Predicting Companies Delisting to Improve Mutual Fund Performance

TA-WEI HUANG

EUGENE YANG

PO-WEI HUANG

BADM

BADM GROUP 6

Executive Summary

Stock is removed from an exchange because the company for which the stock is issued, whether voluntarily or involuntarily, is not in compliance with the listing requirements of the exchange. Companies that are delisted are not necessarily bankrupt, but most of bankrupt company will be finally delisted from the exchange. To earn extra high returns on the stock market, mutual fund managers in Taiwan sometimes invest in high risk companies that might to be delisted in one year. However, once those companies get delisted, mutual funds managers will suffer from significant losses because most of those companies will confront a drastic decline in stock prices before delisted from the exchange.

To prevent mutual funds managers from investing in those potentially-delisted stocks, it is definitely very useful to build a system that predict whether a company will be delisted after one years. Therefore, our goal is to predict whether a company would be delisted in one year. We use 2012 financial reports of non-deliested companies and financial reports of companies one year before its delisted in Taiwan stock market to derive a supervised classification model predicting whether the company will be delisted in 1 year.

By trying K-nearest neighbors, ada-boosting classification tree, and logistic regression and embedding a cost function, we finally chose the logistic regression with cutoff probability 0.65 as our final model. We also use the "130-30" portfolio strategy to compare our prediction result with the market, and we outperform the market in expected return. However, we could not guarantee the stableness of this model during the financial crisis. Investors should still be aware of major economic situations that could cause the model fail. In addition, investor's psychology could also misuse this model and self-fill the delist of a misclassified company.

1 Problem Description

Business Goal

Stock is removed from an exchange because the company for which the stock is issued, whether voluntarily or involuntarily, is not in compliance with the listing requirements of the exchange. Companies that are delisted are not necessarily bankrupt, but most of bankrupt company will be finally delisted from the exchange. To earn extra high returns on the stock market, mutual fund managers in Taiwan sometimes invest in high risk companies that might to be delisted in one year. However, once those companies get delisted, mutual funds managers will suffer from significant losses because most of those companies will confront a drastic decline in stock prices before delisted from the exchange. To prevent mutual funds managers from investing in those potentially-delisted stocks, it is definitely very useful to build a system that predict whether a company will be delisted after one years.

Data Mining Goal

Our job is predicting whether or not a company in Taiwan will be delisted after one year. Therefore, we will build a **supervised classification model**, and the output of our model is the dummy variable $\texttt{Delisted}$, where $\texttt{Delisted} = 1$ if delisted and $\texttt{Delisted} = 0$ otherwise. Ranking probabilities of getting delisted also helps fund managers to find interesting stocks and improve investment decisions.

2 Data Description

Our data comes form the TEJ database, which is the largest financial database in Taiwan. All NTHU students and faculties have a free access to that database, and most of companies in financial industires register this database.

At the first time, we have total 23 columns. The first column is the name of one company. The second column is the date of a companies' financial report. The third column our output variable $\texttt{Delisted}$. The third column is the date of that records. By our domain knowledge, the following 15 columns contain important performance measures of one company. All of the financial variables are ratios in order to avoid scale-varying problem. We also use 5 important macroeconomic variables as our columns to solve the time-varying problem. Later we will explain why we drop out the five economic variables in our final model.

Our records have 830 non-delisted companies in 2012, and 91 delisted companies from 2006 to 2012. The sample of our data is shown below, and the full name of each variable are shown in Appendix A.

Table 1: Sample Data (5 rows and 10 columns)

Company	Date	Delisted	EPS	ROE	GPM	PM	GPMGR	TRGR	RGDP GR	WPI GR
3651 F-天鵬	2012/12/28	1	0.45	3.47	11.26	1.5	96.91	72.36	4.07	-1.16
5296 台矽能	2012/12/28	1	0.02	-2.92	-26.98	-193.33	-110.8	-86.52	4.07	-1.16
1101 台泥	2012/12/28	0	2.09	7.63	14.82	10.94	-13.93	0.81	4.07	-1.16
1102 亞泥	2012/12/28	0	1.93	6.85	8.87	5.69	-48.6	-6.75	4.07	-1.16
1103 嘉泥	2012/12/28	0	-1.13	-3.96	3.99	-7.04	-81.19	-10.17	4.07	-1.16

3 Data Preparation

First, there are missing values in some columns, and we try to handle them by checking companies' real financial report and calculate those ratios by ourselves. There are still 8 records, however, have missing values since companies' disclosing policies. So we use median of financial ratios of companies in the same industry to handle missing values. Some visualizations are shown below.

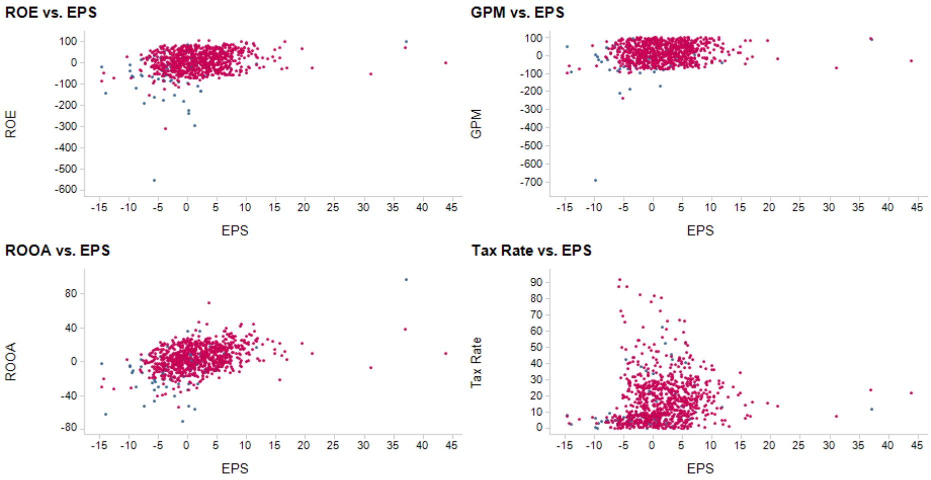


Figure 1: Visualizations of Some Selected Variables

The second problem is the variable selection problem. In the beginning we apply 20 variables, including 15 companies' financial ratios and 5 economic indexes, as our inputs. After running of some algorithms, we find out the predict accuracy is quite great in both training and validation sets, but after running on the 2013 test data, we find out the predict accuracy is dramatically low, that is, the there are over-fitting problems. The results of running that dataset are shown in Appendix B. After dropping out the 5 economic variables, we have a more robust result. Therefore, we use only the 16 financial ratios as our inputs.

4 Data Mining Solution

Algorithms

First, we partition our dataset into two subsets, 60% training data and 40% validation data. We also have a holdout test set, which is the set of financial ratios in 2012, with 842 listed companies and 7 delisted companies in 2013. Our output to predict, `dslisted`, is a categorical variable, and therefore we apply supervised classification models, including *K-nearest neighbors*, *ada-boosting classification tree*, and *logistic regression*. We exclude the Naive Bayes method because we have many numerical inputs that are hard to be binned. The confusion matrices, lift charts, and ROC curves of each algorithm on different datasets are shown in Appendix C, with all the same cutoff probability 0.50. However, there are still three problems: What are the optimal cutoff probabilities of each algorithm? If we change the cutoff, does logistic regression outperform? Is there any asymmetry of costs of misclassification?

Cost Function

To evaluate performances of different cutoff probabilities and mining methods, we define the cost function of misclassification as

$$C(p) = \mathbb{E}(R_0)P(C_0)err_0(p) + \mathbb{E}(R_1)P(C_1)err_1(p),$$

where $\mathbb{E}(R_i)$ is the historical average return of companies' stocks with `delisted` $= i$, $P(C_i)$ is the estimate proportion of companies with `delisted` $= i$, and $err_i(p)$ is the classification error rate of class i, which is a function of the cutoff probability p.

The logic behind the cost function is a simple investment strategy. If one company is predicted as potentially-delisted in 1 year, we will short sell 1 share of its stock for one year; on the other hand, we will buy 1 share of its stock for one year.

Under this rule, we are able to determine the misclassification costs. If the company is predicted as non-delisted while the actual result is non-delisted, the misclassification cost is the negative return of one-year average return on stocks of delisted companies, $\mathbb{E}(R_1)$, because of the long position. Similarly, if one company is predicted as delisted while the actual result is non-delisted, the misclassification cost would be the one-year average return on stocks of non-delisted companies, $\mathbb{E}(R_0)$, because of the short position. Using the historical estimation in 2012, we have $\mathbb{E}(R_0) = 6.26\%$ and $\mathbb{E}(R_1) = 52.1\%$.

We also use the historical estimation from 2006 to 2012 to get the approximate proportion $P(C_0)$ and $P(C_1)$. Then we find that 1% of the current listed companies will get delisted after one year and the other 99% will still survive, that is, $P(C_0) = 99\%$ and $P(C_1) = 1\%$. From above information, we can write the determinist form of the target cost function as

$$C(p) = 6.26\% \times 99\% \times err_0(p) + 52.1\% \times 1\% \times err_1(p).$$

Performance Evaluation

First, we minimize the cost function on the validation datasets to determine the optimal cutoff probabilities of each algorithm, and then compared the minimized costs of each algorithm. The results of optimal cutoff probabilities and minimized misclassification costs are shown in Table 2. Details of cost functions are given in Appendix C. The final model we choose is the logistic regression with cutoff probability 0.65 since it has the smallest cost of misclassification.

Table 2: Optimal Cutoff Probabilities and Costs of Algorithms

Algorithm	Cutoff Prob.	Cost
K-nearest Neighbors	0.35	0.49%
Ads-boosting Tree	0.6	0.40%
Logistic Regression	0.65	0.39%

Model Deployment

Here we give a simple example of deploying this model on the validation dataset. We use the $130 - 30$ trading rules. This method suggest that we can buy 130% of the undervalued stocks and short 30% of the overvalued stocks. If we consider potentially-non-delisted as undervalued and potentially-delisted as overvalued companies, we will long 130% of potentially-non-delisted stocks and short 30% of potentially-delisted stocks. The expected capital gain on the portfolio is

$$Expected\ Capital\ Gain = 6.26\% \times 130\% + 52.1\% \times 30\% = 15.71\%,$$

and the misclassification cost of the logistic regression model with cutoff probability 0.65 is

$$Misclassification\ Cost = 6.26\% \times 30\% \times 67.64\% + 52.1\% \times 30\% = 15.71\% \times 1.67\%,$$

where 67.64% is the error rate of companies with `delisted` $= 1$ and 0.60% is the error rate of companies with `delisted` $= 0$.

Here we still need to consider the taxes and transaction costs, and so we get the total one-year expected return on the portfolio is $15.71\% - 1.67\% - 0.185\% = 13.855\%$, where the market return is 11.655% (have considered taxes and transaction costs) in 2012. This method beats the market.

Note that this strategy has two major concerns. First, it requires a very diversified portfolio containing every stocks so that it can reach the expected return. Second, in this model we use a historical estimation on most important parameters, but it might be not robust throughout time. If we meet a financial crisis, this model will cause a lot of losses!

5 Recommendations

By using the ratios derived from financial report, we are able to predict whether a company would be delisted from the stock market in Taiwan or not. With managing a "130-30" strategy portfolio, we could outperform the market in the end. However, it is hard to predict the delisting from a longer period of time. The company could have window-dressed their financial report in the previous years or solve their financial problem after our prediction.

In terms of portfolio management, managers could use this model to find some high-risk company for their investment interest. However, the opportunity of short selling is not unlimited. Sometimes, there would not have enough stocks for short selling even if our have correctly predicted the delisting. In addition, this model could only be used in a normal year. In some financial crisis like 2008 world financial crisis, we could not guarantee the stableness of this model. Investors should still be aware of major economic situations that could cause the model failed.

A major concern of this model is the investor's psychology. If the prediction is widely spread, most investors would tend to short the predicted delisted companies, even if it is an error prediction. In this situation, even a healthy company could face a financial problem and finally become delisted. The "self-filling" phenomenon could increase our prediction accuracy, but it is not our original goal.

Appendix A Full Names of Variables

Varialble	Full Name	Varialble	Full Name
EPS	Earning per Share	CR	Current Ratio
ROE	Return on Equity	NWGR	Net Wealth Growth Rate
GPM	Gross Profit Margin	TAGR	Total Asset Growth Rate
PM	Profit Margin	PMGR	Profit Margin Growth Rate
Exp/Rev	Expense/Revenue	GPMGR	Gross Profit Margin Growth Rate
Exp Ratio	Expense Ratio	RGDP GR	Real GDP Growth Rate
Tax Rate	Tax Rate	WPI GR	WPI Growth Rate
ROOA	Return on Operating Asset	CPI GR	CPI Growth Rate
D/A	Debt-to-Asset Ratio	TB IR	Treaury Bill Interest Rate
D/E	Debt-to-Equity Ratio	SR	Short-term Interest Rate

Appendix B Confusion Matrices of Models with Economic Variables

Logistic Regression

Training Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	55	0
0	496	0

Validation Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	31	3
0	1	332

2013 Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	4	3
0	803	39

K-nearest Neighbors

Training Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	13	42
0	6	490

Validation Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	7	27
0	3	330

2013 Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	1	6
0	3	839

Ada-boosting Tree

Training Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	55	0
0	0	496

Validation Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	32	2
0	1	332

2013 Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	2	5
0	831	11

Appendix C Performances of Three Models

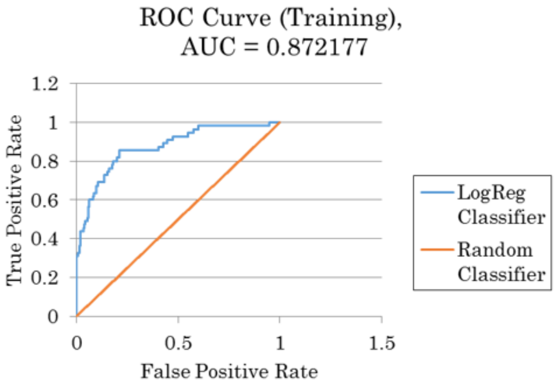
Confusion Matrix (Logistic Regression)

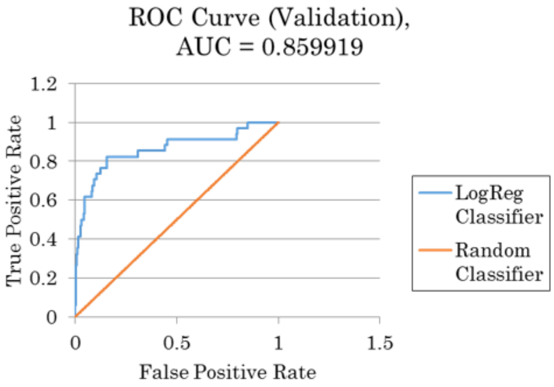
Training Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	18	37
0	7	489

Validation Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	12	22
0	4	329

2013 Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	4	3
0	21	821

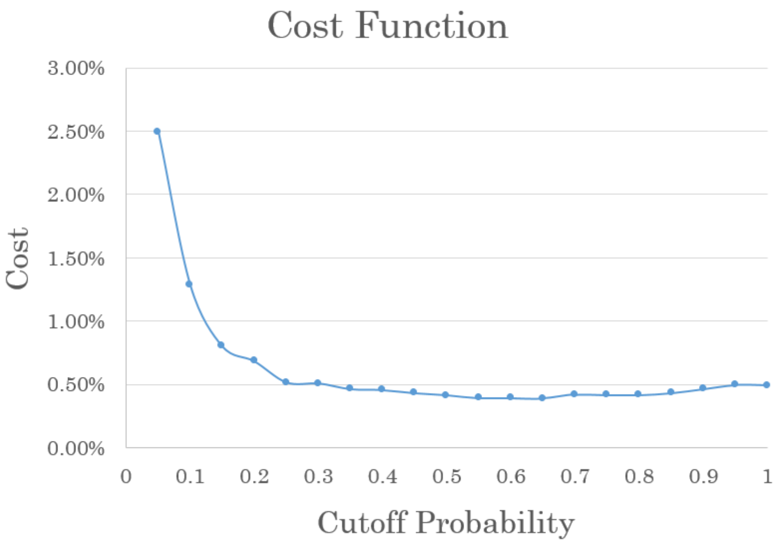
ROC Curve (Logistic Regression)





Optimal Cutoff Probability and Minimized Cost (Logistic Regression)

Cutoff	Cost
0.05	2.50%
0.10	1.28%
0.15	0.80%
0.20	0.69%
0.25	0.52%
0.30	0.51%
0.35	0.46%
0.40	0.46%
0.45	0.43%
0.50	0.41%
0.55	0.39%
0.60	0.39%
0.65	0.39%
0.70	0.42%
0.75	0.42%
0.80	0.42%
0.85	0.43%
0.90	0.46%
0.95	0.49%
1.00	0.49%



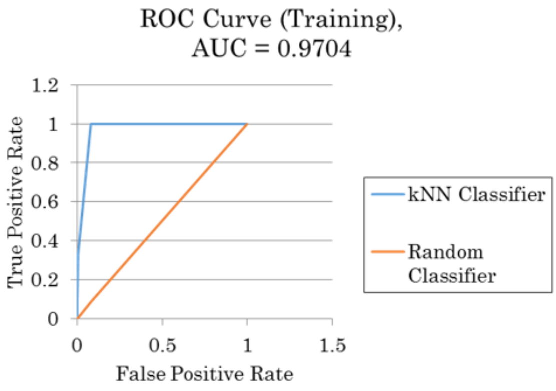
Confusion Matrix (K-nearest Neighbors)

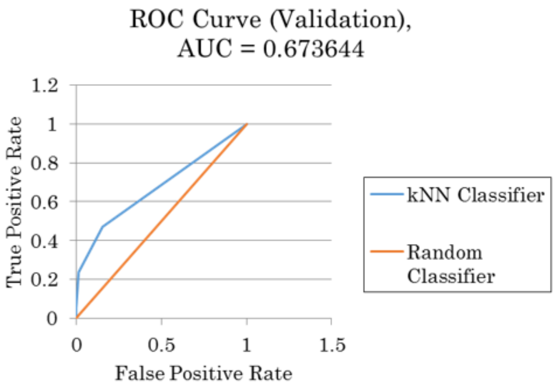
Training Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	18	37
0	2	494

Validation Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	8	26
0	5	328

2013 Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	1	6
0	3	839

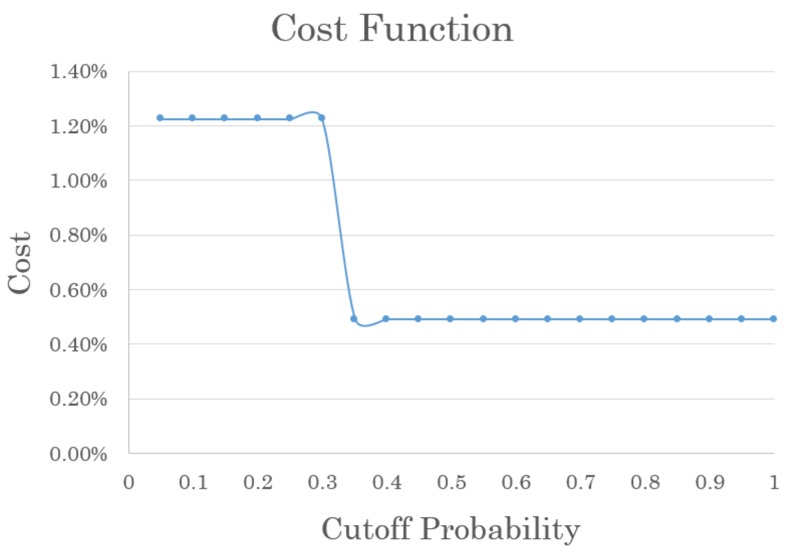
ROC Curve (K-nearest Neighbors)





Optimal Cutoff Probability and Minimized Cost (K-nearest Neighbors)

Cutoff	Cost
0.05	1.22%
0.10	1.22%
0.15	1.22%
0.20	1.22%
0.25	1.22%
0.30	1.22%
0.35	0.49%
0.40	0.49%
0.45	0.49%
0.50	0.49%
0.55	0.49%
0.60	0.49%
0.65	0.49%
0.70	0.49%
0.75	0.49%
0.80	0.49%
0.85	0.49%
0.90	0.49%
0.95	0.49%
1.00	0.49%



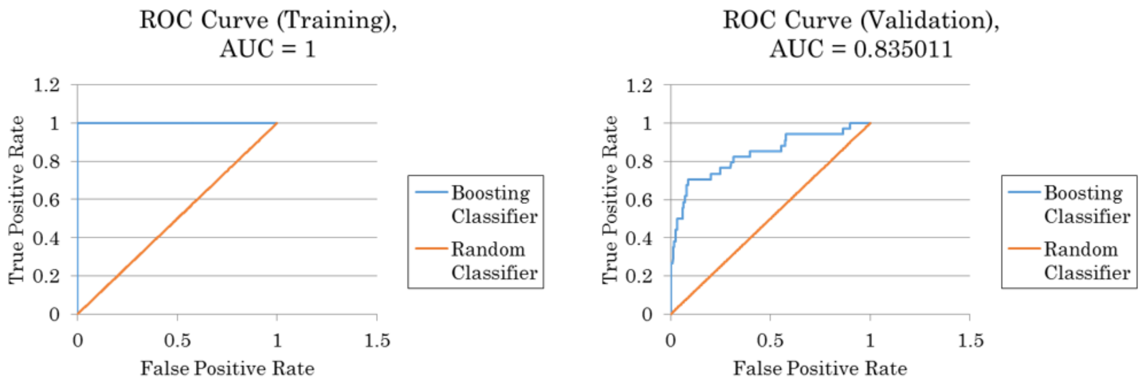
Confusion Matrix (Ada-boosting Tree)

Training Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	55	0
0	0	496

Validation Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	13	21
0	8	325

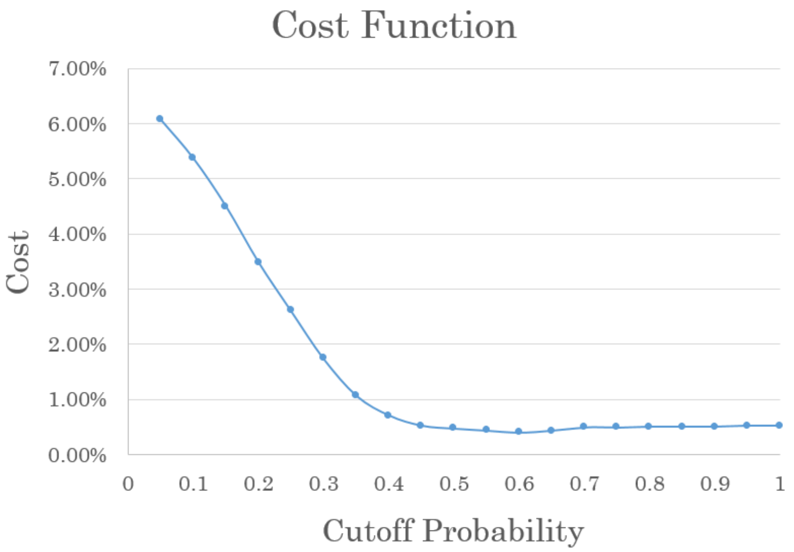
2013 Data (Cutoff = 0.5)		
	Predicted Class	
Actual Class	1	0
1	2	5
0	27	815

ROC Curve (Ada-boosting Tree)



Optimal Cutoff Probability and Minimized Cost (Ada-boosting Tree)

Cutoff	Cost
0.05	6.07%
0.10	5.37%
0.15	4.50%
0.20	3.49%
0.25	2.61%
0.30	1.74%
0.35	1.07%
0.40	0.70%
0.45	0.52%
0.50	0.47%
0.55	0.44%
0.60	0.40%
0.65	0.43%
0.70	0.49%
0.75	0.49%
0.80	0.51%
0.85	0.51%
0.90	0.51%
0.95	0.52%
1.00	0.52%


Cost Function
Cost
Cutoff Probability