TOOLS FOR EVALUATING CLASSIFICATION MODELS

Professor Thomas B. Fomby
Department of Economics
Southern Methodist University
Dallas, Texas

February, 2008
Revised March, 2010
Revised April, 2010
Revised April, 2012

There are several tools that practitioners frequently use to evaluate the performances of competing classification methods. The particular tools used depend on how much information the evaluator has concerning the consequences (gains and loses) associated with using classification methods. We will pursue our discussion assuming, first, that we have complete information on these consequences, then proceed to discuss methods of evaluation based on less and less information concerning consequences as we proceed through our discussion. But first we need to define some concepts that relate to characterizing the consequences of making choices concerning classifications.

1. Binary Choice Problems and Decision Tables

In the following discussion we will be focusing exclusively on the evaluation of **binary** classifiers. That is, we will be interested in evaluating the performances of classifiers that deal with predicting one of two possible outcomes: 1 = a "success" or 0 = a "failure." Of course, one should not take the terms "success" too literally. For example, a success could be the graduation of a high school student versus a failure when a student does not graduate from high school. In contrast, a "success" could be the default of a person on his/her credit card debt versus a "failure" of a person never defaulting on his /her credit card debt. Usually the event of primary interest to the investigator is defined as the "success" event.

Given this binary classification problem, we might use the following **Outcome Table** to specify the possible outcomes of a given prediction.

BINARY CHOICE

OUTCOME TABLE

		Actual Value	
		1	0
Predicted value	1	True Positive	False Positive (Type I Error)
	0	False Negative (Type II Error)	True Negative

Here we use the somewhat more neutral terms of positive (=1) and negative (=0). From the above Outcome Table we can see that there are **four** possible outcomes that go along with a binary prediction. You predict a 1 and a 1 actually occurs (a true positive); you predict a 0 and a 0 actually occurs (a true negative); you predict a 0 and a 1 actually occurs (a false negative); and, finally, you predict a 1 and a 0 actually occurs (a false positive). If we take the null hypothesis to be the negative (0) case, we can view the false positive outcome as being a Type I error in the statistical sense. On the other hand, if we take the alternative hypothesis to be the positive (1) case, we can view the false negative outcome as being a Type II error in the statistical sense. Note this labeling in the above outcome table.

Corresponding to this Outcome Table, we have a so-called **Cost/Gain Table** represented below:

COST/GAIN
TABLE

		Actual Value	
		1	0
Predicted value	1	P_{11}	P_{10}
	0	P_{01}	P_{00}

Here we use the more comprehensive term “payoff” to represent a gain (a positive monetary outcome) as well as a cost (a negative monetary outcome). That is, the P_{ij} ’s in the Cost/Gain table can either be positive or negative, depending on the actual problem being investigated. Then P_{11} represents the payoff arising from a correct classification of a 1; P_{00} represents the payoff arising from a correct classification of a 0; P_{01} represents the payoff (usually negative) arising from incorrectly classifying a 1 as a 0; and, finally, P_{10} represents the payoff (again usually negative) at incorrectly classifying a 0 as a 1.

Now let us consider how we might go about comparing the predictive capabilities of competing classifiers which have been built on a **training data set**. Suppose there are N cases in the **validation data set**. Furthermore, suppose that we “score” the validation data set using a particular classifier and we get the following outcomes represented in the below, so-called, **“Confusion” Table**:

CONFUSION TABLE

		Actual Value	
		1	0
Predicted value	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

In this table n_{11} represents the number of cases that the classifier correctly classified as 1's in the validation data; n_{00} represents the number of validation cases that the classifier correctly classified as a 0; n_{01} represents the number of validation cases that were 1's but were incorrectly classified as 0's; and finally, n_{10} represents the number of validation cases that were 0's but incorrectly classified as 1's. Of course these outcomes are exhaustive and we have $N = n_{11} + n_{00} + n_{01} + n_{10}$.

2. Classification Accuracy Measures

Now consider the following definitions of various Classification Accuracy Measures:

ACC = Accuracy Rate

= proportion of the total number of predictions that were correct

$$= \frac{n_{11} + n_{00}}{N}$$

ERR = Error Rate (1 – Accuracy Rate)

= proportion of the total number of predictions that were incorrect

$$= \frac{n_{01} + n_{10}}{N}$$

TPR = Total Positive Rate (Sensitivity)

= proportion of positive (1) cases that were correctly classified

$$= \frac{n_{11}}{n_{11} + n_{01}}$$

FNR = False Negative Rate (1 – Sensitivity)

= proportion of positive (1) cases that were incorrectly classified as negative (0)

$$= \frac{n_{01}}{n_{11} + n_{01}}$$

TNR = Total Negative Rate (Specificity)

= proportion of negative (0) cases that were classified correctly

$$= \frac{n_{00}}{n_{00} + n_{10}}$$

FPR = False Positive Rate (1 – Specificity)

= proportion of negative (0) cases that were incorrectly classified as positive (1)

$$= \frac{n_{10}}{n_{00} + n_{10}}$$

P = Precision

= proportion of the predicted positive (1) cases that were correct

$$= \frac{n_{11}}{n_{11} + n_{10}}$$

3. The Naïve Classifier

In discussions of classifier performances there are often frequent references to the so-called **Naïve Classifier**. The Naïve Classifier is that classifier that one would use if one had no information available on the input variables associated with the individual and, instead, only had information on the proportion of successes (and hence proportion of failures) in the **validation data set**. For example, suppose that the proportion of successes in a validation data set is 20%. Then, conceptually, given an individual and a hat with five poker chips in it, one that is red and four that are blue, we could, in the absence of any available information on individuals, use the following **naïve classification rule**: If we draw a red chip from the hat, the individual is predicted to be a success (=1); otherwise the individual is predicted to be a failure (= 0). Obviously this is a very crude classifier. In contrast, if we have some useful information on the individual and if this information is adeptly incorporated into a classifier, we should do a better job in classifying individuals in a validation data set than the Naïve Classifier would do. Therefore, in the **Cumulative Gain Charts, Lift Charts,** and **ROC curves** you will study later, you will often see the Naïve Classifier's performance serving **a benchmark to beat**. If the structured classifiers don't, in some sense, outperform the Naïve Classifier, then the construction of new and more efficient classifiers should be undertaken.

Another way of envisioning the Naïve Classifier is to take the majority classification of the data and apply it in the scoring of **all** of the new data. For example, in a binary choice problem with the "failures" being in the majority (the 0s if you will) one could classify all new cases simply as being failures. In so doing, this would guarantee that all failures in the new data set would be classified correctly but, on the other hand, all successes would necessarily be misclassified. The error rate of the Naïve Classifier would then be equal to the proportion of successes that exist in the new data set. Correspondingly, the accuracy rate of the Naïve Classifier would be equal to the number of failures in the new data set.

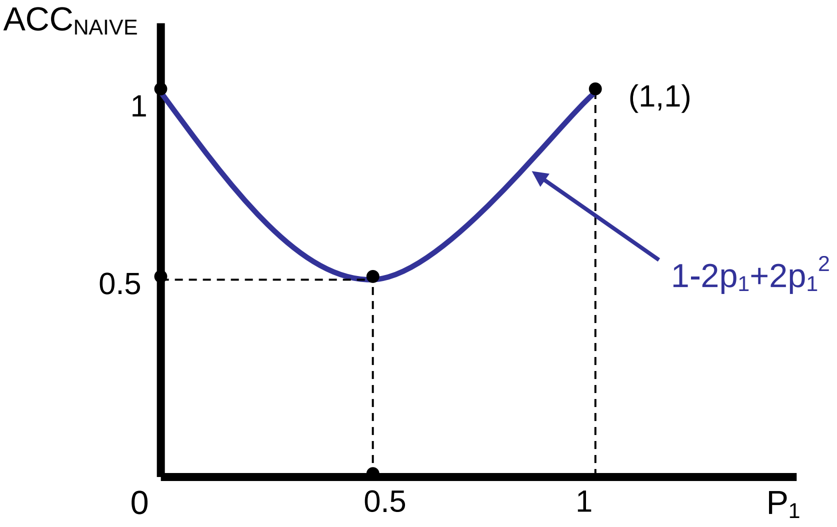
Now if one is comparing potential classifiers with the Naïve Classifier and the **population** proportion of successes **in the validation data** set is p_1, then the **expected** Naïve Classifier's accuracy rate can be computed as

$$
\begin{aligned}
ACC_{NAIVE} &= \frac{E(n_{11}) + E(n_{00})}{N} \\
&= \frac{N \cdot p_1 \cdot p_1 + N \cdot p_0 \cdot p_0}{N} \\
&= p_1^2 + p_0^2 = p_1^2 + (1 - p_1)^2
\end{aligned}
$$

$$= 1 - 2p_1 + 2p_1^2 .$$

In calculating the above formula we have calculated the expected number of correct classifications of successes as $E(n_{11}) = N \cdot p_1 \cdot p_1$ and the expected number of correct classifications of failures as $E(n_{00}) = N \cdot p_0 \cdot p_0$, where $p_0 = 1 - p_1$ is the probability of classifying an individual as a failure. In the calculations of these expectations we have assumed that the actual classification of an individual is independent of the naïve classification of the individual. Then $p_1 \cdot p_1 = p_1^2$ represents the probability that any one success will be correctly classified as a success and $p_0 \cdot p_0 = p_0^2$ represents the probability that any one failure will be correctly classified as a failure. Given these considerations, the **expected** accuracy rate of the Naïve Classifier is plotted in the graph below.

EXPECTED ACCURACY RATE
OF THE NAÏVE CLASSIFIER



P_1 = proportion of successes

Then the **expected** benchmark accuracy rate for comparison with other more sophisticated classifiers is given by the formula $1 - 2p_1 + 2p_1^2$. Obviously, the Naïve Classifier's expected accuracy rate is at a minimum when $p_1 = 0.5$. In contrast, when the proportion of successes becomes more "extreme" $(p_1 \to 0$ or $p_1 \to 1)$ the Naïve Classifier's expected accuracy rate becomes more stringent and tough to beat.

Of course, one rarely has information on the population proportion of successes p_1. Thus, one has to rely on a consistent estimate of it. One such consistent estimate is the sample proportion of successes in the **validation data set**. Let this sample proportion be denoted by $\hat{p}_1$. Then a consistent estimate of the expected accuracy rate of the Naïve Classifier is $A\hat{C}C_{NAIVE} = 1 - 2\hat{p}_1 + 2\hat{p}_1^2$. Now, in comparisons of the accuracy rate of a proposed classifier obtained from the validation data set, say, $ACC_{PROPOSED}$, and that of the Naïve Classifier, $A\hat{C}C_{NAIVE}$, we would initially prefer the proposed classifier if $ACC_{PROPOSED} > A\hat{C}C_{NAIVE}$ and not, otherwise.

However, even if the accuracy rate of the proposed classifier is better than the accuracy rate of the Naïve Classifier in the validation data set, we might ask the following question: "Is the accuracy rate of the proposed classifier statistically superior to that of the Naïve Classifier?" This question can be answered by conducting a one-tailed test of proportions using the following N(0,1) statistic

$$Z = \frac{ACC_{PROPOSED} - A\hat{C}C_{NAIVE}}{\sqrt{\dfrac{A\hat{C}C_{NAIVE}(1 - A\hat{C}C_{NAIVE})}{N}}} \quad .$$

The null hypothesis of the test is

H_0: The Accuracy Rate of the Proposed Classifier Is No
Better than the Accuracy Rate of the Naïve Classifier

versus the alternative hypothesis that

H_1: The Accuracy Rate of the Proposed Classifier Is Statistically
Better than the Accuracy Rate of the Naïve Classifier.

If the Z statistic is positive and if its right-tail p-value is less than a pre-specified size, say $\alpha\%$, one would reject the null hypothesis and accept the alternative hypothesis that the proposed classifier is statistically superior to the Naïve Classifier. Otherwise we would accept the null hypothesis that the two classifiers are not statistically distinguishable.

For example, suppose that the validation data set contains N = 200 cases and that the number of successes in the validation data set is 40. Then the proportion of successes is $\hat{p}_1 = 0.20$. It follows that $A\hat{C}C_{NAIVE} = 1 - 2(0.2) + 2(0.2)^2 = 0.68$. Furthermore, assume that, in the validation data set $ACC_{PROPOSED} = 0.76$. The corresponding Z-statistic then is

$$Z = \frac{0.76 - 0.68}{\sqrt{\frac{0.68(1-0.68)}{200}}} = 2.425.$$

The corresponding right-tail probability for this statistic is p = 0.008 and the difference is highly significant. Thus the proposed classifier is statistically superior to the Naïve Classifier in the examined validation data set.

4. Evaluating Classification Methods and Choosing Optimal Cut-off Probabilities

Let us now consider the following four scenarios concerning **varying degrees of ignorance** with respect to Cost/Gain information, all for the purpose of distinguishing between good and bad classifiers and choosing optimal cut-off probabilities for classifiers based on scoring validation data sets. That is, the following cases are intended to help us come to understand the **crucial role** that the Payoff structure of a classification "game" plays in distinguishing between competing classifiers and between competing cutoff probabilities for a given classifier, all based on the scoring of validation data sets.

- **Case I**: The Payoffs in the Cost/Gain Table are **Completely Known**.
- **Case II**: The Costs (Payoffs) of Misclassification are **Equal** while the Payoffs associated with Correct Classifications are assumed to be **zero**.
- **Case III**: The **Ratio of the Costs** (Payoffs) of Misclassification is **Known** while the Payoffs associated with Correct Classifications are assumed to be **zero**.
- **Case IV**: The Payoffs in the Cost/Gain Table are **Completely Unknown**.

What we must keep in mind at this point is that classifiers depend upon the choice of a "**tuning**" parameter, the so-called "cut-off" probability of the classifier. As we will see, classifiers, given the attributes (inputs) of an individual, generate a probability of "success" for that individual. If that probability is greater than the cut-off probability, the individual is classified as a 1 (success); otherwise, the individual is classified as a 0 (failure). Importantly, the payoff-performance of classifiers is **critically dependent on the choice of the cutoff probability**.

Let $R = \left(\frac{P_{01}}{P_{10}}\right)$ be the ratio of the payoff (cost) of incorrectly classifying a success as a failure, P_{01} , to the payoff (cost) of incorrectly classifying a failure as a success, P_{10}. Then given the payoffs of correct classification P_{11} and P_{00}, the **greater** (less) the value of R, the **lower** (higher) the optimal cut-off probability should be. **Certainly the cut-off probability of 0.5 is not sacrosanct**. It needs to be determined on a case-by-case taking into account the available information concerning the costs and gains of classification decisions. In this sense, the cut-off probability is a **tuning parameter** for classifiers.

Case I: The Payoffs in the Cost/Gain Table are Completely Known

In this case it is quite easy to evaluate the performance of classifiers when scored on the validation data set. All we have to do is calculate the total payoff (P) associated with scoring the classifier on the validation data set,

$$P = n_{11} \cdot P_{11} + n_{10} \cdot P_{10} + n_{01} \cdot P_{01} + n_{00} \cdot P_{00} \text{ .}$$

With respect to a given classifier, **the optimal cut-off probability** is that cut-off probability that produces the largest total payoff for the classifier in the validation data set. With respect to the **comparison of classifiers**, one would prefer the classifier whose optimal cut-off probability produced the largest total payoff in the validation data set among the competing classifiers using their own optimal cut-off probabilities.

Case II: The Costs (Payoffs) of Misclassification are Equal while the Payoffs associated with Correct Classifications are assumed to be zero.

Now consider the case where the costs of misclassification are equal while the payoffs associated with correct classifications are assumed to be **zero**. In this case, $P_{11} = P_{00} = 0$ and $P_{10} = P_{01} = \bar{P}$, say. Therefore, the Payoff function is of the form

$$P = n_{11} \cdot P_{11} + n_{00} \cdot P_{00} + \bar{P} \cdot (n_{10} + n_{01}) = \bar{P} \cdot (n_{10} + n_{01}) \quad . \qquad (1)$$

Then the misclassification rate is the only thing that matters in this case when comparing the performance of two competing classifiers. The classifier that has the smallest sum of false positives and false negatives $(n_{10} + n_{01})$ is the better classifier. Equivalently, in this case, the classifier with the smallest error rate,$ERR = (n_{10} + n_{01})/N$, (and hence the largest accuracy rate $ACC = (n_{11} + n_{00})/N$) is the better classifier.

Now suppose, instead of comparing the performance of two competing classifiers, we are interested in comparing the performance of a given classifier with different cutoff probabilities Invariably, when choosing a classifier based solely on maximizing the accuracy rate (or equivalently minimizing the error rate), **the cut-off probability that should be used is 0.5**. There is no advantage is deviating from 0.5 since the costs of the misclassifications are equal to each other, $P_{01} = P_{10} = \bar{P}$ and R = 1. Since the payoffs of correct classification are assumed to be zero and the costs of misclassification are equal, the manipulation of the cutoff probability to trade off false positives with false negatives and vice versa is a totally offsetting exercise. **Therefore, in computer programs that, by default, only use a 0.5 cut-off probability and only report confusion tables and accuracy and error rates, it is implicitly being assumed that the Case II circumstances (zero payoffs from correct classification and symmetric costs of misclassification) are applicable in the case under investigation.**

Case III: The Ratio of the Costs (Payoffs) of Misclassification are Known while the Payoffs associated with Correct Classifications are assumed to be zero

In this case we assume that the ratio of the misclassification payoffs, say $R = (P_{01} / P_{10})$, is known while the payoffs of correct classifications are assumed to be zero. Therefore, the payoff function becomes

$$P = P_{10} \cdot (n_{10} + R \cdot n_{01}) \quad . \qquad (2)$$

Assuming an unknown $P_{10} < 0$, we can see that a good classifier is one that minimizes the weighted average of the misclassifications, $(n_{10} + R \cdot n_{01})$, or equivalently, minimizes the **weighted error rate**, ERR(weighted) = $(n_{10} + R \cdot n_{01})/N$.

Case IV: The Payoffs in the Cost/Gain Table are All Unknown

Assume that the payoffs in the Cost/Gain table are all unknown. In this case specific payoffs cannot be computed over the validation data set. One pragmatic thing to do is simply to revert to choosing between classifiers using a 0.5 cut-off probability based solely on the Accuracy Rate (ACC) previously defined. Then the classifier that produces the largest accuracy rate, when scored over the validation data set, is declared the winner. The greater the accuracy rate of the classifier, the better the classifier. Also, in this case of no information, we can consider additional performance tools for evaluating classifiers, namely,

- Cumulative Gains Charts
- Lift Charts
- The ROC Space and ROC Curves

For further discussion of these evaluation tools, see the following pdf files posted on the class website:

- Lift Charts.pdf
- ROC.pdf

The above two pdf files were obtained from the website for the Department of Computer Science, University of Regina, Regina, Saskatchewan, Canada and their computer science class CS 831. (Authors: Howard Hamilton, Ergun Gurak, Leah Findlater, and Wayne Olive, July 2003) http://www2.cs.uregina.ca/~dbd/cs831/

5. The Evaluation and Utilization of Classifiers in the Typical Target Marketing Scenario

In this section we are going to consider the evaluation and proper utilization of a classifier in the situation one might describe as a **target marketing problem**. Suppose that we work for a catalog sales company that wants to maximize the return it gets by sending customers product catalogs in the anticipation of getting orders from them. When considering the evaluation of competing classifiers over a validation data set, we are, in this case, more concerned with the total profits we can generate given a sometimes limited budget for producing and mailing the catalogs.

To make matters concrete, let us assume that, conditional on purchase, our customers spend an average of R (revenue) dollars with us. Moreover, assume that the cost of producing and mailing our catalogs per customer is C dollars. Then the Cost/Gain (Profit) Table is given by (here we assume R > C):

COST/GAIN TABLE
FOR TARGET MARKETING CASE
(1 = Purchase, 0 = No Purchase)

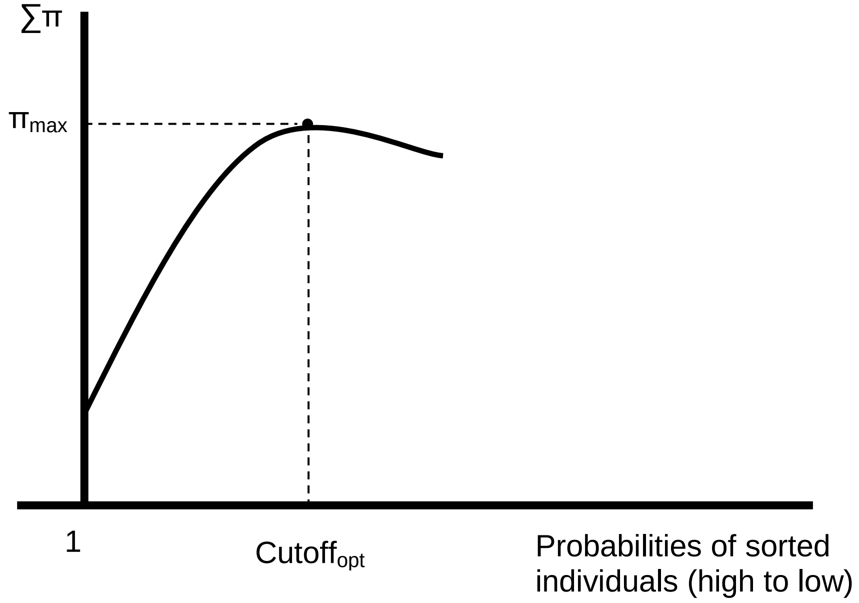
		Actual Value	
		1	0
Predicted value	1	R-C	-C
	0	0	0

Notice that we make the assumption that the cost of missing a potential customer is equal to 0.

Given the above profit matrix we can evaluate competing classifiers as well as "tune" the cut-off probability of a given classifier by simply calculating the cumulative profits generated over the validation data set in the following way: (1) Take the probabilities of purchase (=success) produced by the classifier for the individuals in the validation data set and then sort these individuals from the highest probability of purchase to the lowest probability of purchase. Then using these sorted individuals, score them cumulatively vis-à-vis the above profit matrix. An **idealized** version of the scored cumulative profit curve might look something like the below graph:

CUMULATIVE PROFIT CURVE

SCORED ON VALIDATION DATA SET



$\sum\pi$ = Cumulative Profit

This is, of course, the cumulative profit graph for a given classifier as the individuals are sorted from the highest probability of success to the lowest probability of success. As the highest probability individuals are successively chosen, their prevailing successes in the validation data set lead to a rapid accumulation of profits for the company. However, there is a point of diminishing returns as the lower probability individuals will begin to "fail" and not purchase from the catalogue. Then for choosing an optimal cut-off probability for the given classifier, we should choose as the cut-off probability the probability that maximizes the cumulative (total) profit of the classifier in the validation data set.

Of course, this process can be repeated for other classifiers as well. Then the **overall best classifier** among a set of competing classifiers is the one that, while using its optimal cut-off probability, produces the greatest cumulative profit among the competing classifiers over the validation data set. We then have our best classifier for the Target Marketing problem.

Using a Classifier to Score a New Data Set of Potential Customers

Finally, let us consider the target marketing problem where we have M new potential customers that we are considering sending our catalogs to. What do we do? Well, we take what we believe to be our best known classifier and its optimal cut-off probability and we generate the probabilities of purchase for new potential customers and then we sort them from the highest probability of purchase to the lowest probability of purchase. Before we proceed in our solicitation we need to consider two possible cases:

- **Case 1**: Assume we know, from previous sales experience, the average revenue we can expect from a purchase, say R. Also assume the cost of producing and mailing each catalog is C. Refer to the Cost/Gain Table for the Target Marketing Case above.
- **Case 2**: Assume, from previous sales records, we are able to specify (estimate) an **expected expenditures equation** for purchasers that is a function of the purchasers' attributes (input variables). Also assume the cost of producing and mailing each catalog is C. This Case is sometimes called the **Two-Part Target Marketing problem**.

Case 1: Average Revenue Per Purchase is Known

Let us first consider Case 1. From our previous exercise we know which classifier produces the maximum cumulative profit in the validation data set. We also know the optimal cut-off probability that generated the maximum profit. Assuming that the chosen classifier will do just as well in scoring the new data set as was done in scoring the validation data set, we score the probabilities of purchase of the individuals in the new data set and order them from highest probability of purchase to lowest probability of purchase. Then we entertain the possibility of advertizing to all individuals whose probability of purchase is equal to are greater than the optimal cut-off probability determined in the validation data set for the classifier. Let that number of people be denoted by J_{cutoff} .

Of course being able to send catalogs to the entire J_{cutoff} sorted-set of people depends on the size of the available advertising budget. If the advertising budget exceeds the cost of sending out catalogs to all J_{cutoff} people, i.e. Advertising Budget $\geq J_{cutoff} \cdot C$, then we will send out J_{cutoff} catalogs and return the unused advertising money to the company. In contrast, if our advertising budget is less than $J_{cutoff} \cdot C$ then we advertise to as many individuals (from highest to lowest probability) as our budget will allow.

Case 2: An Expected Expenditure Equation is Available: Two-Part Targeting Problem

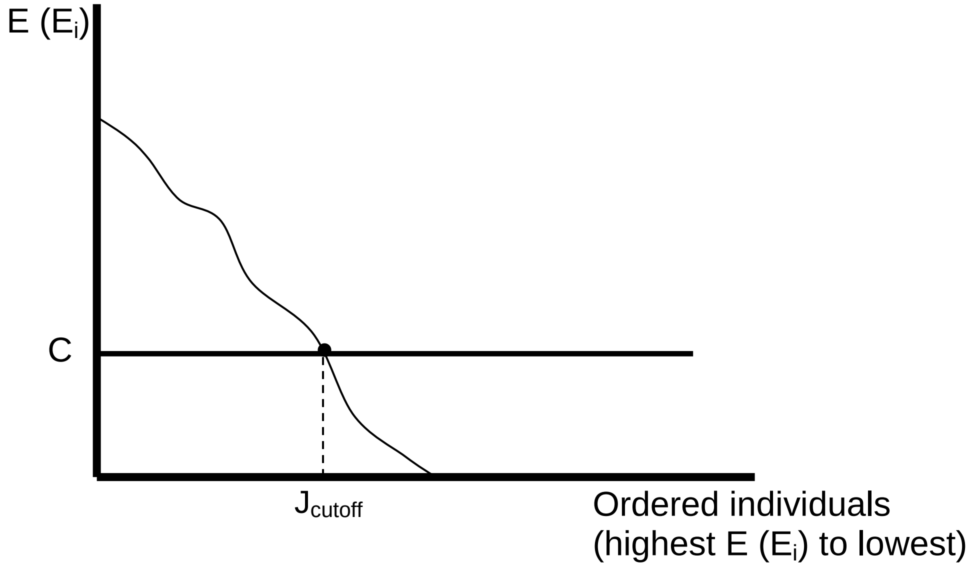
Unlike Case 1, there may be instances when we have more information on how a person's characteristics are likely to affect the amount of merchandise that he/she orders from the company. Suppose, from previous experience with the expenditures of other customers, we are able to determine an expected expenditure equation of the form

$$E_i = f(X_{i1}, X_{i2}, \cdots, X_{iK})$$

where, conditional on purchase, E_i represents the expected expenditure of the i-th individual with the personal characteristics $X_{i1}, X_{i2}, \cdots, X_{iK}$. Then, given the probabilities of purchase of the new potential customers, we can sort over individuals, from highest to lowest, by their **unconditional expected expenditure**, say, $E(E_i) = E_i \cdot P_i$. Then, assuming a sufficient advertising budget, we would advertise sequentially through our sorted individuals until the unconditional expected expenditure of the last individual is just equal to the cost of soliciting that individual. One can represent this decision rule in the below idealized graph.

We should note that the "best" classifier to use in this case may not necessarily be the best classifier for the Case 1 problem above. Recall in Case 1 we choose the classifier that produced the maximum cumulative profit in the validation data set based on the postulated Cost/Gain table above. In the present Case we are simply interested in choosing a classifier that does very well in the Two-Part Targeting problem. Therefore, when choosing a "best" classifier for the Two-Part Targeting problem, we should probably run several two-part experiments on the validation data set, one for each proposed classifier. That is, in conjunction with the adopted expenditure equation, we should determine which classifier produces the most accurate estimate of actual expenditures by customers in the validation data set and then use that classifier in subsequent two-part target marketing problems.

UNCONDITIONAL EXPECTED EXPENDITURE GRAPH


E (E$_i$)
C
J$_{cutoff}$
Ordered individuals
(highest E (E$_i$) to lowest)