KDD-Cup 99 :
Knowledge Discovery In a Charitable Organization's Donor Database

Saharon Rosset and Aron Inger
Amdocs (Israel) Ltd.
8 Hapnina St.
Raanana, Israel, 43000

{saharonr, aroni}@amdocs.com

1. INTRODUCTION

This report describes the results of our knowledge discovery and modeling on the data of the 1997 donation campaign of an American charitable organization.

The two data sets (training and evaluation) contained about 95000 customers each, with an average net donation of slightly over 11 cents per customer, hence a total net donation of around $10500 results from the "mail to all" policy.

The main tool we utilized for the knowledge discovery task is Amdocs' Information Analysis Environment, which allows standard 2-class knowledge discovery and modeling, but also Value Weighted Analysis (VWA). In VWA, the discovered segments and models attempt to optimize the value and class membership simultaneously.

Thus, our modeling was based on a 1-stage model rather than a separate analysis for donation probability and expected donation (the approach taken by all of KDD-Cup 98's reported modeling efforts except our own).

We concentrate the first two parts of the report on introducing the knowledge and models we have discovered. The third part deals with the methods, algorithms and comments about the results.

In doing the analysis and modeling we used only the training data set of KDD-Cup 98, reserving the evaluation data set for final unbiased model evaluation for our 5 suggested models only.

If our goal had been only knowledge discovery, it might have been useful to utilize the evaluation data too, especially the donors. It is probably possible to find more interesting phenomena with almost 10000 donors than with under 5000.

2. MAIN RESULTS

1. Our best prediction model has achieved a maximal net profit of $15515 when checked against the evaluation data set. At its pre-determined cutoff point it achieved a net profit of $15040, compared to KDD-Cup 98's best result of $14712 net profit.

2. We have built a "white-box" model comprised of a collection of 11 customer segments. A policy of mailing only to the customers in these segments brings a combined net donation of $13397 for the evaluation data set. This "white box" model has the advantage that it is robust, understandable and can be implemented easily within the database, without need for additional tools.

3. Donation segments that are both highly profitable and actionable can be identified and utilized in the data. Two examples of these:

   People whose last donation was over $35 donate on average 3.5 times as much as the average donor, although their donation probability is lower than the average by about 30% (see figure 1). This segment's net donation for the training data set is $4100 (for 3500 people only!).

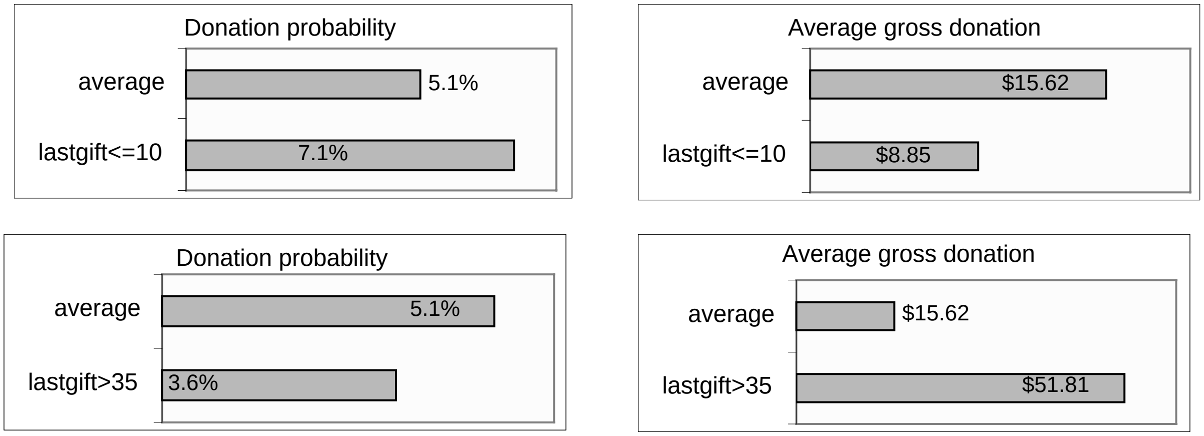


Figure 1: Donation probability and expected donation for customers with LASTGIFT>35 (below) and customers with LASTGIFT<=10 (above)

The approximately 14000 people who:
- live in an area where over 5% of renters pay over $400 per
- month (urban neighborhoods?)
- have donated over $100 in the past,
- have an average donation of over $12,

Account for $8200 net donation in the training data set.

4. Identifying donors is a thoroughly different task than maximizing donation. This can be illustrated in multiple ways:
   Some of the best donation models, when viewed as donation probability models, turn out to be almost random, i.e. at certain cutoff points the number of donors is approximately their average in the population, even though the overall donation is high (see example in modeling section below).

   Many segments can be identified which have a high net donation with less-than-average donation probability and vice versa. A striking example is illustrated in figure 1. It shows the segment described in section 3.a and its counterpart - people whose last donation was no more than $10 donate on average almost 45% less than the average donor, but their donation probability is 40% higher!

5. To examine the variability of profit gained by different models, we experimented with equivalent models on identical data and with identical models on equivalent data. Our results indicate that a difference of less than $500 in evaluation-set profit between models cannot be considered significant. Furthermore, it seems that even a difference of $2000 in profit is not significant if the models are evaluated on different data sets. This indicates the huge extent to which future performance of models can vary from their evaluation-set performance.

Our main discovery & modeling approach was a one-stage 2-class model based on value-weighted analysis. This approach accommodates the combination of knowledge discovery and modeling within the same process - so discovered knowledge is the foundation for the prediction models built. Figure 2 shows a screen capture of the application's display of discovered segments. It shows the same segment in the 2 views – weighted by total value and by number of customers.

3. DETAILED RESULTS

Discovered Knowledge

In this sub-section we describe some of the new understanding and insights about the data which we gained during our analysis. We concentrate on meaningful and potentially useful knowledge. The use of this knowledge for modeling is discussed in the next sub-section.

1. The most significant variables for predicting a customer's donation behavior are the previous donation behavior summaries. This can be seen in the form of correlation between the variables and the donation amount, and also in the best segments discovered by our algorithms. Some additional examples:

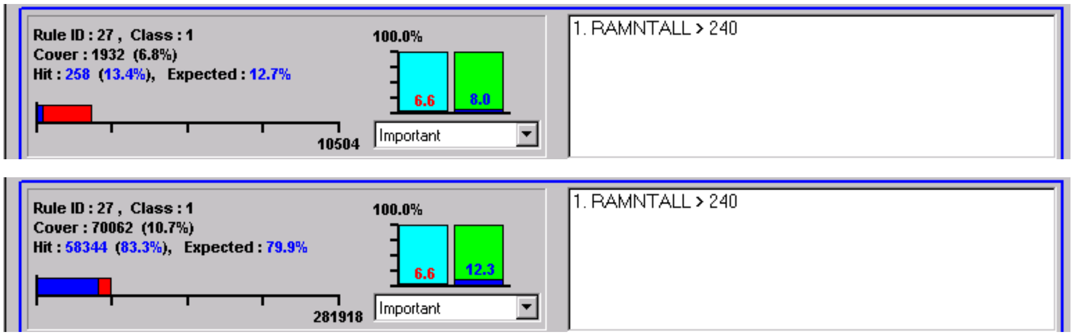
- The 6871 customers whose maximal donation exceeds $30 account for a net donation of $5608 in the training data (265 donors).
- The 5921 customers whose total past donation exceeds $250 account for a net donation of $4426 in the training data (343 donors).

The overlap between these two segments is surprisingly small - only 2412 customers who account for $2924 net donation (105 donors).

2. The NK phenomenon:

- The 2805 customers who have donated over $20 in the 95NK campaign account for $2705 of net profit in the current (97NK) data set.
- The people who have donated non-negligibly (over $3.50) in the 96NK campaign have a 5 times higher probability of donating than the average. Their average donation, however, is the same as the average.

Variables describing other campaigns did not form such powerful patterns in the results of our discovery algorithms. This may imply a need to investigate the connection between the donations in the different NK campaigns. Is there really a unique statistical connection here, as compared with other types of campaigns? Do these people like the NK mailing a lot, or are they once-a-year donors, donating every June, which just happens to be the yearly NK campaign? Time limitations have prevented us from looking for these answers, some of which can certainly be reached from the available data.



Rule ID : 27 , Class : 1
Cover : 1932 (6.8%)
Hit : 258 (13.4%), Expected : 12.7%

1. RAMNTALL > 240

Rule ID : 27 , Class : 1
Cover : 70062 (10.7%)
Hit : 58344 (83.3%), Expected : 79.9%

1. RAMNTALL > 240

Figure 2 : Visualization of discovered segment – by number of donors (above) and value-weighted (below)

3. The US-census (neighborhood level demographics) data turns out to be quite strongly connected to the donation performance of the population.

The variables which describe "richness", such as:

- HV2 - Average home value
- IC2 - Median family income

etc., have a strong positive correlation to donation, although in some cases the connection is weaker to donation probability. For example, the 12833 people defined by HV2 > 2000 account for a net positive donation of about $5750 in the training data. Their donation probability is 27% higher than the average, and their donation amount is 20% higher than the average donation.

It should be noted, however, that the best single model we have built (see below) uses very few of these demographic variables, and it seems that almost all the relevant information contained in them can be inferred from the individual customer attributes - mainly the donation history.

Modeling

Modeling was done on the training data, and only the final chosen models were then evaluated on the evaluation data set, to gain a reliable measure of their "true" performance.

In this section we describe the chosen models and their "knowledge" value. Technical discussion of modeling techniques is deferred to the next section.

The total number of final models built was 5, and they are described below: two “white box” interpretable and easy-to-use models; one relatively simple model, based on 40 variables only; and two candidates for “best overall” model which indeed turned out to be the best by far.

Building the white-box model

To build this model, we have collected 11 "good" segments from the different analyses we have run. We then "rounded" them to create more meaningful patterns. The total net donation of these 11 segments is $13397 for 55086 customers.

These segments are:

1. MAXRAMNT > 30
2. RAMNTALL > 250
3. HV2 > 2000
4. RAMNT_14 > 25
5. IC15<=45 & LASTGIFT>5 & LASTDAT>9606
6. RP2>5 & RAMNTALL>100 & AVGGIFT>12.0
7. RP2>5 & LASTGIFT>15 & LASTDATE>9500 & RICH* > 250 & POOR* <= 500 & JOBS2* <= 45
8. STATE in ("CA", "MI") & NUMPROM > 30 & LASTGIFT > 10 & LASTDATE >9504
9. MAJOR = "X"
10. HV2 > 1500 & LASTGIFT > 5
11. IC4 > 450 & LASTDATE > 9503

* - calculated fields summarizing demographic information

The use of this model has several major advantages for direct mailing campaigns:

- The reasons for the scores are obvious. Thus if there is a change in conditions (e.g. a change in average income) it might be possible to adjust the model without the need for re-modeling.
- The total number of variables figuring in to the model is relatively small.
- The model can be implemented within the operational database, with no need for external scoring procedures.

We can look at this model as a simple way to improve profits by over 25% compared to the full mailing without much effort.

Interestingly, the training data net profit for this model was $14500, so although overfitted, it is not wildly so (and much less than the more complex models – see discussion of this point below).

Figure 3 shows the incremental net profit, which the different segments provide.

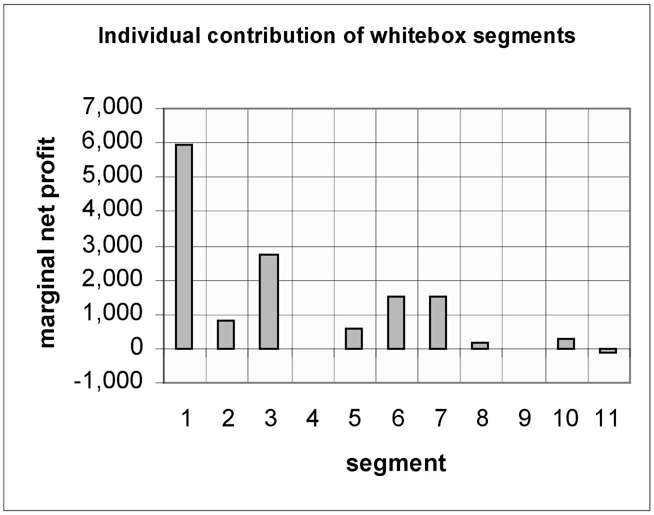


Figure 3: Incremental net donation on evaluation data for white-box model segments

In addition to this 11-segment white-box model, we have also built a 7-segment model, by using an automated selection algorithm, dropping segments 1,9,10,11 above. This model is more compact, and resulted, for the evaluation data set, in mailing to only 40251 customers and a net profit of $12913. From figure 3 it is evident that a 7-segment model of segments 1-7 would have done slightly better, netting $13087.

Best single model

Our chosen model, based on leave-out test-set performance, was generated from a run that used a much reduced group of predictors. After rigorous variable selection (both automated and manual) we selected a group of 31 original variables, plus 9 additional demographic summary variables, such as a "Rich" indicator summarizing demographic variables relevant to economic status.

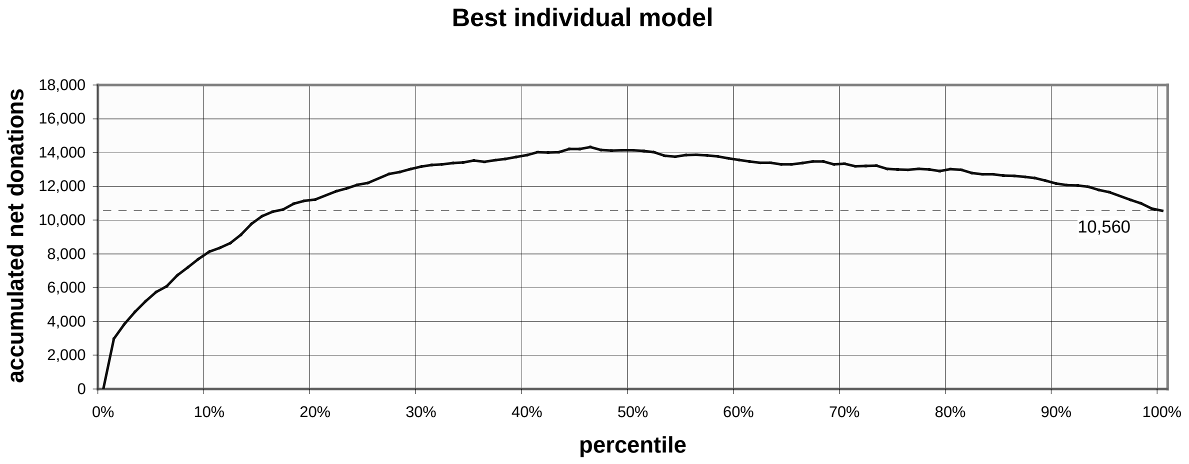


***Figure 4**:* Cumulative donation for suggested “best individual model” on the evaluation data

The modeling technique was a "hybrid" logistic regression model, utilizing the 109 discovered segments as binary variables, as well as the 40 predictors.

Calculating the expected donation for the customers on the evaluation data set, we got a cutoff point (where expected donation crosses $0.68) of about 52%. The actual net profit for the evaluation data set at this point was $14067. Looking at figure 4, we can see that this is a sub-optimal cutoff point for the actual performance. The observed best profit on the evaluation data set is $14377 at 44%. Figure 5 shows the lift graph for the evaluation of this model as a “non value weighted” model. It can be seen clearly that it is hardly any better than random at most points. This is an example of the profound difference between the problem of identifying donors and maximizing profit.

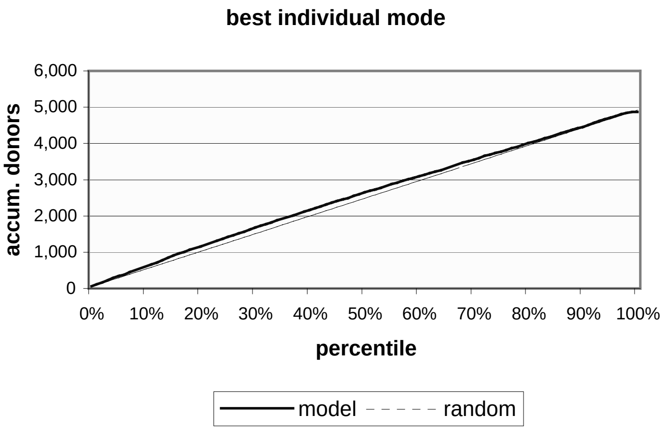


Figure 5: Evaluation of “Best individual model” as a donation probability model

Improving prediction by averaging

Our experience, as well as the literature, shows that the performance of the single models can be enhanced by averaging them, creating new ensemble models.

Our suggested ensemble models were based on 6 models generated in 2 different runs of our knowledge discovery and modeling system. The runs differed in the set of predictors used (all variables versus the reduced set of the "Best single model" in item 2 above), while models within each run differ in the modeling technique.

One of the suggested models was based on the average of one model from each run, and the second on the average of all 6 models.

Both models achieved a maximal donation of over $14712 on the evaluation data, which was KDD-Cup 98’s best result. The first, simpler model achieved a maximal net donation of $15515 at 47%, and a donation of $15040 at our pre-determined cutoff point of 49% (see figure 6). Indeed, at all 20 of the possible cutoff points between 41% and 60% this model achieves a higher net donation than $14712.

The second model achieved a maximal net donation of $14899 at 48% and $14439 at our pre-determined cutoff point of 60%.

Variance of model profit

It is evident from our experiments that the net profit which models generate has a very large variance. This is also intuitively clear from the fact that donations can be rather large and a few large donors can change the net profit significantly.

To illustrate the dependence on random effects in the data, we ran 10 bootstrap-95412 samples from the training data through our selected single model (item 2). The difference between the minimal and the maximal net profit at a fixed cutoff point (50%) was over $4000, with standard deviation over $1000.

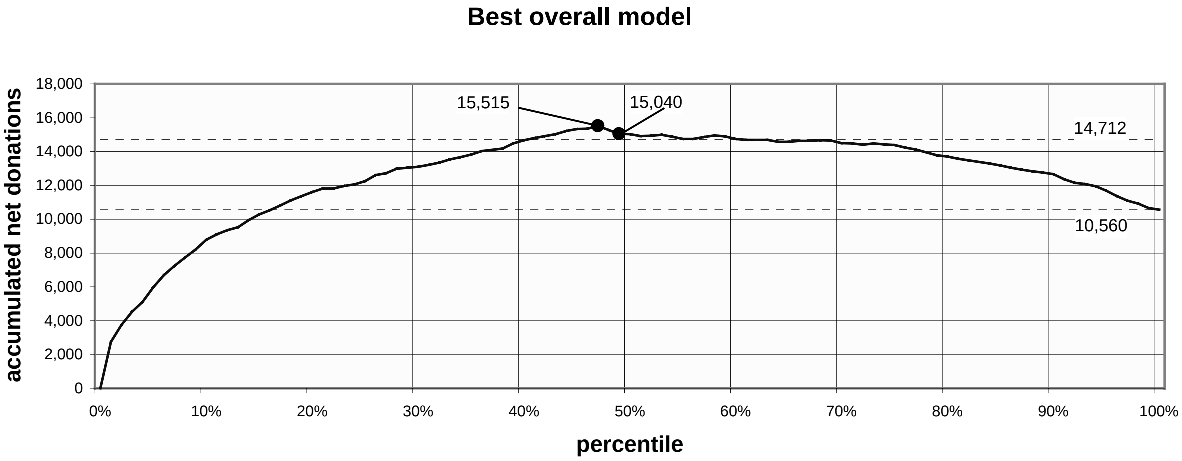


Figure 6: Evaluation of best ensemble model

The main conclusion from this experiment is that good leave-out test set performance can hardly be considered a reliable indication of good future performance - or in our case, good evaluation data-set performance.

Our second experiment attempts to show the difference in performance of "similar" models on the same data. For that purpose we created 10 "different" models by adding random noise to our chosen model scores. The range of results at 50% of evaluation data was over $1000, with a standard deviation of over $400.

The obvious conclusion here is that a significant difference in model performance should definitely exceed $500, and thus we can safely say, for example, that for the KDD-Cup 98 results, the results of places 1 through 2 and 3 through 5 were not significantly different.

4. TECHNICAL DETAILS & COMMENTS

Value-weighted rule-discovery

2-class Value Weighted Analysis deals with 2-class problems, where the question of interest is not "what is the probability of this customer belonging to class 1" but rather "how much are we likely to gain/lose from this customer". The answer to the second question depends both on his class membership and on his value as a customer. Our first encounter with this setup has been in churn analysis, where just identifying likely churners was not sufficient, and the real goal of the analysis process was defined as "finding segments of customers where we are losing a lot of money due to churn".

As can be seen from the results, Value Weighted Analysis aims at and succeeds in finding rules which are interesting from a value weighted point of view, rather than from a "customer weighted" point of view.

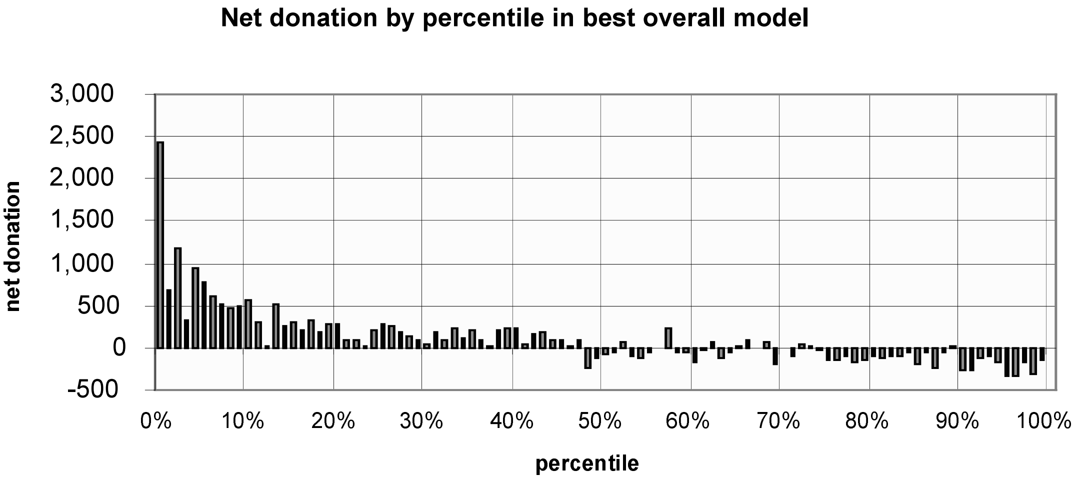


Figure 7: Donation by percentile in best ensemble model – all percentiles up to 48% have a positive net donation. All but 9 of the percentiles from 49% on have a negative net donation.

For our problem, customer value was defined as his net donation. Thus, all non-donors were given a value of $0.68 (cost of mailing) and class 0, and the donors were given a value of their donation minus $0.68 and class 1.

Our Value Weighted rule discovery method is based on a standard C4.5-like tree algorithm, where the splitting and pruning criteria have been modified to accommodate value-weighted analysis.

In the splitting criterion, we look for splits that create unbalanced groups in terms of total value, rather than in terms of number of records.

The pruning process, whose basic component is the "pessimistic approximation" mechanism, is modified by identifying that a rule (or segment) is now a collection of records with different weights. Hence the variance of its empirical accuracy can be calculated, and the "pessimistic approximation" is simply the lower end of the resulting confidence interval.

For the analysis of rule-discovery results, and integration of human and machine knowledge, we have developed a visualization & analysis tool (see figure 2). This tool has also been modified to be able to display rules, segments and customer attributes either in a value-weighted manner or in the normal ("number weighted") manner.

Model building process

Within our tool, we have an array of self-developed modeling techniques, using the combination of discovered segments, user-generated segments and original predictors as the building blocks for the models. The models generated are logistic regression or neural network models.

To complement value-weighted discovery, we have developed value-weighted modeling as well. The scores given by the models thus reflect a "generalized probability", conveying in our case, the balance between donation and non-donation, which this customer represents.

Calculating expected donation

The scores which Value-Weighted modeling gives to the customers approximates the "generalized probability", i.e. the ratio between the customer's expected net donation and his expected "total donation":

$$s \cong \frac{ed - 0.68p}{ed - 0.68p + 0.68(1-p)}$$

With ed denoting the expected net donation of the customer and p denoting the customer's donation probability.

From this we can generate a formula for estimating ed for our customers:

$$ed \cong \frac{0.68s + 0.68p - 1.36sp}{1-s}$$

We utilized this formula in calculating expected donation for the models on the evaluation data. The p's were taken from a separate model, estimating the customer's donation probability in a non-weighted manner.

It seems that our estimations of the ed's were a little too high, leading to suggested cutoff points (where ed=0.68) that were too high for all of the models described above. This may be due to the effects of overfitting in the creation of the models, causing them to generate scores that are "optimistic".

Overfitting

A surprising result is the large amount of overfitting, which the models display on the training data compared to the evaluation data. For our 5 chosen models, we found overfitting of the net profit between $1100 (for the simplest model , i.e. the white-box model) to over $6000 (for the second of the two average models, which had a net profit of almost $21000 on the training data). This re-iterates the importance of limiting the use of the evaluation data, to achieve reliable predictions of future profit.