ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Iwona Schab**

RANKING-BASED CHOICE OF REGRESSORS IN PROBABILITY MODELS

Abstract. The article presents a proposal of using the Receiver Operating Characteristic (ROC) and Cumulative Accuracy Profile (CAP) curves as a ranking-based method for a choice of regressors in probability model. The criterion of reggresors' choice uses the value of summary statistics of discrimination based on ROC/CAP curves as well as it takes into account the shape of the curves itself.

Key words: Receiver Operating Curve (ROC), Cumulative Accuracy Profile (CAP), AUROC statistics, Gini statistics, choice of regressors.

I. THE ISSUE AND THE ASSUMPTIONS

One of the key issue in the statistical modeling is the choice of regressors. Let us consider a probability model explaining the occurrence of an event described by a binary variable Y and a potential regressor to specify the model X. Let us assume X is a continous variable negatively correlated with Y taking value $y_j = 1$ for occurrence of the event and $y_j = 0$ for non occurrence. Negative correlation means the bigger X value is the less probable is the event. Occurrence of an event will also be called a positive event (regardless of the nature of the even, it can be for example default or failure) and non occurrence – the negative event respectively.

Modeling the probability of an event can be regarded as a classification problem. Let us consider an example of bank's client defaulting on a credit obligation. Every client belongs to one of two populations: Π_0 for those who repay the debt contractually (negative event) or Π_1 for those who default (positive event). Therefore a default variable Y is defined as:

$$y_j = \begin{cases} 1 & \text{if } j \in \Pi_1 \\ 0 & \text{if } j \in \Pi_0 \end{cases} \tag{1}$$

* M.Sc., Institute of Statistics and Demography, Warsaw School of Economics.

where:
y_j – default indication for j-th client, $j = 1, \ldots n$.

At the moment of credit decision only regressor X is known (in practice the set of X's variables) also called a diagnostic variable. Client's classification to Π_1 or Π_0 is know only *a posteriori*, whereas *a priori* – at the moment of credit decision – the probability of default event $Y = 1$ can be estimated and therefore presumptions concerning client's assignment to Π_1 or Π_0 can be made. The better the model is the better is the *a priori* classification and therefore more correct decisions are made. One of the necessary conditions for estimation of a good model is its correct specification in terms of variables. It applies to the situation in which the set of regressors is not known from the economic theory and must be decided by the researcher on the basis of empirical data.

The basic postulate against the regressor is its ability to explain the phenomena being modeled which results in stochastic dependence between regressor X and dependent variable Y. Stochastic dependence is defined by the difference in the conditional distributions, e.g. cumulative density functions $F(X|Y)$. In case of binary variable Y the stochastic dependence between X and Y as well as the degree of the difference of between $F(X|Y = 0)$ and $F(X|Y = 1)$ means the strength of X discrimination in respect to Y. It can be assessed by the measures of discrimination which can be used as an alternative way to chose the regressors for a probability model $P(Y) = f(X)$.

II. THE RECEIVER CHARACTERISTOC CURVE AND THE CUMULATIVE ACCURACY PROFILE

The concept of Receiver Operating Curve (ROC) was first introduced in signal detection theory. It originates also from psychology and especially medicine, Hanley and McNail (1982). Over the last few years the concept of ROC curves found interest in machine learning and data mining area as a tool for model evaluation.

The ROC curve plots values of conditional cumulative distribution functions: $F(X|Y = 0)$ against $F(X|Y = 1)$ over varying threshold x. The empirical ROC curve is a plot of empirical CDFs: $F_n(X|Y=0)$ against $F_n(X|Y = 1)$. In different notation it is a plot of the true positive rate ($F_n(x|Y=1)$ – rate of correctly classified $Y = 1$ with threshold x) over the false positive rate ($F_n(x|Y = 0)$ – rate of incorrectly classified $Y = 0$). The ROC curve shows the ability of X variable to discriminate between two classes of Y. The higher the ability is the more concave is the curve since high rates of correctly classified $Y = 1$ are matched to low rates of incorrectly classified $Y = 0$ for the same threshold x. The example of the empirical ROC plot is presented on the Figure 1.

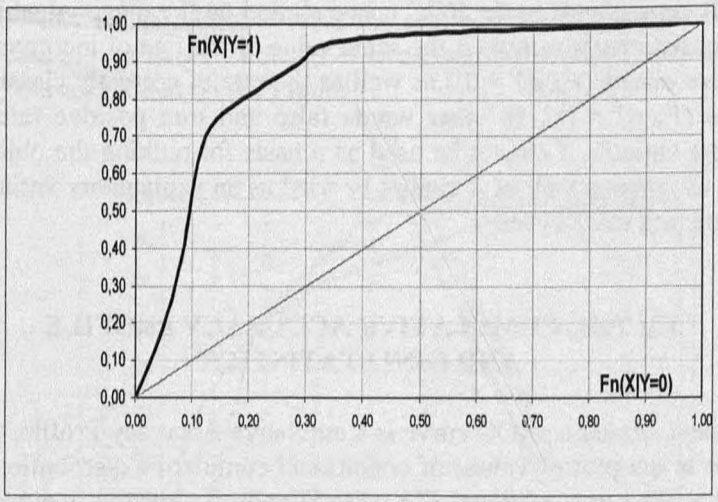

Figure 1. ROC curves for a continuous X variable

The summary statistics measuring the discriminative power of X is the area under the curve, called AUROC statistics. It measures the ranking quality of the variables X in respect to Y. The higher the value of X is the more probable the $Y = 1$ events, so the ranking based on the X values alone is the same as the ranking of probabilities P(Y=1)=f(X) based on a probability model.

The AUROC statistics takes values from 0 to 1 and can be interpreted in probabilistic terms. It is an estimate of the probability that a randomly chosen pair (i, k) of objects with $y_i = 0$, $y_k = 1$ will be correctly ranked by X values, i.e. $x_i > x_k \rightarrow P(Y = 1|X = x_i) < P(Y = 1|X = x_k)$, which means the X value will allow for correct classification over Y for that specific pair (i, k). A perfect ranking gives AUROC statistics equal 1 which means that all positive examples with $y_j = 1$ are ranked lower that the negative ones with $y_j = 0$. On the other hand the minimum value of AUROC statistics of 0 shows a perfect reversed ranking, which means that all positive examples with $y_j = 1$ are ranked higher that the negative ones with $y_j = 0$. In that case: $x_i > x_k \rightarrow P(Y = 1|X = x_i) > P(Y = 1|X = x_k)$ which shows the positive correlation between X and Y. Regardless of the sign of the dependence between both variables the values of AUROC statistics near to 1 or 0 show (extremely) strong ability of X to discriminate over Y which supports the decision of inclusion X variable in probability model specification.

In case of positive correlation between X and Y the ROC curve will be plotted below the diagonal and the ROC statistics will takes values <0; 0,5), negative correlation will give the curve over the diagonal with statistics' values (0,5; 1>. The value of 0,5 shows that a ranking of objects based on X values is a ran-

dom one. It corresponds to the ROC curve plotted on the diagonal which shows for each consecutive threshold x the same value of the rate of incorrectly classified negative events ($F_n(x|Y=0)$) as well as the rate of correctly classified positive events ($F_n(x|Y=1)$). In other words false and true positive rates are the same, so the variable X cannot be used as a basis for ranking the objects in respect of Y variable as well as X cannot be used as an explanatory variable in the model predicting the Y event.

III. THE CUMULATIVE ACCURACY PROFILE AND GINI STATISTICS

A concept similar to ROC curve is Cumulative Accuracy Profile, CAP. The CAP curve is the plot of values of conditional cumulative distribution function $F(X|Y=1)$ against unconditional $F(X)$. Empirically it plots the true positive rate $F_n(x|Y=1)$ – rate of correctly classified positive examples with $Y=1$ against the overall rate $F_n(x)$ of examples cut off by the same x threshold.

A summary statistics derived on the basis of CAP curve is Gini measure which is defined as the ratio of area between CAP and the diagonal to the area between perfect model and the diagonal. Gini measure ranges from 0 to 1. In the case of positive correlation between X and Y which results in reversed ranking the Gini statistics takes values <–1; 0). The Gini statistics and the AUROC are linked via the formula, Engelmann (2006):

$$G = 2 \cdot AUROC - 1 \tag{2}$$

where:
G – Gini statistics.

Since both curves ROC and CAP as well as their summary statistics are closely related to each other further considerations will use ROC curve only. All the conclusions will apply to CAP curve and Gini statistics as well.

IV. AUROC AND ROC CURVE SHAPE IN CHOICE OF REGRESSORS

As mentioned above the AUROC statistics significantly different from the 0,5 value confirms the ability of X variable to differentiate over Y. As a consequence X can be a good predictor for modeling the probability of Y event and therefore can be used in probability model $P(Y) = f(X)$.

Additionally to the AUROC value the shape of the ROC curve can be used in the process of regressors' choice. Let us assume two variables $X1$ and $X2$ with their ROC curves plotted on the Figure 2.

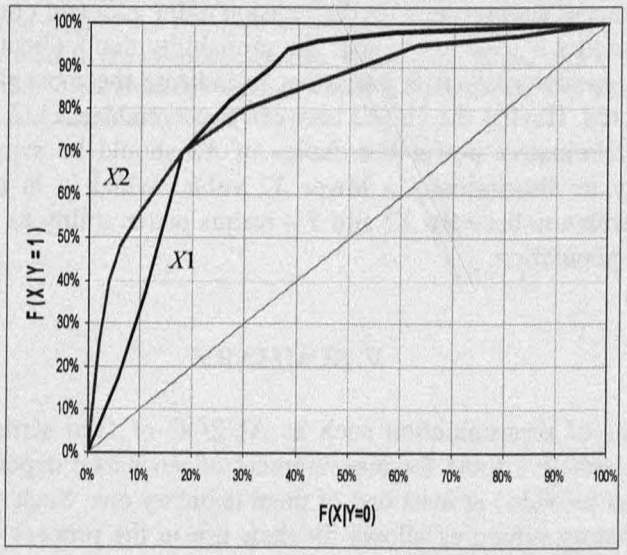

Figure 2. ROC curves for $X1$ and $X2$ variables, AUROC($X1$)=AUROC($X2$).

Let us assume – for simplicity – that both variables have the same domain, so the respective thresholds x take the same values in absolute terms. If it is not the case both variables should be compared by relative thresholds represented by the same distribution quantiles instead of absolute values. Furtherer considerations are valid also for $X1$, $X2$ variables with different domains.

Both variables $X1$, $X2$ from the Figure 2 have the same value of AUROC statistics in respect of the ability to discriminate over Y. Therefore both of them seem to be of the same quality predictors in the probability model. However their ability to discriminate is different on the subparts of $X1$, $X2$ domain. In the range of lower $X1$, $X2$ values the ROC curve for $X2$ shows its better discriminative power. The threshold of X1 which cuts off 10% of the negative ($Y = 0$) events assigns correctly only 30% of all positive events to the Π_1 population, whereas the same 10% threshold of $X2$ assigns correctly 55% of the all positive events to the Π_1 population. On the other hand $X1$ discriminates better in the range of higher $X1$, $X2$ values. In that subpart of $X1$, $X2$ domains the variable $X1$ reaches high true positive rate with lower threshold than $X2$. For example in order to assign 95% of the positive events correctly to Π_1 population as much as

70% of negative events are incorrectly assigned to Π_1 basing on the same $X2$ threshold whereas only 40% using the $X1$ variables.

On average both $X1$, $X2$ variables have the same discriminative power (over all possible thresholds) but their ability to discriminate is different in different subparts of $X1$, $X2$ domains. It has practical implications. Let us consider the probability model supporting a credit decision at the moment client applies for a credit. The model is used to estimate the probability that a client will default on a credit but operationally it is important to indicate the most risky clients that will be rejected. Having the choice between two variables $X1$, $X2$ with the same average discriminative power the choice of $X2$ should be supported since its better ability to discriminate in lower $X2$ values which – in connection with negative correlation between $X2$ and Y – means better ability to indicate the client from Π_1 population.

V. SUMMARY

Measures of discrimination such as AUROC or Gini statistics can be regarded as alternative tools for measurement of stochastic dependency between two variables provided at least one of them is binary one. Such an interpretation of discriminatory measures allows for their use in the process of choosing covariates in probability model.

Additional criterion of covariate choice is the shape of ROC/CAP curves. Depending on degree of concavity and the character of the event being modeled (e.g. default on a debt, product purchase) one of two covariates can be found superior, although the value of discriminatory measure is the same.

REFERENCES

Cortes C., Mohri M. (2005), Confidence Intervals for the Area under the ROC Curve, *Advances in Neural Information Processing Systems*, 17, 305–313.

Engelman B., Hayden E., Tasche D. (2003), Measuring the Discriminative Power of Rating Systems, *Deutsche Bundesbank Discussion Paper*, 1/2003, 1–24.

Engelman B., Rauchmeier R. (2006), *The Basel II Risk Parameters. Estimation, Validation and Stress Testing*, Springer Verlag, Heidelberg.

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

Rossa A. (2004), Classification tree based on Receiver Operating Characteristic Curves, *Acta Universitatis Lodziensis Folia Oeconomica*, 2004, 113–121.

Iwona Schab

DOBÓR ZMIENNYCH OBJAŚNIAJĄCYCH W MODELACH PRAWDOPODOBIEŃSTWA W OPARCIU O KRZYWE ROC ORAZ CAP

W artykule przedstawiono propozycję wykorzystania krzywych ROC (Receiver Operating Characteristic) i CAP (Cummulative Accuracy Profile) w doborze zmiennych objaśniających w modelu prawdopodobieństwa. Kryterium doboru zmiennych opiera się na wartościach miar dyskryminacji wyznaczonych na podstawie krzywych ROC/CAP jak i uwzględnia sam kształt krzywych.