Homework Assignment 4
CSE 291 Spring 2011
Maximizing Donation Profits

Jihoon Kim
UCSD Division of Biomedical Informatics
j5kim@ucsd.edu

Hari Damineni
UCSD Computer Enginering & Science
hdaminen@cs.ucsd.edu

André Christoffer Andersen
NTNU Industrial Economics
andrecan@stud.ntnu.no

May 10, 2011

Abstract

In this assignment we use data mining techniques coupled with economic decision theory in order to optimize a postal mail solicitation campaign using the KDD98 contest data-set. Linear regression is used to predict donation amount, while logistic regression is used to predict willingness to donate as a probability. Combining these models yields a expected individual donation amounts. Under a specific cost profile we can thus use this information to discern the economic benefit of initiating solicitations on a per-person basis. The implementation of our proposed strategy contributed to the total profit with an additional $2,750 compared to the naive solicit-everybody strategy. This is a 29% improvement.

1 Data-set Preparation

1.1 About the dataset

The KDD98 dataset is obtained from the The Second International Knowledge Discovery and Data Mining Tools Competition. It is a collection of data on potential donors who previously have been solicited for monetary gifts to a national veterans organization. The goal is to maximize donations for this veterans organizations 1997 blank-cards-with-labels donation campaign, also known as the (97NK) campaign. More succinctly we aim to generate a subset of the people (i.e., instances) in KDD98, who if solicited will maximize donations while taking in to account a \$0.68 cost per solicitation.

We use two datasets. One main dataset, the training dataset, which the models are based on and a final validation dataset that is only used once for fair performance measuring. The training dataset has 95,412 records while the validation dataset has 96,367 records. Both datasets operate with the same 479 features and two labels. Labels are of course left out as features to be learned.

1.2 Pre-processing before CV

The pre-processing phase goes through add, discard, dummify, and split. R is used for these task as programming is needed.

Add. Introduces a binary missingness feature (1 if missing, otherwise 0) for each existing feature that has at least one missing value will make sure that latent information is not discarded.

Discard. Removes instances whose outcome value is missing, because these obviously have no predictive value.

Dummify. Introduces $k-1$ new binary features for each k-valued categorical feature. One is dropped in order to prevent linear dependence.

Split. Replaces multi-modal features into separate categorical features then dummify each categorical features.

1.3 Pre-processing within CV

To avoid information leakage, impute and normalize are performed within each trainingset per CV iteration.

Impute. The missing value are replaced with the column mean for numeric feature and the mode for binary feature.

Normalize. Performs min-max scaling to numeric feature to have the range $[0, 1]$. This ensures SVM model does not get biased toward a feature with extreme values.

1.4 Feature selection

We use a univariate logistic regression to screen the most predictive features. We select features whose P-values are less than 0.05. Then we select features by the descending order of odd-ratio. Based on previous homeworks, we choose 13 features as below:

Numeric. `INCOME`, `NGIFTALL`, `NUMPROM`, `FISTDATE`, `LASTDATE`

Binary. `PEPSTRFL`, `RECP3`, `RFA_2A_E`, `RFA_2A_F`, `RFA_2A_G`, `RFA_2F_2`, `RFA_2F_3`, `RFA_2F_4`

The raw data of `FISTDATE` and `LASTDATE` had the format "YYMM". But they are converted to number of days till March 1, 1997, arbitrarily chosen last date later than the latest day observed in a data-set.

2 Model

2.1 Combined model

In short, the expected gift amount in excess of \$0.68 will be selected for mailing. A combined model was constructed for this selection. Firstly, we apply logistic regression to model the probability of an event occurrence of donation. Next we fit a ridge regression to model the donation amount. Then the expected donation from each person can be estimated by

$$r(x) = p(x)a(x)$$

where x is a vector features about an individual, $p(x)$ is the gift probability, and $a(x)$ is the gift amount. Later we will expand on how the expected donation can be used for economic decision-making.

2.2 Logistic regression for donation willingness

We estimate the willingness to donate p(x) by training a logistic regression model. Logistic regression is a linear regression method mapping features to a logit. The linear relationship between input x to an output probability p is formulated as following:

$$\log \frac{p}{1-p} = w^T x \Rightarrow p(x) = \frac{1}{1 + exp\{-w^T x\}}$$

Here x is augmented with a leading 1 to allow for an intercept. Learning is then done by changing the weights w. The former identity is how we would use the input x to produce probability p. Learning is done by the same principles as linear regression, using regularization.

$$\hat{w} = \text{argmax}_w \, C \sum_{x \in X} (p(x; w) - y_p(x))^2 + w^T w$$

Here $y_p(x)$ is the true binary labelindicatingif donation was given. Thedifferencebetween the C parameter and a ridge parameter is more of notation than anything else since one parameters only has the inverse effect of the other. In order to optimize the performance of the model we use a grid search with cross-validation to find a good parameter C. The performance metric is the area under the ROC curve (AUC). The model with the maximum AUC is chosen.

2.3 Ridge regression for donation amount

In order to estimate the donation amount $a(x)$ we implement a ridge regression model. For each prediction task, a linear regression model maps features of real values into a real valued label. This is done with a vector of weights w as follows.

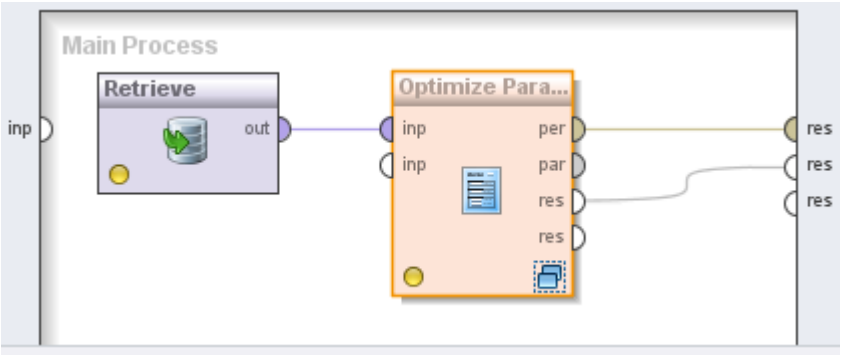

Figure 1: Logistic regression process tree part 1

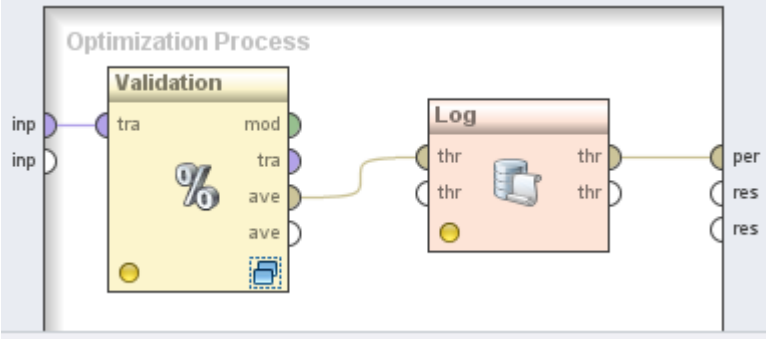

Figure 2: Logistic regression process tree part 2

$$a(x; w) = w^T x$$

Again, x is augmented with a leading 1 to allow for an intercept. These weights are learned by minimizing the error between the predicted label and the true label of the training set in conjunction with a regularization term. The error used in our case is the MSE. Thus the learning task reduces to the following

$$\hat{w} = \text{argmax}_w \frac{1}{|X|} \sum_{x \in X} (a(x; w) - y_a(x))^2 + \lambda w^T w$$

where $y_a(x)$ is the true donation amount and X is the set of feature vectors x

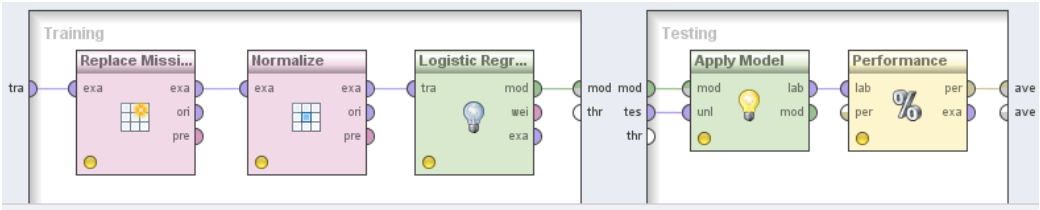

Figure 3: Logistic regression process tree part 3

attributed to each person of the training set. The actual minimization task is done by RapidMiner.

Regularization is again used in order to make the solution unique and tractable. This imposes a global model parameter called the regularization strength, or ridge parameter, denoted by λ. A good λ is found by grid search using cross validation for each parameter permutation. Finally, RMSE is used as a parameter selection metric. The model that minimizes RMSE is selected.

2.4 Cross validation and the final sub-models

In order to avoid overfitting when looking for the learning parameters C and λ we use cross-validation. Cross-validation works by partitioning the available dataset in to multiple folds (or bins if you will) and does training on all but one fold. The held-out fold is used as a test set for assessing classifier performance. This is done for all folds such that every fold is used once as a test set. The resulting performance measurements are then aggregated, by averaging in our case, in to an a robust and fair underestimate of the performance of the final model for a given learning parameter. When optimal learning parameters are found we retrain the model on the entire available dataset. Of course, this does not include the validation dataset which is not touch before final economic performance measuring.

3 Decision Making

3.1 Maximizing profit

To make an optimal decision it is not enough to just make predictions. We also need to put this in to a decision framework and account for the benefit of taking each possible action. Benefit can be an elusive concept, especially for non-profit organizations. However, in our case the object is quite clear. We will be assuming that the cost $c = \$0.68$ of each solicitation is appropriate and congruent with the charitys objectives and that the charity is in fact rational. That is, rational in its strictest economic sense. This means that we wish to maximize the expected overall profit of the donation campaign.

The available action space, i.e., possible actions, for each person is whether to solicit that person or not. If we let X be the set of all people in the data set `KDD98` and X be a subset hereunder we can state our objective function, the total donation profits, as follows

$$\pi^* = \max_{X'} \sum_{x \in X'} \pi_x, X' \subseteq X$$

where π_x is the donation profit from each individual, that is,

$$\pi_x = r(x) - c = p(x)a(x) - \$0.68$$

What we have stated here is that by deducting the cost of solicitation from the expected donation amount yields the total profit from a single donor. We can thus see that every non-negative individual contribution to the overall profit is an improvement. We want to solicit a donation from a person if and only if

$$\pi_x = p(x)a(x) - \$0.68 \geq 0$$

This paves the way to a dual and implementable way of stating our objective function

$$\pi^* = \sum_{x:\pi_x \geq 0} \pi_x$$

This means that we only want to send solicitations to individuals who we expect will yield a per-person profit.

3.2 Alternative Strategies

The campaign management can be presented with to alternative strategies for the solicitation campaign. Strategy `A` proposes to solicit all the donors, while strategy `B` limits itself soliciting only those donors for whom the expected profit is greater than zero. Another alternative, strategy `C`, could be considered where no soliciting mail is sent at all. While this saves the management the solicitation costs of around $65,530, it is difficult to calculate how many would still donate, even without solicitation. Thus we have no good way of estimating a profit we can measure strategy `C` up against the other two strategies. The total profit could be calculated from similar modeling on data based on donors who would have donated irrespective of the solicitation mail. Based on the given data set it is difficult to transform or carve-out a data set which identifies these kind of donors as the features such as `AVG_GIFT` and `LAST_GIFT` factor the effects of previous mail campaigns. Also its difficult to obtain data without the mail solicitation influence. One option of modeling in this scenario would be to model based on pure demographics but we did not pursue that route as it would mean throwing away a lot of data at hand.

4 Results and Analysis

4.1 Logistic Regression

In logistic regression, the goal is to estimate $p(x)$ which represents the probability that a donor x would contribute. Our objective is to find the model with the best area under the curve (AUC) and also a reasonable F1 score. We conducted a grid-search for parameter C of the Logistic Regression operator with values 1, 2, 4, 8, and 16. For each C value, a logistic regression model is built based on the training data after a 3-fold cross-validation. We measured AUC and F1 scores of each such model for every C and are mentioned in table 1

We picked the model obtained using $C = 8$ as parameter value as it has the best AUC and F1 scores. Using $C = 8$ as the guiding parameter we trained a new logistic regression model based on the entire training data and used the newly constructed model to estimate $p(x)$ on the test data. Our dual training strategy, enabled us to obtain an optimum C and use that C to build a model upon the entire data set and not just on the training set in a single fold during cross-validation. Using the model from the cross validation

C	AUC	F1
1	0.623	56.89
2	0.621	56.70
4	0.639	57.01
8	**0.645**	**58.69**
16	0.630	56.23

Table 1: Logistic regression parameter grid search results.

Min Prob.	Max Prob.	Mean Prob.	Variance	RMSE
0.0051	0.7639	0.0163	0.0718	0.2521

Table 2: Statistics and performance of the chosen logistic regression model.

could have a limitation of only using 2/3 of the training data in a single fold. The newly constructed model is used to estimate p(x) based on the test data. Table 2 depicts the results obtained after applying the calibrated model on the test data. RMSE values are calculated from the predicted and actual labels (which are given) for the test data set.

4.2 Ridge Regression

In ridge regression, the goal is to estimate the donation amount, $a(x)$. The linear regression model is constructed based on 10-fold cross validation on the training data. The grid search for the ridge parameter is run with values $0.000125, 0.00125$, and 0.0125 to train the model with different learning rates. The RMSE obtained for each ridge-parameter is shown below.

Similar to logistic regression, we pick the ridge parameter, 0.00125 as

Ridge parameter	RMSE
0.000125	15.23
0.00125	**13.82**
0.0125	14.56

Table 3: Ridge regression parameter grid search results.

Min Amount	Max Amount	Mean Amount	Variance	RMSE
0.0009	89.4821	0.7504	32.0399	14.5832

Table 4: Statistics and performance of the chosen ridge regression model.

Strategy	Solicitations	Profit	Uplift	Percent improvement
A	96,367 (all)	9,471	-	-
B	31,140	12,221	2,750	29%

Table 5: Profit profile for the alternative strategies A and B.

a guidance parameter to build our final linear regression model using the complete training set. The resultant model is used to estimate a(x) on the training data. Table 4 shows the statistics of the obtained estimates of a(x). RMSE values are calculated from the predicted and actual labels (which are given) for the test data set.

4.3 Economic results from prediction

From table 5, though the solicitation campaign can obtain a total profit of $9,470.86 if all the donors are solicited, it can be seen that sending solicitation mail to the selected subset of 31,140 potential donors yields a total profit of 12,220.61 and stretagy B uplifts the total profit by 2,749.75 over strategy A.

5 Conclusion and limitations

In this assignment we use data mining techniques to create an economic model to maximizing profits for a charity. We combined two models, a ridge regression model for amount estimation and a logistical regression model for donation likelihood estimation. The the ridge regression model for amount estimation yielded a RMSE of x and variance of y. And, the logistical regression model for donation likelihood estimation yielded a RMSE of x and variance of y. The implementation of our proposed strategy B would contributed to the total profit with an additional $2,750 compared to the naive solicit-everybody strategy A. This is a 29% improvement.

We apply isotonic regression to post-process prediction probabilities using R package called *Iso*. Unfortunately, we got less improvement. Thus, we stick to the raw probabilities. Fine tuning of calibration would have improved the results.

The main issue with our model is that both our models are either linear or just linear transformation, that is, a change in one of the features of an example can only contribute monotonically to the prediction. This limits the model such that it cant handle non-monotonic feature-label relationships such as u-shaped relations. Also, even with monotonic relationships, any other than linear and sigmoidal relationships will incur inevitable errors for the linear and logistical regression models, respectively.

Finally, a general limitation of our approach is that our objective functions is not directly tied to the charity's objectives. While the charity wants to maximize profits, we do this only indirectly by minimizing the error of our two sub-model errors. These are obviously congruent goals, but to what extent is not a trivial question. When all is said and done it is not clear how we would go about imposing a single direct objective function in this assignments context.