Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2009, Article ID 868215, 14 pages
doi:10.1155/2009/868215

Research Article

Cumulative Gains Model Quality Metric

Thomas Brandenburger and Alfred Furth

Department of Mathematics and Statistics, South Dakota State University, Box 2220, Brookings, SD 57007, USA

Correspondence should be addressed to
Thomas Brandenburger, thomas.brandenburger@sdstate.edu

Received 1 December 2008; Accepted 13 April 2009

Recommended by Lean Yu

This paper proposes a more comprehensive look at the ideas of KS and Area Under the Curve (AUC) of a cumulative gains chart to develop a model quality statistic which can be used agnostically to evaluate the quality of a wide range of models in a standardized fashion. It can be either used holistically on the entire range of the model or at a given decision threshold of the model. Further it can be extended into the model learning process.

1. Introduction

In developing risk models, developers employ a number of graphical and numerical tools to evaluate the quality of candidate models. These traditionally involve numerous measures including the KS statistic or one of many Area Under the Curve (AUC) methodologies on ROC and cumulative Gains charts. Typical employment of these methodologies involves one of two scenarios. The first is as a tool to evaluate one or more models and ascertain the effectiveness of that model. Second however is the inclusion of such a metric in the model building process itself such as the way Ferri et al. [1] proposed to use Area Under the ROC curve in the splitting criterion of a decision tree.

However, these methods fail to address situations involving competing models where one model is not strictly above the other. Nor do they address differing values of end points as the magnitudes of these typical measures may vary depending on target definition making standardization difficult. Some of these problems are starting to be addressed. Marcade [2] Chief Technology officer of the software vendor KXEN gives an overview of several metric techniques and proposes a new solution to the problem in data mining techniques. Their software uses two statistics called KI and KR. We will examine the shortfalls he addresses more thoroughly and propose a new metric which can be used as an improvement to the KI and KR statistics. Although useful in a machine learning sense of developing a model, these same issues and solutions apply to evaluating a single model's performance as related by Siddiqi [3, chapter 6] and Mays [4, section 2] with respect to risk scorecards. We will not specifically give examples of each application of the new statistics but rather make the claim that it is useful in most situations where an AUC or model separation statistic (such as KS) is used.

2. Problems with Current Methods

2.1. Overview

As previously mentioned, the references indicate that there are many ways to assess the way in which a model classifies the outcome. Mays [4, chapter 6] separates these into a few general categories. We are concerned with two of these.

(i) Separation statistics. Within this specifically we are concerned with the KS statistic. Its advantages include that it is fairly easy to understand. In the context in which we use Kolmogorov-Smirnov (KS) statistic, it is defined as the maximum separation (deviation) between the cumulative distributions of “goods” and ”bads” as both Mays [4, chapter 6] and Siddiqi [3, page 123] outline it.

(ii) Ranking statistics. Siddiqi [3] outlines the C statistic or area under the ROC curve along with the Lorenz Curve, Gini index, Lift Curve, and Gains Chart. These definitions vary somewhat from source to source in industry literature. We will concern ourselves with the AUC of a Cumulative Gains chart similar to that used for the KI statistic by Marcade [2].

2.2. Details of the Cumulative Gains Chart

To illustrate some of the flaws of KS and AUC statistics, let us use two graphical examples. The figures represent example models built from actual test data of a random mailing of potential credit card recipients in the sub-prime credit market. The sample includes approximately 1400 cardholders who responded to credit offers. Models were built using logistic regression with several covariates. Figures 1 and 2 are charts from 2 candidate models.

The construction of the chart is as follows.

(1) Create a logistic model. It does not need to be a logistic model but the ability to define a level of the dependent or target variable as a successful prediction is necessary. In the case of risk the target is often a “bad” account since bad accounts have the greatest financial cost. Whereas if you were doing a model for a response to a mailing marketing campaign, a response would be “good” and that would be your target. To simplify for this example, the value of the dependent variable = 1 in this risk model will constitute a “bad” account risk. A value of the target dependent variable = 0 will be “good”.

(2) Score the data set on the model and rank them in order from highest to lowest probability of being bad (target = 1).

(3) Go through the ranked data set (highest to lowest probability) in a loop counting the cumulative number of actual points in the data set which were bad (value = 1) and good (value = 0).

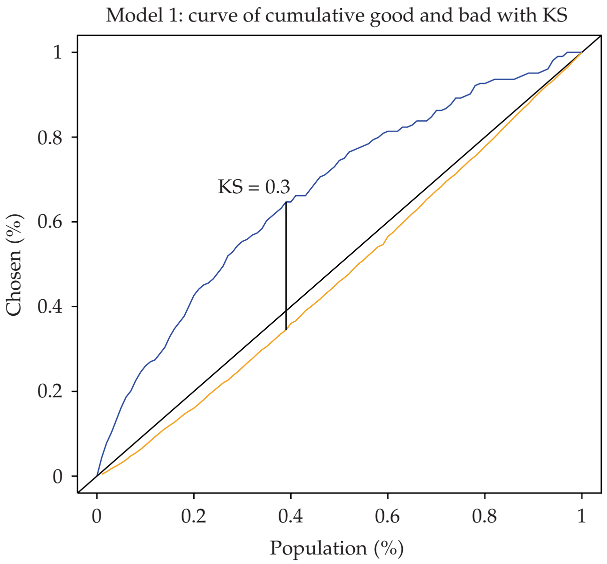

Figure 1: Model 1 Gains Chart

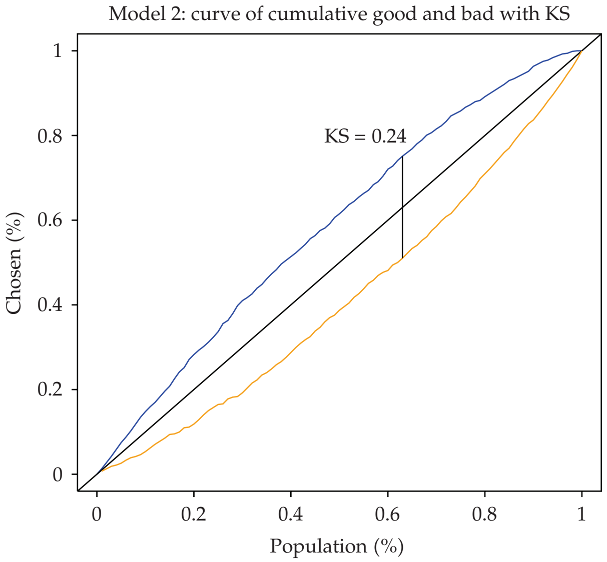

Figure 2: Model 2 Gains Chart

(4) Plot these 2 sets of values in 2 curves as a percentage of the proportion of the bad and good populations, respectively. In our risk model example the percent of the population in Model 1 which are “bad” is approximately 15 percent. In Model 2 the definition of the risk target variable is different so even though it is the same data set the bad rate is approximately 50 percent for the entire sample. So if there are 1400 data points and the proportion of bad in Model 1 is .15 then the data point on the graph corresponding to (.4,.63) would mean that by successfully picking the riskiest accounts (calculated by the model) in decreasing order it

would take 40 percent or the accounts (or 560) to find 63 percent of the bad accounts in the sample (or approximately 132 out of the 210 bad accounts in the 1400 sample size).

(5) Plot a 45-degree line. The meaning of this line is often glossed over quickly in literature and often misunderstood by analysts; however, it is very key to the development of descriptive statistics for model quality so we will detail its meaning. This line represents if you were truly random guessing. Imagine a bag full of some green balls and some red balls. A random ball is successively picked at random out of the hat without replacement. Its value is cataloged as red or green. If you pick zero balls out of the bag you would get zero red. Picking all the balls out of the hat would result in having all the red balls accounted for. Picking half of the balls out of the bag should on average net you half of the red balls which were in the bag regardless of the proportion of red balls. Hence the vertical axis is “as a percentage of the bad accounts”. Choosing randomly you should get on average half the bad accounts with half the population chosen regardless of the proportion of bad accounts.

(6) Calculate KS by taking the maximum difference between these good and bad curves.

2.3. Shortcomings

When observing the cumulative gains charts in Figures 1 and 2, it is intuitive that the greater the separation between the good and bad curves, the better the model. If one model has separation across the entire ranking greater than another model with the same definition of the target variable, then as pointed out by Piatetsky-Shapiro and Steingold [5, section 1] for any cost versus benefit matrix the dominate model wins. However in practical models you may not get the case where one model is strictly dominating over the other. Consider Model 2. It is easy to imagine another model which has the same KS but the distribution of the curve curve is such that the KS occurs significantly further to the left and the curves may cross. In fact they may cross more than once. Piatetsky-Shapiro and Steingold [5, section 1], Mays [4, page 110], Siddiqi [3, page123], and Marcade [2] all point out that practically speaking a model has a functional decision cutoff, and the KS may fall outside the range of risk values used to extend credit therefore making its usefulness suspected. Further they indicate that the performance of the model in the intended range of use is highly important.

The reality of the intended practical use of the model cutoff is also important in the calculation of any AUC type statistic. Notice that in Figure 1 not only is the KS greater but it occurs further to the left. This creates a bulbous shape in the curve to the left side of the graph. When comparing 2 models it is often more important what happens in the lower deciles. For instance in a mailing campaign for marketing the decision may be to only mail the 20 percent of the population which is most likely to respond. Consider 2 models which happen to have the same KS and same AUC under the bad curve. The model which is ”shifted left” with its KS occurring earlier and AUC weighted to the left will have superior financial performance since it is identifying a greater percentage of the target variable with a lower percentage of population. This concept is widely known and is easy to see visually, however there is no good quantification of this. We propose that our model statistic will account for this. A more difficult fact to comprehend from the graph is that of the proportion of bad in the population and its effect on KS and AUC. Consider 2 modeling situations. The first is the one presented here in the difference between Model 1 and Model 2. Model 1 has a much lower proportion of bad in the population as compared to Model 2 (15 percent versus 50 percent) even though the data set is the same. This is due to redefining what constitutes a bad account.

Mays [4, page 122] mentions different bad definitions but does not address its effect on the statistical metrics. The situation here is that the data is entirely the same for both models with the exception of the definiton of bad. Depending on the goal of the model the definition of bad may be different. For instance in a credit situation using the same data set one may predict the probability of an account being “charged off” as bad debt within 3 years of extending a credit offer. However for purposes of expediency (rather than waiting for 3 years for the test data to mature in performance) a surrogate endpoint may be used. As an example an endpoint of “60 days or more in arrears at 12 months” may be used instead. The two different definitions of bad will not have the same “bad rate”. Nor will you be able to have the same ability to predict the indended endpoint of bad debt at 3 years if using the earlier estimate of risk. Inherently 2 models built on these 2 different endpoints would have different values of KS. The second case is of different populations. In prime versus subprime you have entirely different risk levels as a population. Mays [4, chapter 7] addresses an entire chapter on KS indicating that the KS for subprime loans may be different than prime loans but misses the opportunity to explain that much of the difference is arithmetic. The author asserts that KS statistics for subprime loans will be lower because they have a much higher bad rate and concludes that the reason for the lower KS is the inability of models to differentiate between good and bad in higher risk credit markets. As we will demonstrate this is not necessarily the case. The difference in KS between 2 models with largely different KS can be at least partially explained by a rather obvious arithmetic property.

3. Improving KS and AUC

3.1. The Perfect Model

Noting that KS can be misleading Siddiqi [3, page 123] suggusts one way to view the separation between goods and bads is to plot the separation over the whole range of the model at specified increments. This technique gives a better view of the distribution of

Separation as can be seen in Figures 3 and 4.

Notice that just like the gains chart even though there are differences in magnitude the general shape is to increase to a peak somewhere in the middle of the distribution then decrease. Piatetsky-Shapiro and Steingold [5] begin the discussion of creating a statistic which can be used to account for the irregularities we have already noted. They use the concept of an ideal or optimal model to compare the given model through “lift quality”. The commercial software KXEN uses this concept heavily in the development of the statistics called KI and KR. We will concentrate on the explanation of KI as it is explained in their white paper. Consider Figures 1 and 2 redrawn in Figures 5 and 6, this time including a diagonal line with slope $= 1/\text{badrate}$.

As outlined in both Piatetsky-Shapiro and Steingold [5] and Marcade [2], the upper line represents the ideal model or perfect. If you rank the data set from highest to lowest probability of having target value = 1 then the best you could do is be right every time. In the analogy of the bag of red and green balls this line would be equivalent to drawing a red ball every time until the red balls were all gone. So in the case of Model 1, 15 percent of the population is bad and perfect model performance would be to choose 100 percent of these bad accounts in the first 15 percent of the population by picking them all correctly. It is by drawing this line that you then see that the model KS is constrained by the proportion of bad and good accounts in the data set. Again that proportion is determined by two things, the

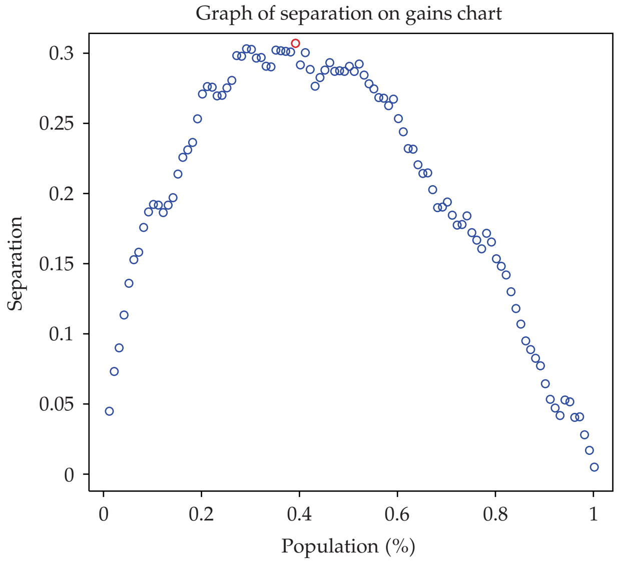

Figure 3: Model 1 Separation

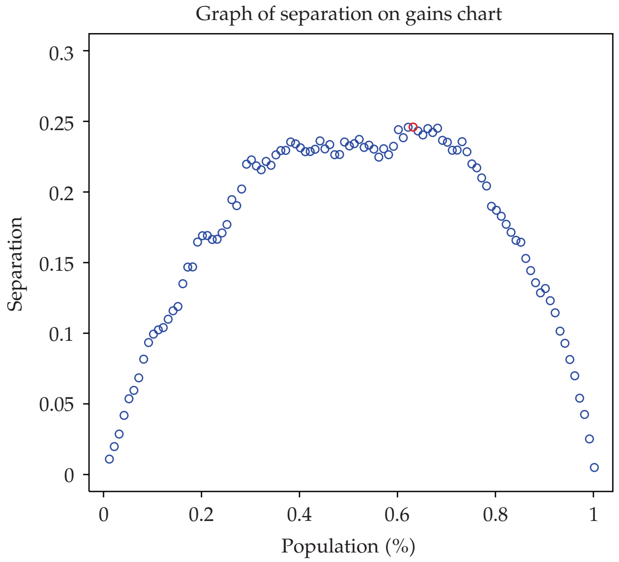

Figure 4: Model 2 Separation

definition of bad and the general risk level of the population. It can be immediately seen that the assertion by Mays [4] is attributing a lower KS to a higher rate of bad due to a model not being which is not as good at rank ordering the bad accounts starts to fall apart. You can also see why the plots of KS in Figures 3 and 4 must go up then come down in the fashion they do. What is less obvious at first is that this also generally means that the KS for the model will be to the right of the corner in the perfect graph. Note in both Model 1 and Model 2 this is true. However, depending on the application, the model's performance may be most crucial at the extremes rather than at the middle due to financial performance considerations.

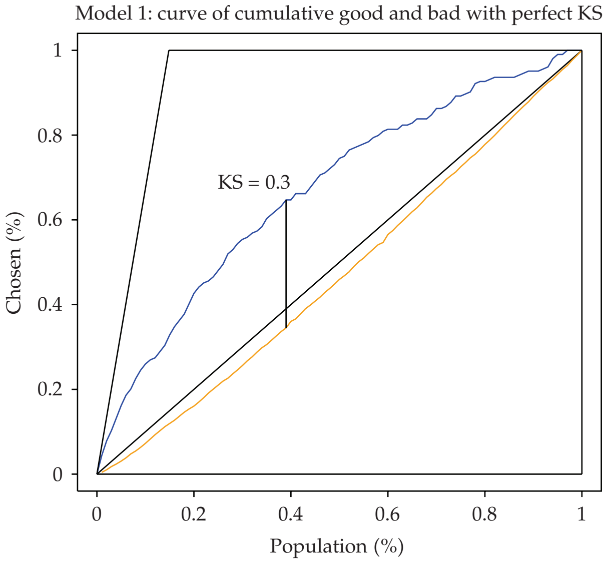

Figure 5: Model 1 Gains with Perfect Line

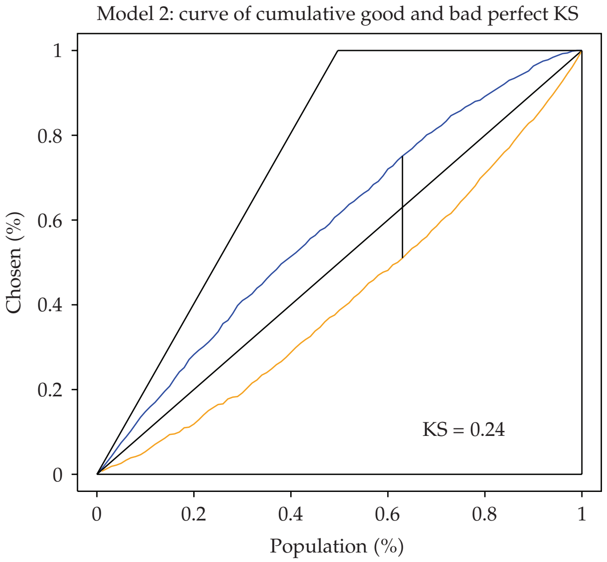

Figure 6: Model 2 Gains with Perfect Line

3.2. KI and KR

KXEN's KI statistic then uses the concept laid out by Piatetsky-Shapiro and Steingold [5] to create two statistics. We have used KXEN's analytics tool and find it to be an outstanding piece of software with powerful results which are easy to interpret. The KI statistic falls between 0 and 1, gives a value of 1 for a perfect model, and gives 0 for a completely random model. This gives it an intuitive feel for a good model metric as Marcade [2] suggests it should. KI is calculated as a "percent of perfect". The target (in this case bad) curve is plotted.

Let $p(x)$ be a function describing the distance between the line of the perfect model and the random diagonal line. Let $a(x)$ be a function describing distance between the curve of the model target and the random line

$$\mathrm{KI} = \frac{\int_0^1 a(x)dx}{\int_0^1 p(x)dx}. \tag{3.1}$$

KR not covered in the reference is a natural extention of KI. It is a measure used on validation data to determine stability of the model. It is essentially the KI of the validation data divided by KI of the model data. The closer this ratio is to 1, the more stable the model. KI is an attempt to offer some measure of "percent of perfect" for the model gains versus a perfect model. It however leaves out a few key ideas.

(i) There is no consideration of the shape of the curve for models whose separation may be shifted left. A model which has a relatively larger separation earlier in the ranking is favorable. KI has no way to reward this.

(ii) It does not incorporate the full consideration of separation between both curves but rather considers only the target. This idea will be more fully explored in our statistic.

(iii) It is only useful as a full model metric. It cannot be used at a cutoff decision point to evaluate a model. That is, KI(x) is not defined.

3.3. Redefining KI

Looking at the definition of KI, 2 of the 3 listed drawbacks of the statistic could be mitigated by a small but significant shift in definition. Let us redefine KI by

$$\mathrm{KI}^* = \int_0^1 \frac{a(x)}{p(x)} dx. \tag{3.2}$$

This change is small but crucial. It does 2 things. First it allows us to define a function which is a point estimate of model quality. We will call it $q(x)$. Second it allows us to create a metric for a range of the model risk rather than just a point or the whole model. Remember that KS was a point estimate, and AUC statistics are typically full model statistics. By taking the integral over a region of the model that we are practically interested in we can achieve a statistic which answers the need,

$$q(x) = \frac{a(x)}{p(x)}, \tag{3.3}$$

$$\mathrm{KI}^* = \int_a^b \frac{a(x)}{p(x)} dx. \tag{3.4}$$

By taking (a, b) to be (0,1) you obtain a statistic on the whole model. It becomes immediately obvious that $q(x)$ can be used as a point estimate of model separation where q is a ratio between actual separation and perfect separation from random. This change by itself would

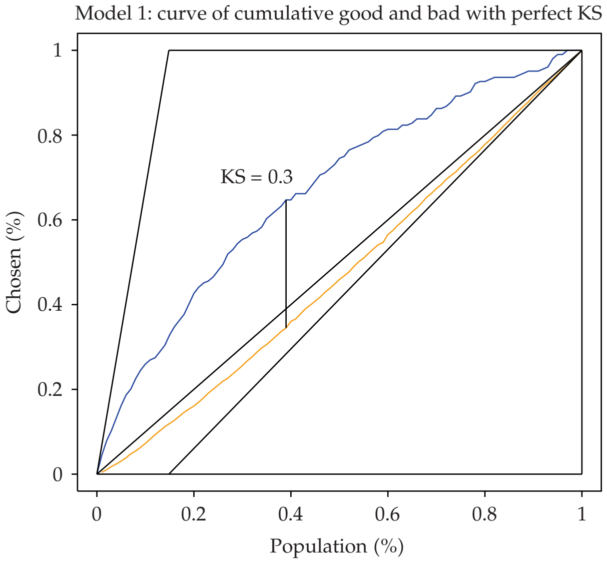

Figure 7: Model 1 Gains with Good and Bad Perfect Lines

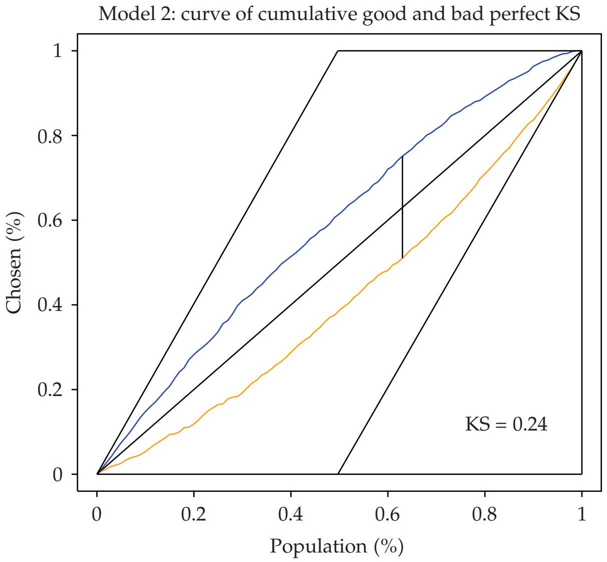

Figure 8: Model 2 Gains with Good and Bad Perfect Lines

be sufficient to be a useful improvement to KI. However there is still one more item not addressed by this metric. That is the idea of separation between goods and bads. To understand the necessity of this portion of the problem, let us once again redraw our two models . In Figures 7 and 8, we insert another diagonal line with slope = 1/(*goodrate*).

Immediately, another visual difference between the two models is explained. Notice that in Model 1 the good curve follows closely under the diagonal line while in Model 2 there appears to be a large separation. Once the second diagonal line is drawn it becomes

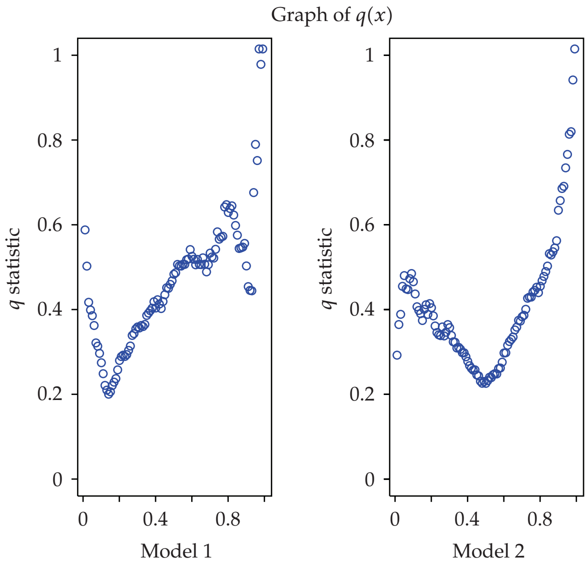

Figure 9: $q(x)$ for Both Models

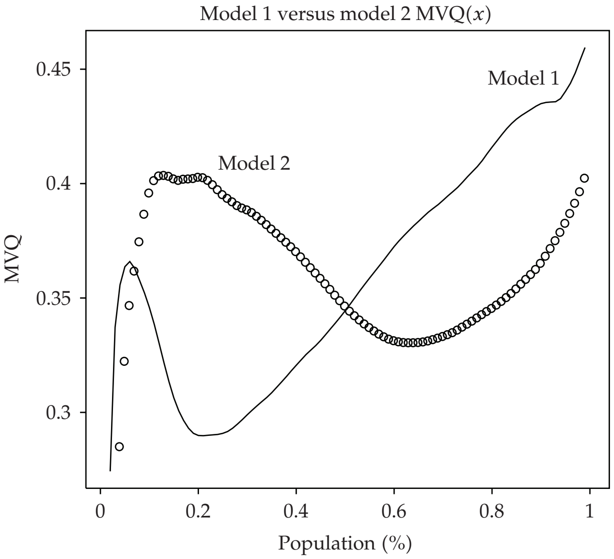

Figure 10: MVQ for Both Models

obvious that the good curve is being constrained by the lower diagonal in the same way the bad curve is constrained by the upper diagonal. What is the meaning of this diagonal? Consider the problem of the red and green balls in the bag in our analogy. The upper perfect curve represents drawing all the red balls out of the bag without error. Once you have drawn them all, the curve must stop at 1 and continue on horizontally as there are no more red balls. Consider the curve of green balls in this circumstance. It would remain at zero until all the red balls had been exhausted. At that point every subsequent ball drawn would be green until you reached 100 percent of the green balls which would naturally be obtained only after the last ball was drawn. Thus the progression of the good curve for a perfect model would follow the x axis until reaching $x = \textit{bad rate}$ then rising in a straight diagonal to the point $(1, 1)$ at a slope of $1/(1 - \textit{badrate})$ or $1/(\textit{goodrate})$.

3.4. Cumulative Gains Model Quality

As can be seen in Figures 7 and 8. KS = 1 for a perfect model is always obtained at $x = bad\ rate$ and everywhere else separation of good and bad must be less than this. Extending the logic used for KI, it then becomes a natural extension to consider the area between the perfect good and bad curves as being the ideal curve. Dividing the area between the actual model good and bad curves by the area between the perfect good and bad curves would seem to be a good extension of KI. However this leads to the same problems discussed with KI. This was solved by altering KI in such a way to offer a function which described a point estimate of the ratio between actual and perfect at any point. In order to do this you simply need to divide the separation at that point which we will call $\mathrm{ks}(x)$ by the perfect separation $p(x)$. Intuitively $q(x)$ becomes a normalized separation statistic always between 0 and 1:

$$q(x) = \frac{\mathrm{ks}(x)}{p(x)}. \tag{3.5}$$

In itself it becomes a good second look at separation to supplant KS. However since we are also interested in the performance of the entire model or the performance of the model over a specified range of risk, integrating this function over the whole range of the model becomes a model statistic improvement over KI and integrating over a specific range of x values becomes an improvement to KI*. We therefore define Q to be

$$Q = \int_a^b \frac{\mathrm{ks}(x)}{p(x)} dx = \int_a^b q(x) dx, \tag{3.6}$$

where (a, b) is a region of risk in which we are interested. If chosen to be (0,1), Q becomes a statistic on the whole model.

There is one last thing to consider. One reason KXEN's KI statistic is so successful is that it is easy to interpret. It is always between 0 and 1. This makes it an excellent statistic when interpreting results. As we have shown in this paper one of the primary problems with KS is that it does not land within any consistent range of values. By not being a standardized statistic, it loses ability to compare one model to another. We have specifically chosen two models which have differing KS values due to different definitions of good and bad. We could have just as easily chosen two models with different risk populations and the same bad definition yielding different bad rates. In both cases the challenge is to strive for a statistic which can compare the relative effectiveness of the models.

In this case we have achieved our goal. The distribution of q is analogous to a normalized KS point estimate of model effectiveness, and if you choose (a, b) to be (0,1) on the whole model you achieve a good metric on the whole model. Choosing (a, b) other than (0,1) also allows us to compare two models easily. However what it loses is the guarantee that Q will fall between 0 and 1 and thus loses some of what we were striving for. To fix this consider the Mean Value Theorem from elementary calculus. By applying this theorem to our problem we can thus guarantee that all stated goals can be achieved by any segment of the model. This statistic is given as

$$\mathrm{MV}Q = \frac{Q}{b-a}. \tag{3.7}$$

4. Discussion

As we have mentioned MVQ becomes a rather useful tool to assess the quality of a model. We plot q in Figure 9, which again is the same as plotting the KS separation at each point but dividing by the separation of a perfect model.

Notice how you can now get a feel for the separation with respect to what was possible. It also levels the playing field in such a way as to indicate that maybe Model 2 actually outperforms Model 1 in certain regions of the model. Specifically in the lower deciles. As already noted there are modeling situations when this performance becomes key. We could then calculate MVQ for the entire model, which turns out to be $\mathrm{MV}Q_1 = 0.46$ and $\mathrm{MV}Q_2 = 0.40$. For illustration to understand the performance of the model over regions of the model, we plot $\mathrm{MV}Q(x)$ taken to be MVQ on interval $(0, x)$, In Figure 10. Note that this plot accentuates the performance advantage of Model 2 in the first half of the population.

4.1. Model Stability Metric

The KXEN statistic KR is an extention of KI as already noted. By similar extension of MVQ we can similarly create a model stability metric. When creating models it is common practice to separate a data set into: data the model is built on, and data the model is validated on. One of the easiest and most elegant ways to test for model stability and overfit is to compare the lift statistics of the model build data versus the lift statistics of the validation data. In this case we want to compare MVQ(build) versus MVQ(validation). The closer the ratio of these two is to 1, the more stable the model. We define the Model Stability Metric (MSM) to be

$$\mathrm{MSM} = \frac{\mathrm{MV}Q(\mathrm{validationdata})}{\mathrm{MV}Q(\mathrm{builddata})}. \tag{4.1}$$

4.2. Applicability

We feel that this is a valuable new tool in model performance metrics. As Marcade [2] notes this type of metric is not only useful as a visual or manual interpretation but can be used in a number of different algorithmic ways in a similar fashion to entropy for machine learning. As an example Ferri et al. [1] pointed out you could use area under the ROC curve to not only assess the quality of a decision tree but in the creation of the splits during the algorithm. By extension MVQ can be used in a similar fashion. MVQ and q offer several key features of a model quality statistics including what follows.

(i) $q(x)$ can be used as a point estimate of separation of model curves which is superior to KS in that it is normalized between 0 and 1. This gives it an inherent advantage to understand model performance in that it is very intuitive. The normalization between 0 and 1 also proves superior in different risk populations of financial modeling where the underlying risk of the population may differ substantially, we have demonstrated that KS would necessarily differ from among scenarios whereas $q(x)$ would be a pure "percent of perfect". Results can more consistently be interpretted.

(ii) MVQ can be used as a whole model lift statistic by choosing the interval (a, b) to be (0,1). It can also be used to measure the useful range of the model where financial decisions will take place on any interval (a, b) without loss of generality. As an example a risk model may have pro forma financial criteria. The cut point in the population where credit will be extended is based on this. We may only be interested in comparing two models below this cut point. Above this performance cut point the two models may actually cross in performance but that is of no concern as we have predetermined that both scenarios are unprofitable above a certain risk.

(iii) MVQ and q both give weight to higher performance early in the population selection ranking, whereas other measures such as KS have no ability to discern a "percent of perfect" rating. This allows for greater weight on the left for higher performing models. This is a very key achievement for practitioners of risk management. In the case of 2 models with similar KS the model where the separation is further to the left will inherently be given a higher value of q and MVQ.

(iv) It is agnostic in its assumptions of the model. Credit risk management does not use traditional measures of model goodness such as P-value. Instead various lift metrics are often used. MVQ is free from any assumptions regarding the model or population itself and is only concerned with how accurately we successfully predicted the endpoint. In the case of a simple logistic regression, the choice of 0 and 1 and the predicted probability is obvious. Many other types of models easily lend themselves to this kind of interpretation however. In a Markov model, often used for risk management, one only has to determine which states of the Markov model are a success and which are failure creating a 0 and 1 outcome. Similarly financial models may have a continuous dependent variable such as profit. Decision criteria of the business will yield a threshold of profit which is acceptable and not acceptable. This leads to a 0 versus 1 endpoint. In practice these financial decisions are regularly made and it is only up to the model to predict which side of the threshold a particular credit account falls. It does not matter whether logistic regression, multinomial regression, Markov models, clustering, neural nets, or trees (all important models in risk management) are used. All may be condensed into whether the predicted outcome meets financial criteria or not and will submit to the 0 versus 1 test. So in fact this allows you to compare the use of say a neural net on a data set versus a logistic model versus a regression tree and have a meaningful way to compare the results.

The ability to better understand the implications of risk in the financial world is of utmost importance. It is precisely the lack of this ability which has been implicated in the current financial crisis. Practitioners are often left however with a confusing myriad of models which do not lend themselves to traditional measures of model quality taught in classic statistics classes. Ultimately the goodness of a model can only be determined by the proportion of the time it is correct in predicting the outcome. Current methods of doing this such as KS are not complete or intuitive and do not have consistent normalized measure from model to model. MVQ answers these issues. At a practical level, the authors feel that q and MVQ are especially useful in interpreting results at the decision level of risk management where nonstatisticians become intimately involved.

References

[1] C. Ferri, P. A. Flach, and J. Hernández-Orallo, "Learning decision trees using the area under the ROC curve," in *Proceedings of the 19th International Conference on Machine Learning (ICML '02)*, pp. 139–146, Sydney, Australia, July 2002.

[2] E. Marcade, "Evaluating modeling techniques," Technical Whitepaper, KXEN, Inc., San Francisco, Calif, USA, 2003.

[3] N. Siddiqi, *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, Hoboken, NJ, USA, 2006.

[4] E. Mays, *Credit Scoring for Risk Managers: The Handbook for Lenders*, South-Western, Mason, Ohio, USA, 2004.

[5] G. Piatetsky-Shapiro and S. Steingold, "Measuring lift quality in database marketing," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 2, pp. 76–80, 2000.