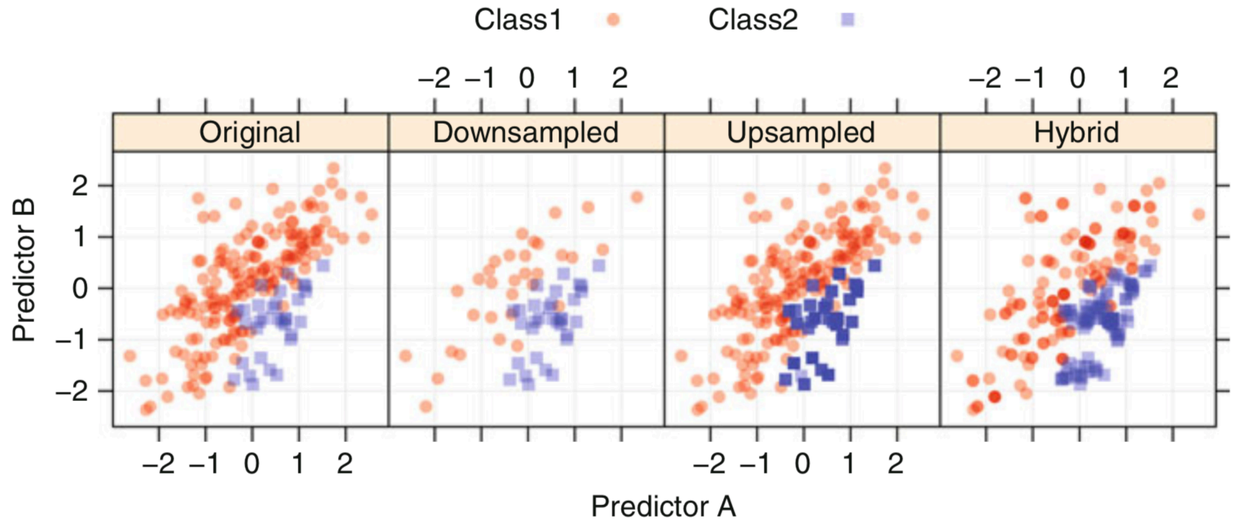
to up-sample for the minority class, SMOTE synthesizes new cases.

to do this, a data point is randomly selected from the minority class and its K-nearest neighbors (KNNs) are determined

the new synthetic data point is a random combination of the predictors of the randomly selected data point and its neighbors

while the SMOTE algorithm adds new samples to the minority class via up-sampling, it also can down-sample cases from the majority class via random sampling in order to help balance the training set

the original data contain 168 samples from the first class and 32 from the second (a 5.2:1 ratio). The down-sampled version of the data reduced the total sample size to 64 cases evenly split between the classes. The up- sampled data have 336 cases, now with 168 events. The SMOTE version of the data has a smaller imbalance (with a 1.3:1 ratio) resulting from having 128 samples from the first class and 96 from the second.



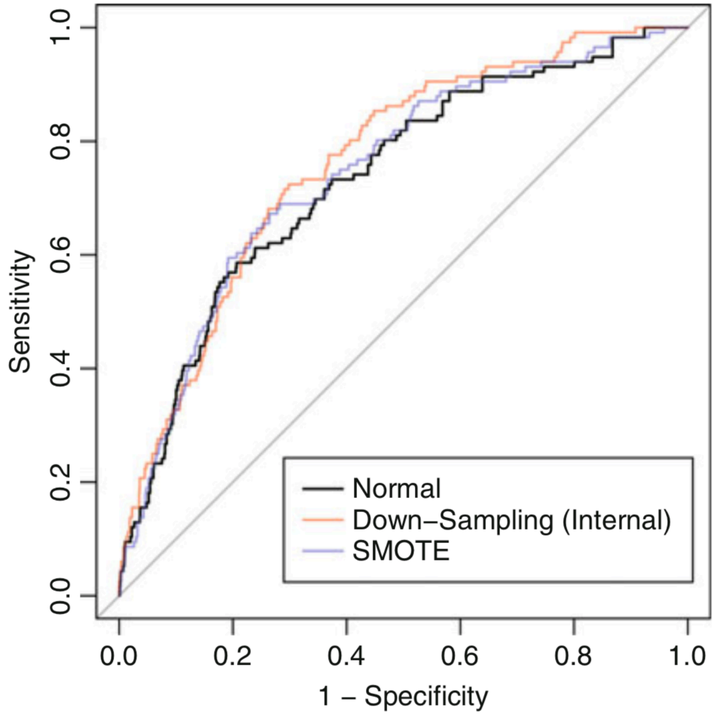
sampling methods

it should be noted that when using modified versions of the training set, resampled estimates of model performance can become biased

for example, if the data are up-sampled, resampling procedures are likely to have the same sample in the cases that are used to build the model as well as the holdout set, leading to optimistic results

these sampling methods were applied to the random forest models for the insurance data using the same tuning process as the original model.

Method	Evaluation	Test		
	ROC	ROC	Sensitivity	Specificity
Original	0.757	0.738	64.4	75.9
Down-sampling	0.794	0.730	81.4	70.3
Down-sampling (Internal)	0.792	0.764	78.0	68.3
Up-sampling	0.755	0.739	71.2	68.1
SMOTE	0.767	0.747	78.0	67.7



sampling methods

the results show that the up-sampling procedure had no real improvement on the area under the curve.

SMOTE showed an improvement in the evaluation set, but the increase in the area under the ROC curve was not reproduced in the larger test set

simple down-sampling of the data also had a limited effect on model performance

down-sampling inside the random forest model had robust areas under the ROC curve in both data sets—this may be due to using independent realizations of the majority class in each tree