Paper 942-2017

Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data

Josephine S Akosa, Oklahoma State University

ABSTRACT

The most commonly reported model evaluation metric is the accuracy. This metric can be misleading when the data are imbalanced. In such cases, other evaluation metrics should be considered in addition to the accuracy. This study reviews alternative evaluation metrics for assessing the effectiveness of a model in highly imbalanced data. We used credit card clients in Taiwan as a case study. The data set contains 30,000 instances (22.12% risky and 77.88% non-risky) assessing the likeliness of a customer defaulting on a payment. Three different techniques were used during the model building process. The first technique involved down-sampling the majority class in the training subset. The second used the original imbalanced data whereas prior probabilities were set to account for oversampling in the third technique. The same sets of predictive models were then built for each technique after which the evaluation metrics were computed. The results suggest that model evaluation metrics might reveal more about distribution of classes than they do about the actual performance of models when the data are imbalanced. Moreover, some of the predictive models were identified to be very sensitive to imbalance. The final decision in model selection should consider a combination of different measures instead of relying on one measure. To minimize imbalance-biased estimates of performance, we recommend reporting both the obtained metric values and the degree of imbalance in the data.

INTRODUCTION

One of the biggest challenges in data mining is dealing with highly imbalanced data sets. We encounter imbalanced data in several real world applications including, credit card fraud detection, churn prediction, customer retention, and medical diagnostics among many others. An imbalance occurs when one or more classes (minority class) have very low proportions in the data as compared to the other classes (majority class). Mostly in these situations, the main interest is in correctly classifying the minority class. However, the most commonly used classification algorithms do not work well for such problems. This is because the classifiers tend to be biased towards the majority class and hence perform poorly on the minority class. Several different techniques have been proposed to solve the problems associated with learning from class-imbalanced data. One of such techniques is based on cost sensitive learning. Here, a high cost is assigned to misclassification of the minority class while trying to minimize the overall cost. For instance, an analyst might have reasons to believe that misclassifying the minority class (false negatives) is *X* times costlier than misclassifying the majority class (false positives). The addition of specific costs during model training will bias the model towards the minority class thereby affecting the model parameters. This therefore has a potential of improving upon the performance of the model. Another technique is to use a sampling technique during model training: either down-sample the majority class or over-sample the minority class. Down-sampling is a technique utilized to reduce the number of samples in the majority class to obtain roughly equal data points across the classes. Up-sampling on the other hand simulates data points to enhance balance across the classes. Even though the use of sampling techniques can introduce bias into the model results, these techniques can still be effective during the tuning of model parameters.

Despite the improvements of the above techniques on model performance during parameter tuning, we note that the best performing model is not chosen based on the performance measure of the training subset but on the testing subset. The distribution of the testing data may differ from that of the training data, and the true misclassification costs may be unknown at learning time. In addition, the testing data needs to be consistent and reflect the state of nature of the real data in order to produce honest estimates of future events. Consequently, sampling techniques cannot be applied to the testing data to fairly balance the class distribution. In such situations, it is the duty of the researcher or practitioner to determine an appropriate performance measure to use when choosing between different classifiers.

EFFECT OF CLASS-IMBALANCE ON PREDICTIVE ACCURACY

For illustration, consider building a classification model for detecting credit card frauds. Assume the data set for building the classifier has 990 genuine events (majority class) and only 10 fraudulent events (minority class). The interest here will be to accurately classify the fraudulent events. Naturally, a classifier will classify all events as genuine to optimize for accuracy; given an accuracy of 99% (Table 1). Unfortunately, this classifier is useless as the events of interest have been misclassified.

	Classified positive	Classified negative
Actual positive	0	10
Actual negative	0	990

Table 1. Confusion matrix for classifier 1 in illustrative example

Now, consider another classifier that provides the results in Table 2. In this scenario, the accuracy of the classifier is 98.6%. Even though the first classifier has a zero predictive power, there is an improvement in accuracy for this classifier over the second classifier. The name given to this exact situation is *accuracy paradox*. It is sometimes the case where the accuracy measure shows an excellent model performance but the accuracy is only reflecting the underlying class distributions.

	Classified positive	Classified negative
Actual positive	6	10
Actual negative	4	980

Table 2. Confusion matrix for classifier 2 in illustrative example

The question that we are faced with in such situation is “What performance measure(s) do I use in choosing between candidate models?” This study analyzed the effect of class-imbalance on the learning and evaluation process of a classifier. The results suggest that in the presence of class-imbalance, the model evaluation measures may reveal more about distribution of classes than they do about the actual performance of models. In addition, we identified some of the classification models (especially the gradient boosting model) to be very sensitive to class-imbalance and perform poorly in such cases. Consequently, the final decision in model selection should consider a combination of different performance measures instead of relying on one. To avoid or minimize imbalance-biased performance estimates, we recommend reporting both the obtained measure values and the degree of imbalance in the data.

MODEL PERFORMANCE MEASURES

In classification analysis, we usually evaluate a classifier by a confusion matrix (Table 3). In Table 3, the columns represent the classifier’s predictions and the rows are the actual classes. TP (True Positive) is the number of positive cases correctly classified as such. FN (False Negative) is the number of positive cases incorrectly classified as negatives. FP (False Positive) is the number of negative cases that are incorrectly identified as positive cases and TN (True Negative) is the number of negative cases correctly classified as such.

	Classified positive	Classified negative
Actual positive	TP	FN
Actual negative	FP	TN

Table 3. Confusion matrix for two classes’ classification

By convention, we consider the minority class in imbalanced data modeling as the positive class whilst the majority class is considered as the negative class. We derive most of the performance measures utilized in classification problems based on the confusion matrix. Some of these performance measures are summarized in Table 4.

Measure	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Misclassification rate (1 – Accuracy)	$\frac{FP + FN}{TP + TN + FP + FN}$
Sensitivity (or Recall)	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Precision (or Positive Predictive Value)	$\frac{TP}{TP + FP}$

Table 4. Some common performance measure based on confusion matrix analysis

The most commonly reported measure of a classifier is the accuracy. This measure evaluates the overall efficiency of an algorithm. However, as illustrated earlier, predictive accuracy can be a misleading evaluation measure when the data is imbalanced. This is because in such cases, more weights are placed on the majority class than on the minority class making it more difficult for a classifier to perform well on the minority class. Sensitivity, another performance measure, measures the accuracy of positive cases whereas specificity measures the accuracy of negative cases. Sensitivity assesses the effectiveness of the classifier on the positive/minority class while specificity assesses the classifier's effectiveness on the negative/majority class. For any given analysis, there is usually a trade-off between the sensitivity and the specificity. Precision on the other hand is a measure of a model's exactness. A higher precision value for a classifier is an indication of a good classifier.

COMBINED PERFORMANCE MEASURES

Generally, analysts would want to balance between both false positive and false negative rates. Previously discussed measures do not provide a good evaluating measure in such cases. Performance measures that try to balance between the false positives and the false negatives are discussed herein.

Geometric Mean

The Geometric Mean (G-Mean) is a metric that measures the balance between classification performances on both the majority and minority classes. A low G-Mean is an indication of a poor performance in the classification of the positive cases even if the negative cases are correctly classified as such. This measure is important in the avoidance of overfitting the negative class and under fitting the positive class.

$$G - Mean = \sqrt{Sensitivity \times Specificity}$$

Discriminant Power

The discriminant power (DP) is another measure that summarizes sensitivity and specificity. The formula is given by:

$$DP = \frac{\sqrt{3}}{\pi}(\log X + \log Y)$$

where X = sensitivity/(1-sensitivity) and Y = specificity/(1-specificity). The discriminant power assesses how well a classifier distinguishes between the positive and negative cases. The classifier is considered a poor classifier if DP < 1, limited if DP < 2, fair if DP < 3 and good in other cases.

F-Measure and Adjusted F-Measure

The F-Measure conveys the balance between the precision and sensitivity. The measure is 0 when either the precision or the sensitivity is 0. The formula for this measure is given by:

$$F - Measure = \frac{2 \times sensitivity \times precision}{sensitivity + precision}$$

This measure however performs well when the data is fairly balanced. The Adjusted F-Measure (AGF) is an improvement over the F-Measure especially when the data is imbalance. The AGF is computed by first computing:

$$F_2 = 5 \times \frac{sensitivity \times precision}{(4 \times sensitivity) + precision}$$

After, the class labels of each case are switched such that positive cases become negative and vice versa. A new confusion matrix with respect to the original labels is created and the quantity:

$$Inv\ F_{0.5} = \frac{5}{4} \times \frac{sensitivity \times precision}{(0.5^2 \times sensitivity) + precision}$$

The AGF is finally computed by taking the geometric mean of F_2 and $Inv\ F_{0.5}$ as

$$AGF = \sqrt{F_2 \times Inv\ F_{0.5}}$$

This measure accounts for all elements of the original confusion matrix and provides more weight to patterns correctly classified in the minority class. A high F- or adjusted F-measure indicates a good performing classifier on the minority class.

Balanced Accuracy

The balanced accuracy is the average between the sensitivity and the specificity, which measures the average accuracy obtained from both the minority and majority classes. This quantity reduces to the traditional accuracy if a classifier performs equally well on either classes. Conversely, if the high value of the traditional accuracy is due to the classifier taking advantage of the distribution of the majority class, then the balanced accuracy will decrease compared to the accuracy.

$$Balanced\ Accuracy = \frac{1}{2}(sensitivity \times specificity)$$

Matthew’s Correlation Coefficient

The Matthews correlation coefficient (MCC) is least influenced by imbalanced data. It is a correlation coefficient between the observed and predicted classifications. The value ranges from -1 to +1 with a value of +1 representing a perfect prediction, 0 as no better than random prediction and -1 the worst possible prediction.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Cohen’s Kappa (or Kappa)

Kappa takes into account the accuracy that would be generated purely by chance. The form of the measure is:

$$kappa = \frac{total\ accuracy - random\ accuracy}{1 - random\ accuracy}$$

where

$$total\ accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

and

$$random\ accuracy = \frac{(TN+FP)(TN+FN)+(FN+TP)(FP+TP)}{(TP+TN+FP+FN)^2}$$

In a similar fashion to the MCC, kappa takes on values from -1 to +1, with a value of 0 meaning there is no agreement between the actual and classified classes. A value of 1 indicates perfect concordance of the model prediction and the actual classes and a value of −1 indicates total disagreement between prediction and the actual.

Youden's Index

Youden's index evaluates the ability of a classifier to avoid misclassifications. This index puts equal weights on a classifier's performance on both the positive and negative cases. Thus:

$$Youden's\ index\ (\gamma) = sensitivity - (1 - specificity)$$

A higher value of γ is an indication of a good performing classifier.

Likelihoods

The positive and negative likelihood ratios are two other good measures for evaluating a classifier's performance. The positive likelihood ratio (LR(+)) is the ratio between the probability of predicting a truly positive case as positive and the probability of predicting a truly negative case as positive. The negative likelihood ratio (LR(-)) is the ratio between the probability of predicting a truly negative case as positive and the probability of predicting a truly negative case as negative. Thus:

$$LR(+) = \frac{sensitivity}{1 - specificity}\ , \qquad LR(-) = \frac{1 - sensitivity}{specificity}$$

Based on the definitions, a higher positive likelihood ratio and a lower negative likelihood ratio is an indication of a good performance on positive and negative classes respectively.

RANK PERFORMANCE MEASURES

Receiver Operating Characteristic (ROC) Charts/ Area under the curve (AUC)

The ROC curve calculates the sensitivity and specificity across a continuum of cutoffs. An appropriate balance can be determined between sensitivity and specificity using the curve. AUC is simply the area under the ROC curve. An area of 1 represents a perfect model and an area of 0.5 represents a worthless model.

Lift Charts

Lift charts are a visualization tool that helps in assessing the ability of a classifier in detecting the events of interest in a data. Assume we want to score *N* sample events using a classifier. We would expect a good classifier to classify the events of interest such that the scores associated to these events are ranked higher than the nonevents. This is exactly what lift charts do. In a perfect classifier, the *N* highest-ranked samples would contain all *N* events of interest. Lift measures the effectiveness of a classifier by calculating the ratio between the outcomes attained with and without the classifier/predictive model.

DATA DESCRIPTION & PREPARATION

We utilized data on credit card clients in Taiwan available within the University of California Irvine (UCI) machine learning repository website as a case study. The data set contains 30,000 instances (22.12% risky and 77.88% non-risky) assessing the likeliness of a customer defaulting on a payment. There are 24 attributes per instance; including the target variable which classifies an instance as default payment (Yes = 1, No = 0). The predictor variables utilized are:

- Limit_bal: Given credit amount (NT dollar). This includes both the individual consumer credit and his/her family (supplementary) credit
- Sex: 1 = male; 2 = female

- Education: 1 = graduate school; 2 = university; 3 = high school; 4 = others
- Marital status: 1 = married; 2 = single; 3 = others
- Age (year)
- Pay_0 – Pay_6: History of past payment. Past monthly payment records were tracked from April to September 2005 as follows: Pay_0 = repayment status in September, 2005; Pay_2 = repayment status in August, 2005; …; Pay_6 = repayment status in April, 2005. The repayment status are measured on a scale of -1 to 9 where: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
- Bill_Amt1 – Bill_Amt6: Bill statement amount (NT dollar). These amounts are recorded as: Bill_Amt1 = September, 2005 bill statement amount; Bill_Amt2 = August, 2005 bill statement amount; …; Bill_Amt6 = April, 2005 bill statement amount
- Pay_Amt1 – Pay_Amt6: Previous payment amount (NT dollar). These are recorded as: Pay_Amt1 = amount paid in September, 2005; Pay_Amt2 = amount paid in August, 2005; …; Pay_Amt6 = amount paid in April, 2005

Initial exploratory analysis identified the continuous variables to have high skewness and kurtosis values. We applied log transformation on these variables to reduce the departures from normality. We also analyzed the categorical variables to check for anomalies in the scale of measurement.

METHODOLOGY

We followed the Cross Industry Standard Process for Data Mining (CRISP-DM) modeling approach. The five phases in this process include understanding the business problem, understanding the data, data preparation, modeling, evaluation and deployment. To ensure honest assessment of the models built, we partitioned the data into training (70%) and validation (30%) subsets. In order to analyze the effect of imbalance on modeling and performance measures three different techniques were utilized during model training. The first technique involved down-sampling the majority class in the training subset so that data points from the majority class is roughly the same size as the minority class. Models were then built and compared using this new dataset. The second involved building the model on the original training subsets of the imbalanced data whiles the last technique involved adjusting the prior probabilities to account for class-imbalance in the training subset. We built and evaluated the same predictive models in each of the three techniques.

All analyses were carried out in SAS® Enterprise Miner™ 13.1 and SAS/GRAPH® 9.4 software. Variable selection techniques were implemented prior to model building to select the most significant input variables. Several different predictive models including decision trees with variation in splitting rule target criteria (default, entropy, Gini, number of branches), logistic regression with variation in variable selection criteria (default, stepwise, backward, decision tree), neural and auto neural networks, gradient boosting, random forest and support vector machine (SVM) with variation in kernel function (linear, sigmoid, polynomial) were considered. Model performances were evaluated based on the measures described above using the validation subset.

RESULTS & DISCUSSION

A number of evaluations measures were executed to review the impact of class-imbalance on performance measures. This section reviews and discusses the results of the three different techniques employed.

ANALYSIS OF RESULTS FOR TECHNIQUE ONE

The results from the first technique are summarized in Table 5 and Figures 1, 2a and 2b. Based on the traditional predictive accuracy one can conclude that the best model is Model 1D (Gradient boosting) with the highest predictive accuracy of 78.11% followed by Model 1B (Neural Network) with predictive accuracy of 77.88%. Model 1F (Random Forest) turned out as the weakest model having the lowest predictive accuracy of 73.57%. However, diving a little deeper into the results and considering other evaluation measures that balance between false positives and false negatives, one can observe that though Model 1A has the second lowest predictive accuracy, this model is practically the best. The G mean, lift, balanced accuracy and adjusted F-measure all allocate the best value to Model 1A but with slight difference across the other results except for the lift which has a difference of about 0.7 between the lowest and highest models. The ROC and lift curves provides further evidence to the choice of Model 1A and even Model 1C over Model 1D (Figures 2a and 2b).

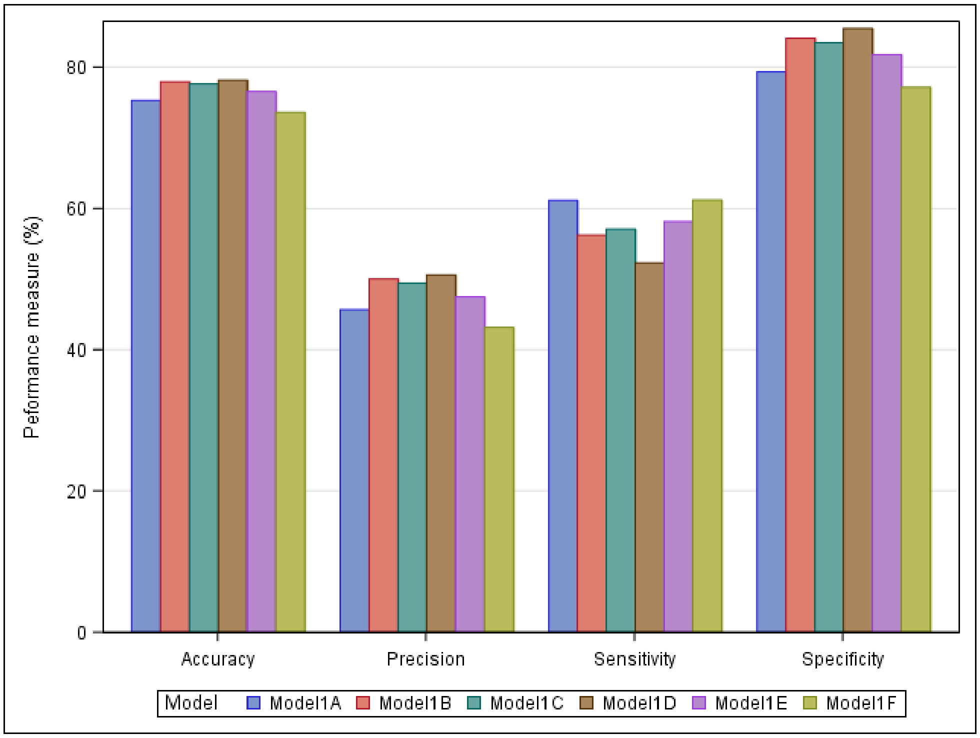

Figure 1. Some model performance metrics for Technique 1

Model 1A: Decision Tree
Model 1B: Neural Network
Model 1C: Logistic Regression
Model 1D: Gradient Boosting
Model 1E: Support Vector Machine
Model 1F: Random Forest

Measure	Model 1A	Model 1B	Model 1C	Model 1D	Model 1E	Model 1F
Kappa	0.36	0.385	0.383	0.373	0.369	0.333
Youden's index	0.404	0.402	0.404	0.377	0.398	0.382
G Mean	0.696	0.687	0.69	0.668	0.689	0.687
MCC	0.367	0.386	0.385	0.373	0.372	0.342
Balanced Accuracy	0.702	0.701	0.702	0.689	0.699	0.691
Adjusted F-Measure	0.701	0.689	0.692	0.669	0.692	0.693
F-Measure	0.522	0.529	0.53	0.514	0.523	0.506
ROC/AUC	0.756	0.757	0.759	0.719	0.711	0.744
Lift	3.010	2.751	2.812	2.283	2.463	2.830

Table 5. Model performance measures using Technique 1

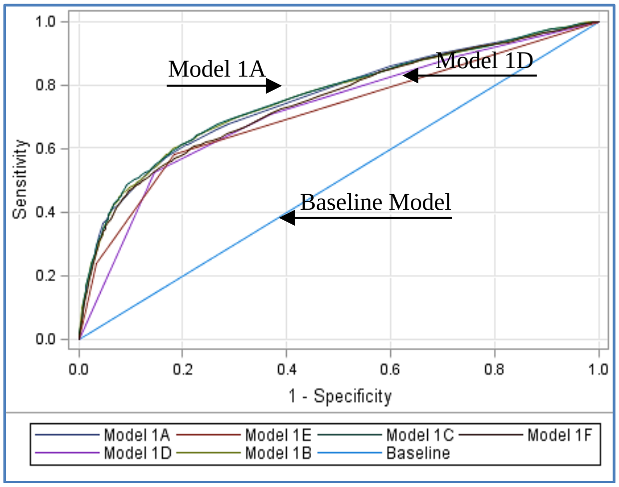

Figure 2a. ROC curves for Technique 1

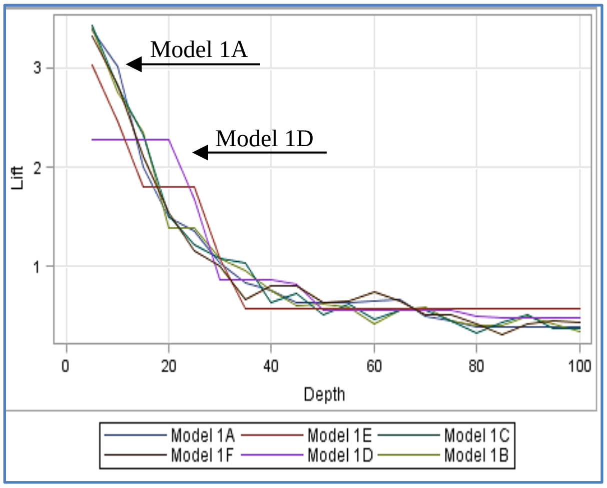

Figure 2b. Lift charts for Technique 1

ANALYSIS OF RESULTS FOR TECHNIQUE TWO

Similarly, the results of the second technique are summarized in Table 6 and Figures 3, 4a and 4b. Unlike the first technique, patterns of best performance metrics are not consistent across the models. Model 2F (Random Forest) shows the highest predictive accuracy of 82.11% but this measure differs slightly across the other models. Furthermore, Model 2E (SVM) has the highest precision and lift (68.99% and 3.104 respectively) whiles Model 2A (Decision Tree) has the highest Youden's index, G mean, balanced accuracy, F- and adjusted F-measures as shown in Table 6. Again, the measures provide evidence of choosing the decision tree model (Model 2A) over the other models despite the fact that the random forest model had the highest predictive accuracy.

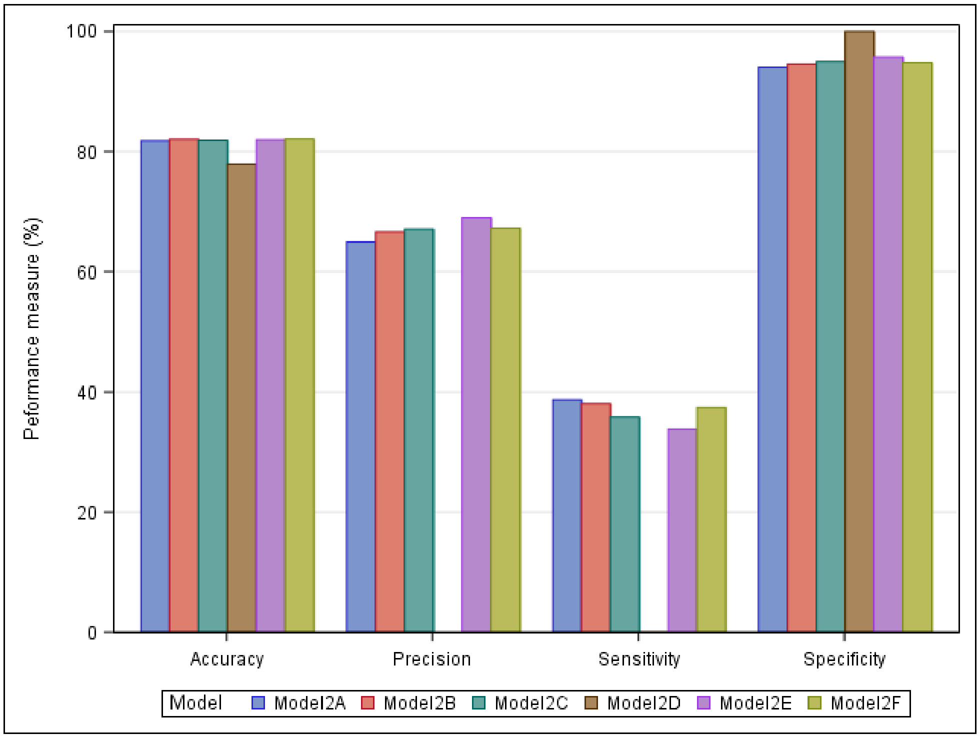

Figure 3. Some model performance metrics for Technique 2

Model 2A: Decision Tree	Model 2C: Logistic Regression	Model 2E: Support Vector Machine
Model 2B: Neural Network	Model 2D: Gradient Boosting	Model 2F: Random Forest

Of particular interest is the performance of Model 2D (Gradient Boosting). This model is virtually predicting every observation as not going to default on payment. However, the model has a somehow good predictive accuracy. This is a real depiction of the accuracy paradox discussed earlier. This shows that gradient boosting models are very sensitive to class-imbalance. If a practitioner or researcher is only considering the predictive accuracy and/or ROC/AUC for model evaluation, he/she may be unable to capture the poor performance of this model. In the presence of class-imbalance, we recommend the practitioner or researcher to utilize gradient boosting models with caution.

Measure	Model 2A	Model 2B	Model 2C	Model 2D	Model 2E	Model 2F
Kappa	0.383	0.385	0.37	0	0.361	0.383
Youden's index	0.328	0.326	0.308	0	0.295	0.322
G Mean	0.603	0.6	0.583	0	0.569	0.596
MCC	0.402	0.408	0.396	–	0.394	0.407
Balanced Accuracy	0.664	0.663	0.654	0.5	0.648	0.661
Adjusted F-Measure	0.602	0.599	0.582	–	0.568	0.595
F-Measure	0.485	0.484	0.467	–	0.454	0.481
ROC/AUC	0.749	0.754	0.755	0.720	0.648	0.753
Lift	2.921	2.892	2.895	2.283	3.104	2.756

Table 6. Model performance measures using Technique 2

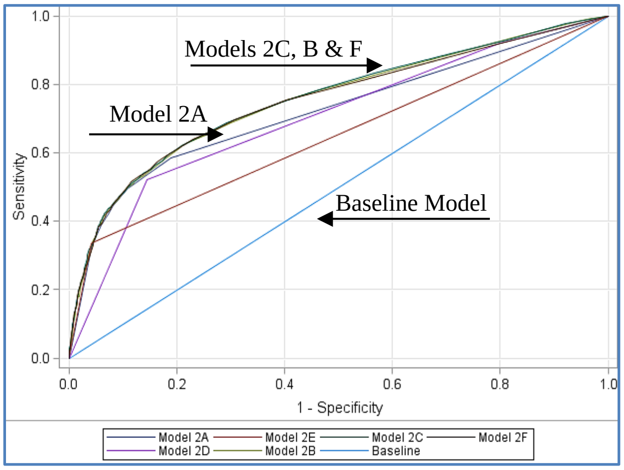

Figure 4a. ROC curves for Technique 2

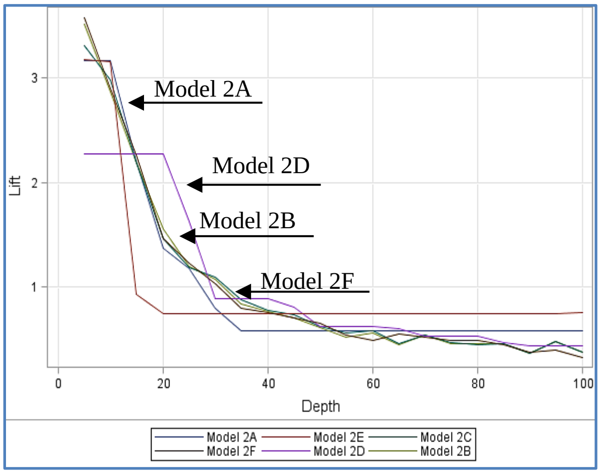

Figure 4b. Lift charts for Technique 2

ANALYSIS OF RESULTS FOR TECHNIQUE THREE

Finally, the results of the last technique are summarized in Table 7 and Figures 5, 6a and 6b. The SVM model (Model 3E) shows the highest predictive accuracy for this technique. Similar to the other techniques, the difference in the predictive measure across the other models is not drastic. Based on the common performance measures (Figure 5), one might conclude that the SVM model is the best model. However, considering the other performance measures, we observe that the neural network model rather balances well between the false positives and the false negatives as seen in the values of Table 7. Since we do not observe a particular model performing well in almost all the cases, making decision for this technique appears complex. One needs to rely on the cost associated with each of the possible misclassifications in making a final decision.

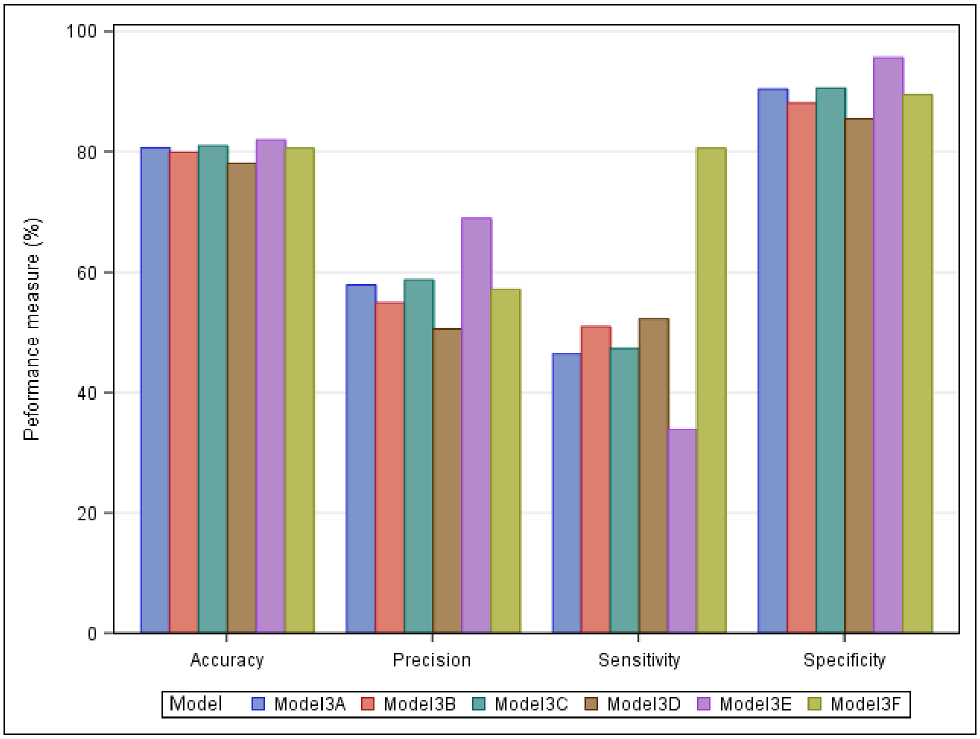

Figure 5. Some model performance metrics for Technique 3

Model 3A: Decision Tree
Model 3B: Neural Network

Model 3C: Logistic Regression
Model 3D: Gradient Boosting

Model 3E: Support Vector Machine
Model 3F: Random Forest

Measure	Model 3A	Model 3B	Model 3C	Model 3D	Model 3E	Model 3F
Kappa	0.396	0.401	0.407	0.373	0.361	0.406
Youden's index	0.368	0.390	0.379	0.377	0.295	0.386
G Mean	0.648	0.670	0.655	0.668	0.569	0.663
MCC	0.400	0.401	0.411	0.373	0.393	0.408
Balanced Accuracy	0.684	0.695	0.689	0.689	0.647	0.693
Adjusted F-Measure	0.647	0.669	0.654	0.669	0.568	0.662
F-Measure	0.515	0.528	0.524	0.514	0.454	0.528
ROC/AUC	0.749	0.754	0.755	0.755	0.648	0.753
Lift	2.921	2.892	2.895	2.895	3.104	2.756

Table 7. Model performance measures using Technique 3

The results reported for all the three techniques indicate that higher predictive accuracy does not necessarily guarantee a better performance of the classifier in general. This does not however applies to only the predictive accuracy; it applies to all the performance measures considered. Relying on only one performance measure to choose between candidate models can be misleading. Additionally, we identified that the degree of class-imbalance in the training data can have an impact on the best performing classifier. This is evident as not the same classifier was selected as the best performing classifier in all the three techniques.

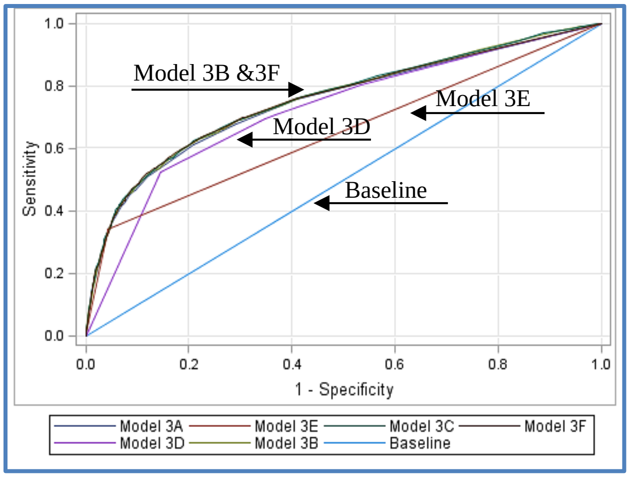

Figure 6a. ROC curves for Technique 3

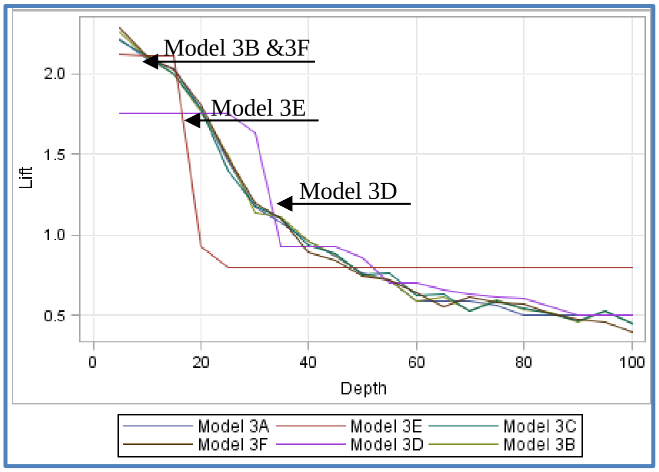

Figure 6b. Lift charts for Technique 3

CONCLUSION

One key factor to the success of any data mining method is in the model assessment technique. However, within the class-imbalanced data context, this subject is more complex since most of the frequently used performance measures could be misleading. Several real-world classification tasks, such as medical diagnosis, customer retention, credit card fraud detection, churn prediction among many others suffer from class-imbalance. This study analyzed the impact class-imbalance have on performance measures by looking at a considerable number of model performance measures. We established that class-imbalance results in a classifier's suboptimal performance. Model evaluation measures may reveal more about distribution of classes than they do about the actual performance of models when the data is imbalanced. We also identified some of the classification models to be very sensitive to class-imbalance and perform poorly in such cases. When dealing with class-imbalance data, the final decision in model selection should consider a combination of different measures instead of relying on only one measure. As a final remark, we recommend reporting both the obtained performance measure values and the degree of class-imbalance in the data to minimize imbalanced-biased performance estimates.

REFERENCES

Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. "Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation." *Australasian Joint Conference on Artificial Intelligence*. Springer Berlin Heidelberg, 2006.

Maratea, Antonio, Alfredo Petrosino, and Mario Manzo. "Adjusted F-measure and kernel scaling for imbalanced data learning." *Information Sciences* 257 (2014): 331-341.

Jeni, László A., Jeffrey F. Cohn, and Fernando De La Torre. "Facing Imbalanced Data--Recommendations for the Use of Performance Metrics."*Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Bekkar, Mohamed, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. "Evaluation Measures for ModelsAssessment over Imbalanced Datasets."*Journal of Information Engineering and Applications* 3.10 (2013).

Kuhn, M. and Johnson, K., 2013. *Applied predictive modeling* (Vol. 26). New York: Springer.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Josephine Sarpong Akosa
Graduate Teaching Associate
Oklahoma State University
josephine.akosa@okstate.edu
http://statistician.wixsite.com/josephine-akosa