CS-BIGS 2(2): 147-156
http://www.bentley.edu/csbigs/vol2-2/wang.pdf

An Analysis of Profit and Customer Satisfaction in Consumer Finance

Chamont Wang
College of New Jersey, USA

Mikhail Zhuravlev
College of New Jersey, USA

This study investigates a set of bank data from a cost/profit perspective. In this specific case, the bank would like to know whether prospective consumers will pay back their credit. The common practice is to calculate the probability that a consumer with certain covariates is to be considered as a potential risk. Our study will go further in the use of the probabilities to maximize the profit. In addition, we will show that the same technique can be used to improve customer satisfaction. The technique should be equally applicable in other consumer markets including auto loans, credit cards, mail catalog orders, home mortgages, and a variety of personal loan products provided by insurance firms, mobile phone companies, and other lending institutions. This case is accessible to readers with an intermediate level of statistics.

1. Introduction

In consumer markets, statistical models are frequently used for two different purposes: (1) to predict the probability that a customer will pay back credit, and (2) to predict the probability that a customer is satisfied with a specific service or should be targeted with a customer-retention program.

Typing "Credit Risk Management" into Google will provide one with more than 13 million examples from the first category. For example, The CreditQuest website, which is found at (http://www.creditquest.com/harland-commercial-lending-resource-library/harland-basel-ii-articles-whitepapers/articles/basel-ii-PD-LGD-EAD-CreditQuest?gclid=___CO_WsJf2tJcCFQslHgod-VLgig), advertises commercial services in the calculation of the probability that a specific customer will default within the next 12 months. Similar calculations, if done properly (a questionable proposition to the 13 million links), may help protect financial institutions from unwanted setbacks.

For an example in the second category, one can see an article titled "Predicting Dissatisfied Credit Card Customers" (Arens and Wegman, 2001). The paper presented a technique to maximize the profit of a customer-retention program. Much of our current research was indeed inspired by the Arens-Wegman paper (http://www.galaxy.gmu.edu/stats/syllabi/inft979/ArensPaper.pdf).

In this study, we will use a set of bank data that was intended for building statistical models to predict the default probability of a customer. In the process, we will impose a cost structure and use a technique to maximize the profit in the decision-making framework. In addition, we will discuss how to use this technique to improve

customer relationship with a concept of *zero revenue* or *minimum profit*. In short, it is a happy instance that it is possible to kill two birds with one stone.

Specifically, we assume that a correct decision of the bank would result in 35% profit at the end of a pre-determined period, say 3-5 years. Here a *correct decision* means that the bank predicts that a customer's credit is in good standing (and hence would grant the loan), and that the customer indeed has good credit. On the other hand, if the model or the manager makes a false prediction that the customer's credit is in good standing, yet the opposite is true, then the bank will result in a unit loss. We summarize the above discussions in the first column of the profit matrix in Table 1.

Table 1. Profit Matrix

	Good Customer (predicted)	Bad Customer (predicted)
Good Customer (observed)	+0.35	0
Bad Customer (observed)	-1.00	0

In the second column of the matrix, the bank predicted that the customer's credit is not in good standing and hence declined the loan. In this situation, there would be no gain or loss in the decision.

Note that the data used in this paper contain 1,000 customers, of which 70% are credit-worthy (good) customers and 30% not-credit-worthy (bad) customers. A manager without any model, who gives every customer a loan would generate the following negative profit per customer:

(700*0.35- 300*1.00)/1000 = -55/1000
= -0.055 unit loss.

This number (-0.055 unit loss) may seem small. But if the average of the loan is $10,000 for this population (n = 1000), then the total loss will be

(-0.055 unit loss)*($10,000 per unit per customer)*
(1,000 customers) = -$550,000,

a whopping five hundred and fifty thousand dollar loss. On the other hand, if a model produced the classification matrix in Table 2, the total profit would be

608*$10,000*0.35 – 192*$10,000
= $208,000

The difference of model vs. no-model is

$208,000–(-$550,000) = $758,000,

about seven hundred and fifty eight thousand dollars of profit. The main goal of this study is to build statistical models to maximize the profit. In the process, we will discuss how to modify the technique to improve customer satisfaction in the zero-revenue framework.

Table 2. Classification Matrix

	Good (predicted)	Bad (predicted)	Row total
Good (observed)	608 customers (76%)	46 customers	700 customers
Bad (observed)	192 customers (24%)	154 customers	300 customers
Column total & percentages	800 customers (100%)	200 customers	1,000 customers

2. Modelling Strategy

Assume that the data are already in order without missing values. Then the following steps may help maximize the profit (and with a small twist to maximize the customer satisfaction). The implementation of the strategy would involve the following steps:

1. In the field of data mining and predictive modeling, a variety of tools are available that may yield different results in terms of different profits and different levels of customer satisfaction. The tools include Decision Tree, Regression, Neural Networks Stochastic Gradient Boosting, Support Vector Machines, Ensemble models and countless variations of these techniques. In this study, we will use SAS Enterprise Miner 5.3 for model building and profit calculation.

2. Variable Selection: The original data set has 20 predictors. In many studies, the number of variables in the study may be hundreds or thousands (or millions in Google data mining; Joseloff and Pozderec, 2005). Some of the predictors may not be as important as others and the exclusion of these variables may improve the model performance in a very significant manner.

3. Bundling the Variables: Some of the predictors may be redundant or correlated to each other. In the statistical literature, a variety of techniques are available to lump these predictors together. In SAS-EM, two different techniques (Variable Clustering and Principle Components) often improve the model performance.

4. Binning, Filtering, and Variable Transformation: Binning is a technique to group variable values into classes that can be used as inputs for subsequent

model building. Sometimes certain observations are corrupted and should be filtered out to improve model performance. Furthermore, transformation of variables may improve the fit of the model to the data.

5. Isolated Events and Cluster Structures: The K-Nearest Neighbor algorithm usually cannot compete with models such as Neural Networks, Regression, or Decision Trees. But if there are isolated events in the data, then the algorithm may be the best to try. On the other hand, if there are cluster structures in the data, then RBF (radial basis function) Neural Networks may be the best.

6. Mega Models: Sometimes it is desirable to build a chain of models such as Consolidation Tree + Selection Tree + Consolidation Neural Network (or Consolidation Regression). This kind of mega model takes skill to build but sometimes the payoff can be rather rewarding.

7. Parameter Tuning: Given a specific data set, almost all data mining tools can be honed for better performance. A few years from now, someone may come up with a meta-algorithm in a super computing machine to incorporate all the strengths of the data mining models for the best outcomes under various criteria. But before that new era, parameter tuning of the models may be needed to improve results.

8. Change Nominal Predictors to Ordinal Variables: In this German credit data, many input variables are *ordinal* in nature but are coded on *nominal* scales. With this kind of coding, Neural Network and Gradient Boosting may fail to run. Our remedy is to convert nominal predictors into ordinal variables and then treat the predictors as interval variables for higher profit.

9. Different Cutoff Values: Given the study population, the model will produce the probabilities of all customers with regard to their credit standing. If the probability of a specific customer is above the cutoff (a.k.a., threshold), then the customer will be placed in the category of good customers; otherwise the customer loan application will be denied. By the adjustment of different cutoff values, we may be able to increase the total profit. In our experience, this technique is one of the most important in the maximization of the profit.

10. Marginal Effects and Decision Rules of Complicated Models: Machine learning techniques such as Neural Network, Support Vector Machine, and Gradient Boosting are often criticized for being black-box model in which it is "impossible to figure out how an individual input is affecting the predicted outcome" (Ayres, 2007, p. 143). This was true in the old days. But with modern computing power, given any Neural Network, one can plot its response surface and calculate the marginal effects (Wang and Liu, 2008). For Boosted Trees, one can also calculate Interaction Effects (Friedman and Popescu, 2005) and draw Partial Dependence Plots for the understanding and the interpretation of the model (Friedman, 2002). Furthermore, one can build a Decision Tree after a Neural Network (or other complicated model) to extract decision rules that can be very helpful for managers or other decision makers in real world applications.

In this study, we will focus on Steps 1, 8 and 9 to illustrate the key components of this technology. Other steps sometimes help and can be used as homework assignments in a data mining class. For budding data miners, it is a thrill when they discover that a specific technique indeed boost the model performance. It is worth trying.

3. Data

The dataset used in this study comes from the Department of Statistics, University of Munich and is available at http://www.stat.uni-muenchen.de/service/datenarchiv/kredit/kredit_e.html. The data contains 1000 cases, representing past borrowers, and 20 variables, representing different attributes of those borrowers. The attributes include financial, personal, and demographic information.

Each case also has a binary variable (Credibility) indicating whether the borrower was a good or poor customer. Credibility is either "Good" or "Bad," with 70% of the cases falling into the former category, and the other 30% into the latter. This value is determined by a variety of factors, including timeliness of repayment, amount over-drafted, and account turnover. The variables can be grouped as in Table 3.

Of the 20 variables, only Duration in Months, Amount of Credit in DM (Deutsche Mark), and Age are coded as interval variables. The other 17 are nominal. However, Balance of Current Account, Value of Savings or Stock, Duration of Current Employment, Installments in % of Available Income, Duration in Current Residence, and Number of Previous Credits at This Bank all lend

themselves well to being assigned numerical values corresponding to the different nominal values.

Table 3. List of the Predictors

Category	Variables
1. Personal	Marital Status, Sex, Age, Number of Dependants
2. Assets	Balance of Current Account, Amount of Credit in DM (Deutsche Mark), Value of Savings or Stock, Most Valuable Available Asset
3. Repayment History	Duration in Months, Payment of Previous Credit, Number of Previous Credits at This Bank
4. Leverage	Purpose of Credit, Installments in % of Available Income, Further Debtors/Guarantors, Further Running Credits
5. Employment	Duration of Current Employment, Occupation, Foreign Worker
6. Household	Duration in Current Residence, Type of Apartment, Telephone

This study seeks to use these variables to discriminate between borrowers with Good Credibility and Bad Credibility. A model that could successfully discriminate between Good and Bad borrowers would help the bank in deciding to whom they can extend credit.

3. Methods

Part-A. In this section we will first present the results using the original data and the following models: Decision Tree, Dmine Regression, and Support Vector Machine. Here Dmine Regression computes a forward stepwise least-squares regression including two-way interactions, binning, and group variables. This tool is similar to the technique used in Foster and Stine (2004). The tool often produces superior results, but sometimes the model is sensitive to outliers and sometimes over-fits the training data and hence needs caution in deployment.

Robust methods against outliers include Decision Trees and Support Vector Machines (SVM). Note that SVM is one of the most important tools in the machine learning community and indeed has many success stories in its applications. For example, in 2008, SVM scored the winning results of the KDD (Knowledge Discovery in Databases) Cup Data Mining competition (http://www.stern.nyu.edu/ioms/Perlich_final-cup-kdd08.pdf, http://www.kddcup2008.com/).

For other examples, Bastos and Wolfinger (2004) reported a 4% error rate by using SVM, as compared to a 27% error rate on the same set of data in a 2002 paper in *The New England Journal of Medicine*. Yu (2005) used SVM and achieved a 2% error rate in a case study on cloud detection, as compared to a 53% error rate by expert labels. Adnan and Bastos (2005) reported substantial advantages of SVM over regression and neural networks. Furthermore, Adnan used SVM to achieve a stunning 99.6% accuracy in the 2004 UCSD data mining competition.

Our SAS-EM modeling process is displayed in Figure 1:

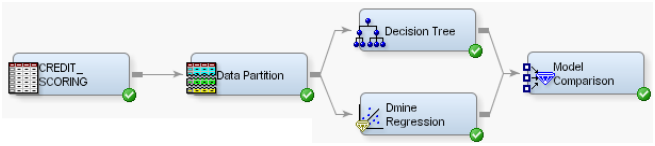

Figure 1. SAS-EM process flow

In the process flow in Figure 1, we use the 40%-30%-30% split for the partition of the original data into Training, Validation, and Test (Hold-out) data sets. We then build and compare the three models (Regression, Tree, and SVM) on their profits. The entire process is self-explanatory and indeed rather straightforward -- with this exception: in the first Data Source node, we need to enter *weight values* for the decision (Figure 2):

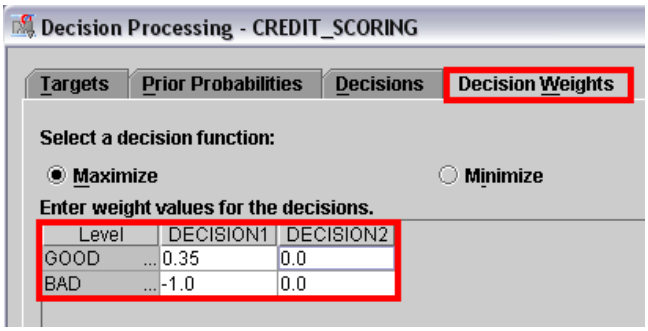

Figure 2. Decision matrix

This process yields the Cumulative Mean Profits of Regression, Dmine Regression, and Decision Tree, displayed in Figures 3, 4 and 5 respectively. The charts indicate that Dmine Regression is consistently better than Regression and that at a 30%-cutoff, Decision Tree is the least desirable model. For Regression, the chart says that if the bank manager uses the model to grant the loans to the top 30% of the customers, then within this group, the average profit would be 0.263 units.

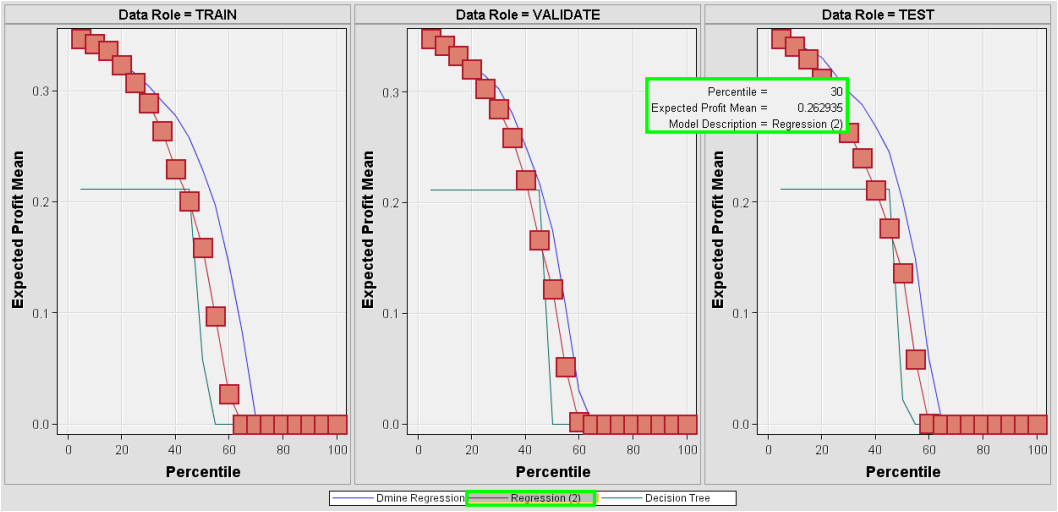

Figure 3. Cumulative mean profits of Regression Model

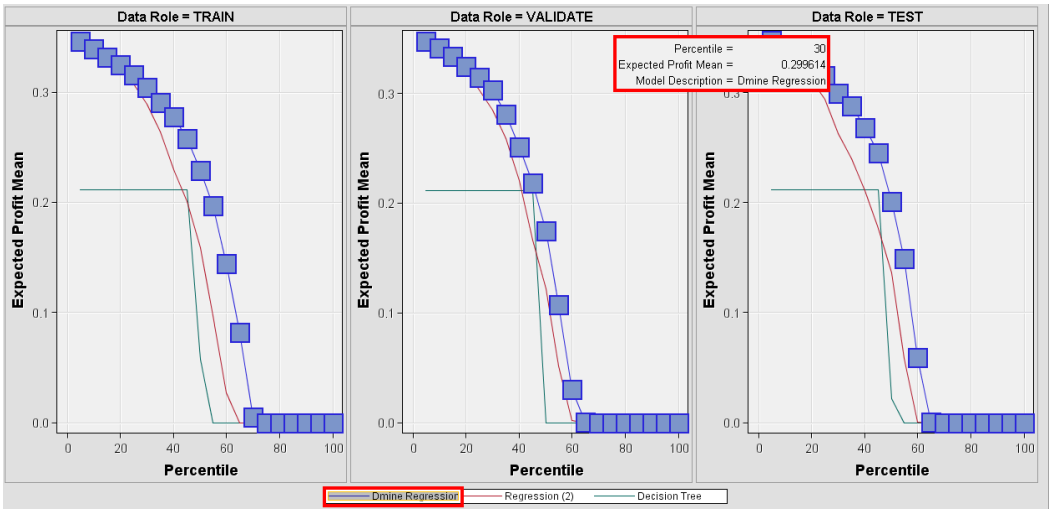

Figure 4. Cumulative mean profits of Dmine Regression

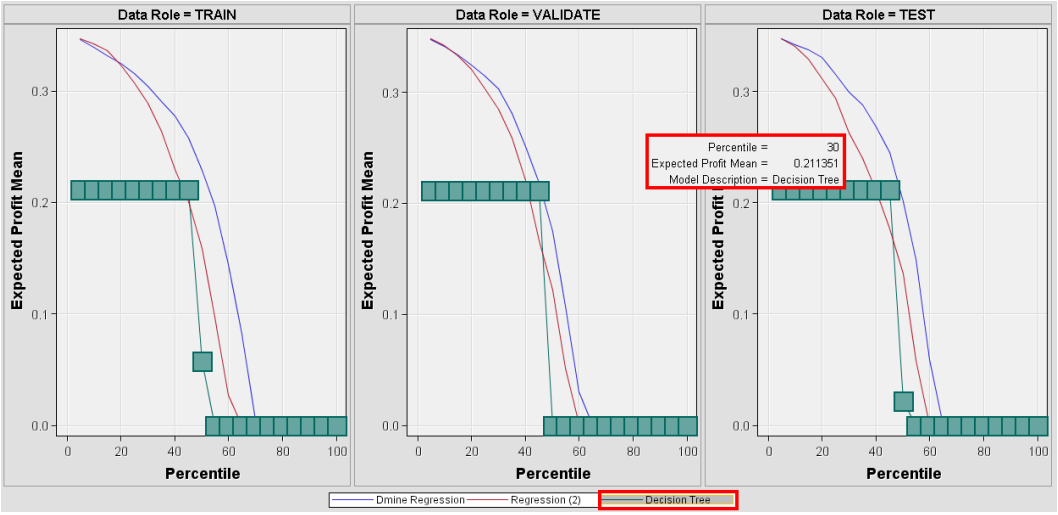

Figure 5. Cumulative mean profits of Decision Tree

Table 4. Mean Profit and Total Profit

Maximum		25%	30%	35%	40%	45%	50%	55%
Mean	Reg	0.29423	0.26294	0.24011	0.21072	0.17647	0.13597	0.05850
Profit	Dm_Reg	0.31544	0.29962	0.28761	0.26818	0.24500	0.20178	0.14912
	Tree	0.21135	0.21135	0.21135	0.21135	0.21135	0.021767	0
Total	Reg	735,575	788,820	840,385	$842,880	794,115	679,850	321,750
Profit	Dm_Reg	788,600	898,860	1,006,635	1,072,720	$1,102,500	1,008,900	820,160
	Tree	528,375	634,050	739,725	845,400	$951,075	108,835	0

Assume that there are 1,000 customers with an average loan application of $10,000, then the total profit would be 1000*0.3*0.26294*$10,000 = $788,820.

In contrast, the mean profit of the Decision tree at the 30th-percentile is 0.211351, which is equivalent to a total profit of 1000*0.3*0.211351*$10,000 = $634,053, a lot less than that of Regression. However, there are surprises when the story unfolds. Table 4 summarizes the Cumulative Profit for the above three models (the maximum profit for each model is high-lighed in red).

Discussion:

(a) Examining Table 4, it follows that the Total Profit

=(number of customers)*(% of selected customers)*(mean profit)*(average loan amount)
=1,000*(% of selected customers)*(mean profit)*$10,000.

So if the concern of the bank manager is maximum profit, then he or she should use Dmine Regression to select the top 45% of the customers.

(b) Note that the original population has 70% of customers with good credit standing. But Dmine Regression would select only 45% of customers. So it is questionable whether the use of the Dmine Regression is a good business practice. From a customer-relationship view point, a cutoff at 60% or 70% may be more desirable. As a matter of fact, in certain industries (airlines, for example), a higher cutoff with zero profit may be preferable in terms of customer retention. In the next Section, we will try other models to see whether a high profit with high cutoff is possible.

Part-B. Neural Networks and recoding of the data.

Categorical predictors often pose problem for parametric models such as neural network, where each categorical level must be coded by an indicator variable and the result would be a huge amount of parameters beyond the capacity of the model. Sometimes a technique of using a Consolidation Tree may be able to group categorical input levels and create new, useful predictors for neural network in the later part of the process flow. On the other hand, if the categorical variables are intrinsically *ordinal*, then a re-coding of the inputs may improve the profit. In this case study, the following predictors are categorical (see Table 5).

Table 5. Nominal Predictors with Multiple Categories

Predictor	Categories
Purpose of credit	11 levels
Amount of credit	10 levels
Duration in months	10 levels
Age in years	5 levels
Payment of previous credits	5 levels
Value of savings or stocks	5 levels
Has been employed by current employer	5 levels
Installment in % of available income	4 levels
Living in current household	4 levels
Most valuable available assets	4 levels
Balance of current account	4 levels
Number of previous credits at this bank	4 levels
Occupation	4 levels
Further running credits	3 levels

For example, one of the inputs has five different categories and is called "Payment of previous credits." In our judgment, it may be better to code the five different categories in either of the following manners (see Table 6).

Table 6. Predictor: "Payment of previous credits"

Categorical level	Ordinal level (option 1)	Ordinal level (option 2)
no previous credits / paid back all previous credits	3	1
paid back previous credits at this bank	5	2
no problems with current credits at this bank	4	2
hesitant payment of previous credits	1	0
problematic running account / there are further credits running but at other banks	2	0

In Option 1 in Table 6, the five categories are coded with five different numbers, while in Option 2, certain categories are lumped together. In the subsequent analysis, we will use Option 1, while readers are urged to use their own judgment to code the variables in sensible ways to achieve higher profit. The re-coding of the 14 predictors in the previous Table takes effort and

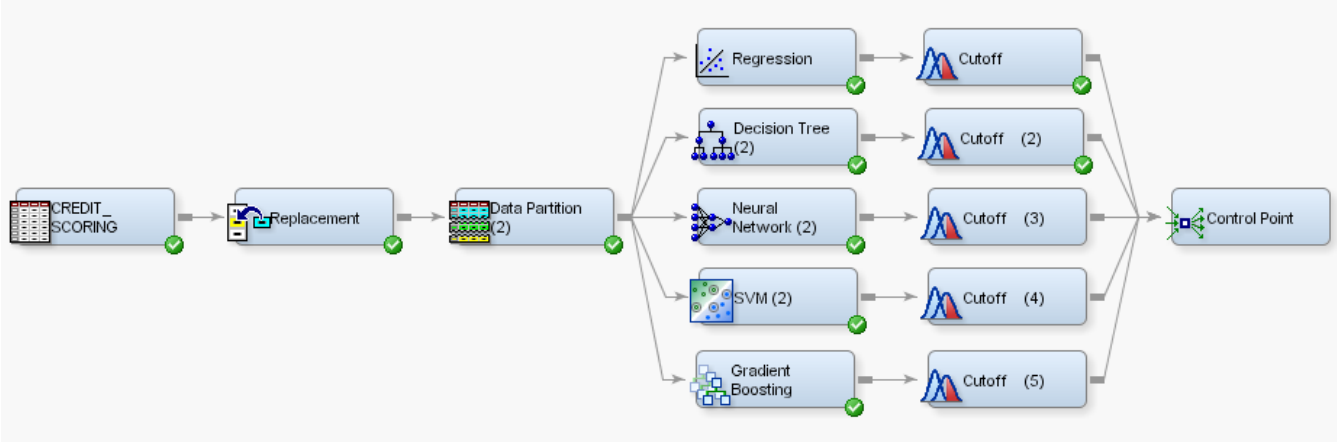

Figure 6. SAS-EM process flow

was accomplished by a SAS-EM **Replacement node**; details are included in the Appendix. After all this hard work, both the Neural Network and the Gradient Boosting will run smoothly. The process flow is as in Figure 6.

In the process flow in Figure 6, the Cutoff node uses resolution at 1%-increment (as compared to the 5%-increment in the Model Comparison node). In the following model comparison, we will skip Dmine Regression for some technical reasons. In one case, we used a min-max transformation and a mixture of ordinal and nominal scales and produced the following results (the coding method is included in the Appendix).

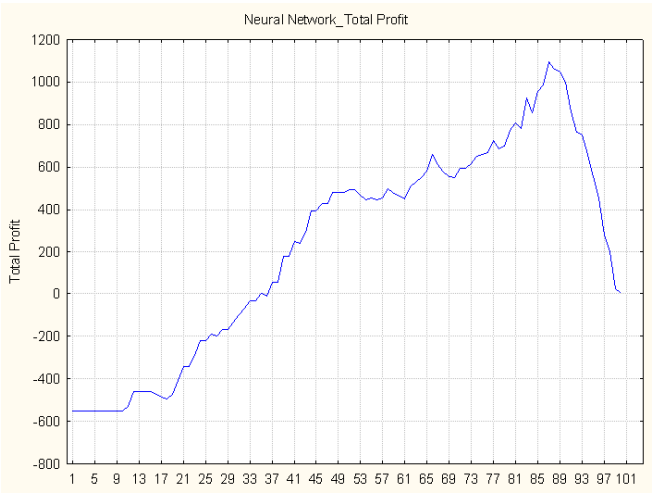

Figure 7. Neural Network: Total Profit vs. cutoff probability

The graph shows that for the Neural Network model, the Total Profit reaches its maximum at a cutoff value near 86%. Similar calculations are summarized in Table 7. Table 7 indicates that the Neural Network model produced the best profit, while SVM selected the highest number of customers. In another case, we abandoned the min-max transformation but converted almost all nominal variables into ordinal scale. Some results are given in Table 8 (the coding method is included in the Appendix).

Table 7. A Comparison of Three Models with Min-Max Transformation

Maximum	Total Profit	Threshold Probability	% of Selected Customers
NN	$1,096,667	0.86	0.40
Gini Tree	$488,333	0.89	0.49
SVM	$921,667	0.55	0.62

Table 8. A Comparison of Two Models without Min-Max Transformation

Maximum	Total Profit	Threshold Probability	% of Selected Customers
Boosting	$883,333.33	0.84	0.41
NN	$935,000.00	0.76	0.64

In both cases, Neural Networks appeared to outperform other models. However, it is important to remember that the models will be applied to future data, subject to fluctuation. Consequently it may not be beneficial to blindly chase the model that produces the best profit. Consider the situation in Table 9.

Table 9. Are the Differences Statistically Different?

Maximum	Total Profit	Threshold Probability	% of Selected Customers
NN-1	$1,096,667	0.86	0.40
NN-2	$935,000.00	0.76	0.64
Statistically Significant?	No		Yes

We conclude from Table 9 that the difference in the first column may be due to chance fluctuation, while that in

the third column is not. NN-2 should then be the best model. We explore this issue in the next section.

Part-C. Investigation of the variability of Total Profit and the % of Selected Customers.

It is a common belief that Neural Networks and other complicated models cannot generate confidence interval of the estimated value (see, e.g., Ayres, 2007, p. 143-144). This was true in the old days. Recall that we used the 40%-30%-30% split for the partition of the original data into Training, Validation, and Test (hold-out) data sets. The process required a random number generator with a specific seed. A change of the seed would change the holdout data and give different values of the Total Profit and the % of Selected Customers. By repeating this process, we should be able to estimate the variations of these two quantities (Total Profit and the % of Selected Customers). Table 10 shows the Total Profits of five different models in twelve runs; here SE = Standard Error which is SD/sqrt(K), K=12.

Table 10. Total Profits in 12 Runs for Each of Four Models

Run	NN-1	NN-2	Regression	SVM
1	$1,006,622.52	$966,887.42	$823,333.33	$1,033,112.58
2	$945,182.72	$885,382.06	$903,973.51	$1,008,305.65
3	$1,197,019.87	$1,119,205.30	$980,066.45	$1,089,403.97
4	$852,649.01	$746,688.74	$783,333.33	$854,304.64
5	$745,847.18	$684,385.38	$733,443.71	$674,418.60
6	$1,053,333.33	$833,333.33	$825,581.40	$983,333.33
7	$1,108,333.33	$490,033.22	$831,125.83	$759,136.21
8	$850,993.38	$780,000.00	$704,318.94	$915,000.00
9	$890,365.45	$729,235.88	$935,430.46	$875,415.28
10	$778,145.70	$923,841.06	$764,900.66	$923,841.06
11	$923,588.04	$1,029,900.33	$887,417.22	$865,448.50
12	$870,860.93	$889,072.85	$978,333.33	$789,735.10
Mean	$935,245.12	$839,830.46	$845,938.18	$897,621.24
Ranking	1	4	3	2
SD	$134,296.67	$168,665.24	$91,928.13	$120,516.72
SE	$38,768.11	$48,689.46	$26,537.37	$34,790.18
t-test	t = 1.90 (NN-1 vs. Regression), df = 19.45, one-tailed P = .036			
F-test	F-ratio = 2.13, one-tailed P = .11, not significant			

Table 10 indicates that SVM is a strong model, but NN-1 is the best. The statistical tests focus on NN-1 and the standard logistic regression. Our feeling is that NN-1 would have more variability in successive runs than Regression, but the F-test fails to register the difference (probably due to the small sample size of 12). A t-test with Satterthwaite approximation (assuming different variances) has degrees of freedom of 19.45 and a one-sided P value of .036. A standard t-test (assuming equal variances) has degrees of freedom of 22 and a one-sided P value of .035. Figure 8 displays the Total Profits in 12 runs for NN-1 and Regression models.

In other words, if we imagine a big bank with 12 branches, then NN-1 would outperform Regression by ($935,245.12 - $845,938.18)*12 = $1,071,683.28, more than one million dollars for the bank. In Table 11, we compare the percent of Selected Customers for various models. The objective is to find a model that is high in customer satisfaction. For each model at each run, the number is selected at cutoff that produced the highest profit. In certain applications (e.g., airline industry) where long-term customer satisfaction is more important than short-term profit, then the cutoff should be selected when the profit is zero or slightly above zero. The technique is the same (but the binary target must contain satisfaction and dissatisfaction).

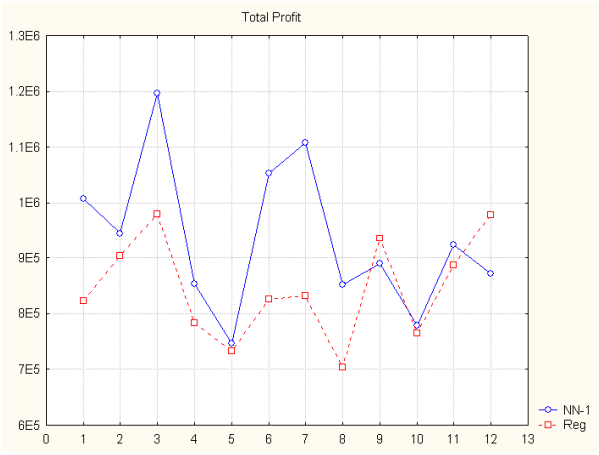

Figure 8. Total Profit: NN-1 model vs. Regression in 12 runs

Table 11. Percent of Selected Customers for Four Models

Run	NN-1	NN-2	Reg	SVM
1	45.36%	40.40%	55.67%	48.68%
2	38.54%	53.49%	34.77%	53.16%
3	63.58%	63.91%	48.51%	49.01%
4	65.23%	43.05%	50.67%	39.74%
5	40.53%	54.15%	50.33%	64.12%
6	50.67%	45.67%	59.47%	48.67%
7	40.67%	69.10%	48.01%	42.19%
8	37.09%	57.00%	52.16%	57.00%
9	48.50%	45.18%	59.93%	45.51%
10	41.39%	50.66%	57.62%	50.66%
11	39.20%	43.52%	57.29%	46.51%
12	44.04%	49.67%	53.67%	59.60%
Mean	46.23%	51.32%	52.34%	50.40%
Ranking	4	2	1	3
SD	9.40%	8.74%	6.89%	7.08%
SE	2.71%	2.52%	1.99%	2.05%
t-test	t = -1.82 (NN-1 vs. Reg), df = 20.18, one-tailed P = .042			
F-test	F-ratio = 1.86, one-tailed P = .16, not significant			

The t-test of the equality of means in Table 11 shows that Regression may be better in terms of picking up more customers and hence may be more profitable than NN-1 in the long term. However, the actual difference (52.34%-46.23%) is not substantial, so NN-1 would be the best choice in terms of its superior Total Profit. Finally, in theory, SVM may be the most robust, but more experiments need to be done.

4. Concluding Remarks

This study investigates a set of bank data from a cost/profit perspective. Our analysis indicates the following: (1) the re-coding of nominal variables in ordinal scale improves the Total Profit, (2) linear transformations of interval variables into the range of [0, 1] improve the performances of Neural Networks and Support Vector Machines, and (3) adjustment of threshold probability boosts the Total Profit.

In the comparisons of the models, we focus on the hold-out data to avoid potential problems of model over-fitting on the training and validation datasets. Given a specific model, we observed significant variations of Total Profits when we changed the seeds in the data partition step. To facilitate a fair comparison of different models, we use a re-sampling scheme that is similar to cross-validation and bootstrapping techniques in modern statistics. The comparison indicates that Neural Network model with suitable re-coding and transformation would be the most profitable of the models in this study.

Appendix

1. SAS code for the resampling to evaluate the variability of the Total Profit and the %_Population:

```
data German_NN (Keep = Cutoff FP_CLASSIFS
     TP_CLASSIFS Pred_Pos Pred_Neg
     Total_Profit Percent_Population
     DataRole);
 Set sasuser.German_NN_Cutoff;

  Total_Profit=(1000/(Pred_Pos+Pred_Neg))*
  (10000)*(.35*TP_CLASSIFS-1*FP_CLASSIFS);
 Percent_Population=
Pred_Pos/(Pred_Neg+Pred_Pos);
 If DataRole ne "TEST" then delete;
 Run;

Proc sort;
 by descending Total_Profit;
Run;
```

2. The data re-coding method for NN-1 model:

(a) Interval Variables that were transformed onto a range between 0 and 1:
DURATION_IN_MONTHS
AMOUNT_OF_CREDIT_IN_DM
AGE

(b) Binary Variables:
CREDITABILITY
FOREIGN_WORKER
TELEPHONE
NUMBER_OF_PERSONS_ENTITLED_TO_MAINTENANCE

(c) Changing Nominal Variables into Ordinal Scales:

The conversion of Nominal variables to Ordinal scale requires a lot of subject-matter judgment and sometimes can be controversial. Readers of this article are urged to examine the conversion given in the figure in the Appendix and use his/her own numbers when deemed necessary. Different assignments of the numeric values may produce higher profits as a result.

Variable	Level					
Balance_of_current_account	>=200 DM	...	C	...	.	4
Balance_of_current_account	no running account	...	C	...	.	1
Balance_of_current_account	no balance	...	C	...	.	2
Balance_of_current_account	<= 200 DM	...	C	...	.	3
Balance_of_current_account	_UNKNOWN_		C		.	_DEFAULT_
Living_in_current_household_for	greater than 7 years	...	C	...	.	4
Living_in_current_household_for	between 1 and 4 years	...	C	...	.	2
Living_in_current_household_for	between 4 and 7 years	...	C	...	.	3
Living_in_current_household_for	less than 1 year	...	C	l...	.	1
Living_in_current_household_for	_UNKNOWN_	.	C		.	_DEFAULT_
Further_debtors_Guarantors	none	...	C	...	.	1
Further_debtors_Guarantors	guarantor	...	C	...	.	2
Further_debtors_Guarantors	co-applicant	...	C	...	.	3
Further_debtors_Guarantors	_UNKNOWN_	.	C		.	_DEFAULT_
Further_running_credits	no further running credits	...	C	...	.	1
Further_running_credits	at other banks	...	C	...	.	3
Further_running_credits	at department store or mail order house	...	C	...	.	2
Further_running_credits	_UNKNOWN_	.	C		.	_DEFAULT_
Has_been_employed_by_current_emp	between 1 and 4 years	...	C	...	.	3
Has_been_employed_by_current_emp	greater than 7 years	...	C	...	.	5
Has_been_employed_by_current_emp	between 4 and 7 years	...	C	...	.	4
Has_been_employed_by_current_emp	less than 1 year	...	C	l...	.	2
Has_been_employed_by_current_emp	unemployed	...	C	...	.	1
Has_been_employed_by_current_emp	_UNKNOWN_	.	C		.	_DEFAULT_
Installment_in___of_available_in	less than 20	...	C	l...	.	1
Installment_in___of_available_in	between 25 and 35	...	C	...	.	3
Installment_in___of_available_in	between 20 and 25	...	C	...	.	2
Installment_in___of_available_in	greater than 35	...	C	...	.	4
Installment_in___of_available_in	_UNKNOWN_	.	C		.	_DEFAULT_
Number_of_previous_credits_at_th	one	...	C	...	.	1
Number_of_previous_credits_at_th	2 or 3	...	C	...	.	2
Number_of_previous_credits_at_th	4 or 5	...	C	...	.	3
Number_of_previous_credits_at_th	6 or more	6	C	...	.	4
Number_of_previous_credits_at_th	_UNKNOWN_	.	C		.	_DEFAULT_
Occupation	skilled worker/skilled employee/minor civil s...	...	C	...	.	3
Occupation	unskilled with permanant residence	...	C	...	.	2
Occupation	executive/self-employed/higher civil servant	...	C	...	.	4
Occupation	unemployed/unskilled with no permaant res...	...	C	...	.	1
Occupation	_UNKNOWN_	.	C		.	_DEFAULT_
Payment_of_previous_credits	no previous credits or paid back	...	C	...	.	3
Payment_of_previous_credits	paid back previous credits at this bank	...	C	...	.	5
Payment_of_previous_credits	no problems with current credits at this bank	...	C	...	.	4
Payment_of_previous_credits	problematic running accounts	...	C	...	.	1
Payment_of_previous_credits	hesistant payment of previous credits	...	C	...	.	2
Payment_of_previous_credits	_UNKNOWN_	.	C		.	_DEFAULT_
Most_valuable_available_assets	savings contract with a building society/Life ...	...	C	...	.	3
Most_valuable_available_assets	no assets	...	C	...	.	1
Most_valuable_available_assets	car/other	...	C	...	.	2
Most_valuable_available_assets	ownership of house or land	...	C	...	.	4
Most_valuable_available_assets	_UNKNOWN_	.	C		.	_DEFAULT_
Type_of_apartment	rented	...	C	...	.	1
Type_of_apartment	free apartment	...	C	f...	.	2
Type_of_apartment	owner	...	C	...	.	3
Type_of_apartment	_UNKNOWN_	.	C		.	_DEFAULT_
Value_of_savings_or_stocks	no savings	...	C	...	.	1
Value_of_savings_or_stocks	greater than 1000 DM	...	C	...	.	5
Value_of_savings_or_stocks	less than 100 DM	...	C	l...	.	2
Value_of_savings_or_stocks	between 100 and 500 DM	...	C	...	.	3
Value_of_savings_or_stocks	between 500 and 1000 DM	...	C	...	.	4
Value_of_savings_or_stocks	_UNKNOWN_	.	C		.	_DEFAULT_

REFERENCES

Adnan, A. and Bastos, E. (2005). A Comparative Estimation of Machine Learning Methods on QSAR Data Sets. *SUGI-30 Proceedings*.

Ayres, I. (2007), *Super Crunchers: Why Thinking-by-Numbers is the New Way to be Smart*. Bantam Books.

Bastos, E. and Wolfinger, R. (2004). Data Mining in Clinical and Genomics Data, presented at M2004, *the SAS 7th annual Data Mining Technology Conference*. Power Point slides are available upon request.

Foster, D.P. and Stine, R.A. (2004). Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. *Journal of the American Statistical Association* 99, 303-313.

Friedman, J.H. (2002). Greedy Function Approximation: a Gradient Boosting Machine. *Annuals of Statistics* 29, 1189-1232.

Correspondence: wang@tcnj.edu

Friedman, J.H. and Popescu, B.E. (2005). Uncovering Interaction Effects, presented in the *Second International Salford Systems Data Mining Conference*.

Jordaan, E. (2008). Support Vector Machines: the New Kid on the Block, presented at M2008, *the SAS 11th annual Data Mining Conference*. Power Point slides are available upon request.

Joseloff, M and Pozderec, G.P. (2005). T*he Google Boys*, a Biography Channel DVD movie.

Wang, C. and Liu, B. (2008). Data Mining for Large Datasets and Hotspot Detection in an Urban Development Project. *Journal of Data Science*, in print.

Yu, Bin (2005). Mining Earth Science Data for Geophysical Structure: A Case Study in Cloud Detection, presented at the 5th SIAM International Conference on Data Mining. Power Point slides are available upon request.

http://www.siam.org/meetings/sdm05/binyu.htm