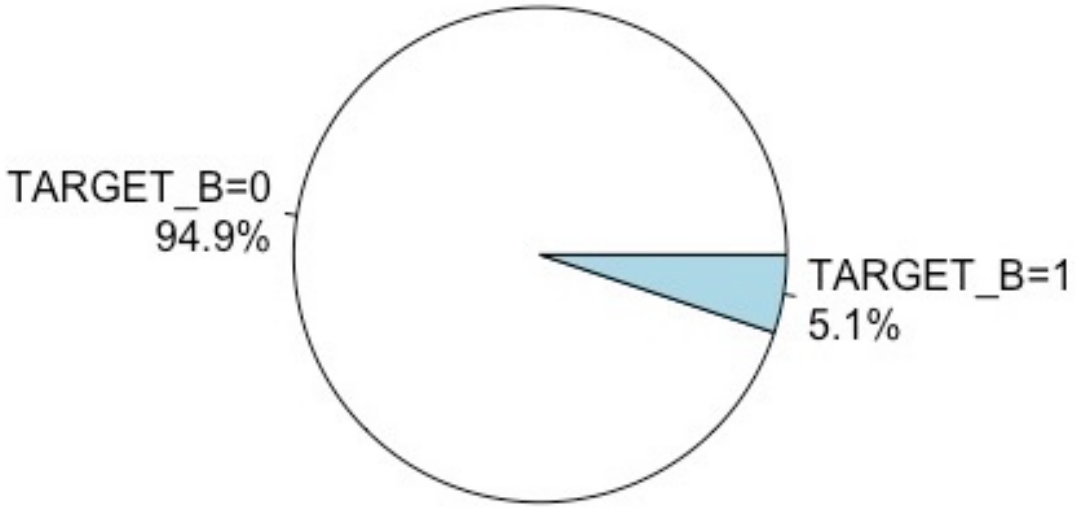
PROJECT

- Source code of your solution.
- Project report (including the interpretation of the algorithm, your implementation details, evaluation strategy and performance analysis results). Novel algorithm or improvement on some existing algorithm is a plus (optional).

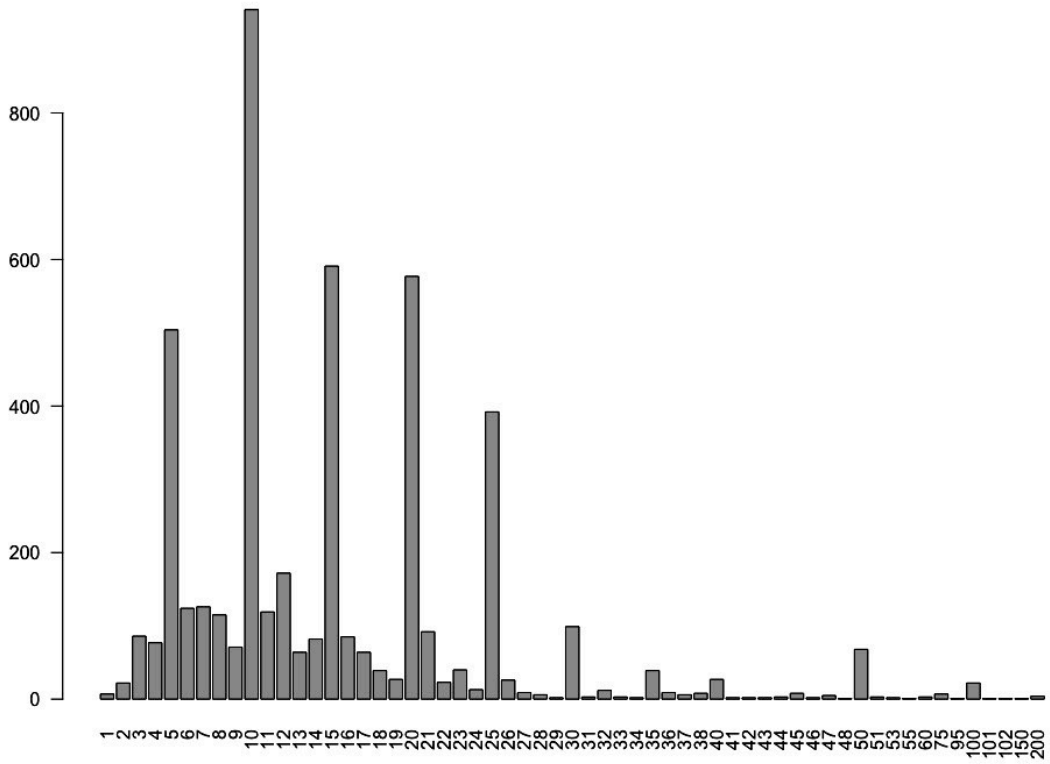
The learning dataset “train.txt” contains 95,412 records and 481 fields, and the validation dataset “validation.txt” contains 96,367 records and 479 variables. Each record has a field CONTROLN, which is a unique record identifier. There are two target variables in the learning dataset, TARGET_B and TARGET_D. TARGET_B is a binary variable indicating whether or not the record responded to mail while TARGET_D contains the donation amount in dollar.

Data inspection and variable selection

First we check the distribution of the two target variables, TARGET_B and TARGET_D. The plot shows that most donations are no more than $25 and are multiples of $5. Based on the distribution shown in the above barplot, we discretize TARGET_D to make a new variable TARGET_D2.



Then we check the number of positive donations not in whole dollars, round the donation amount to whole dollars and then draw a barplot for it. The plot shows that most donations are no more than $25 and are multiples of $5. Based on this barplot, we discretize `TARGET_D` to make a new variable `TARGET_D2`, that the intervals are open on the right and closed on the left.



Variable RFA_2R (recency code for RFA_2) is removed, because all records have the same value of "L" in that field. Around 99.7% of records has a value of "0" in field NOEXCH, so it is also removed. We also excluded characteristics of the donors neighborhood, and variables from the promotion history file and the giving history file. However, the summary variables from the two history files are kept. After the above inspection and exploration, the following variables are selected.

DEMOGRAPHICS

ODATEDW	CDPLAY
OSOURCE	STEREO
STATE	PCOWNERS
ZIP	PHOTO
PVSTATE	CRAFTS
DOB	FISHER
RECINHSE	GARDENIN
MDMAUD	BOATS
DOMAIN	WALKER
CLUSTER	KIDSTUFF
AGE	CARDS
HOMEOWNR	PLATES
CHILD03	INCOME
CHILD07	GENDER
CHILD12	WEALTH1
CHILD18	HIT
NUMCHILD	COLLECT1
VETERANS	HOMEE
BIBLE	PETS
CATLG	

HISTORY INFORMATION

PERSTRFL	CARDGIFT
CARDPROM	MINRAMNT
MAXADATE	MAXRAMNT
NUMPROM	LASTGIFT
CARDPM12	LASTDATE
NUMPRM12	FISTDATE
RAMNTALL	TIMELAG

ID'S & TARGETS

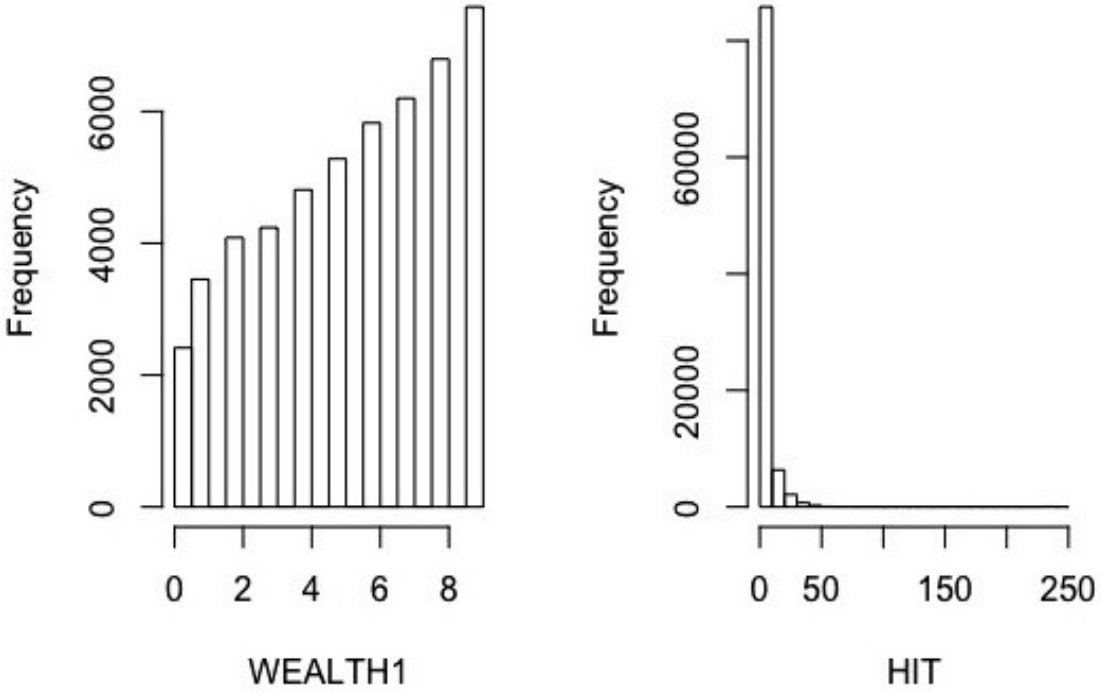
CONTROLN	TARGET_D2
TARGET_D	TARGET_B

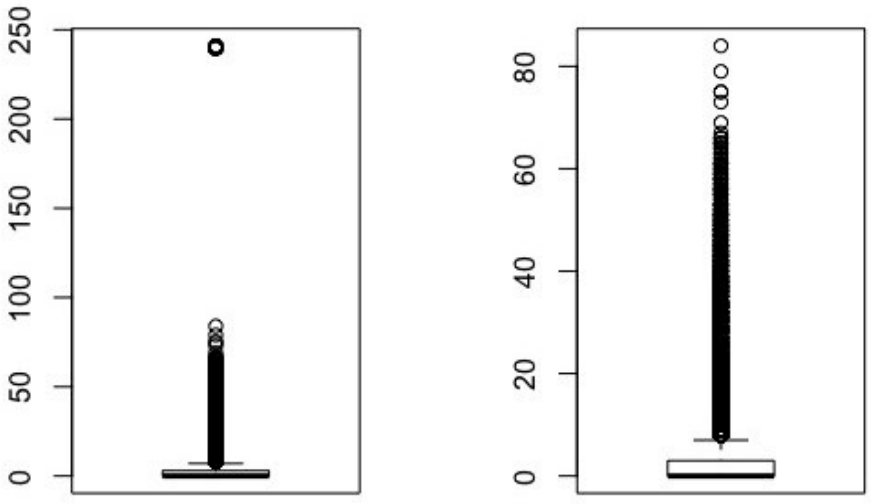
OTHERS

HPHONE_D	CLUSTER_2
RFA_2F	GEOCODE2
RFA_2A	MDMAUD_F
MDMAUD_R	MDMAUD_A

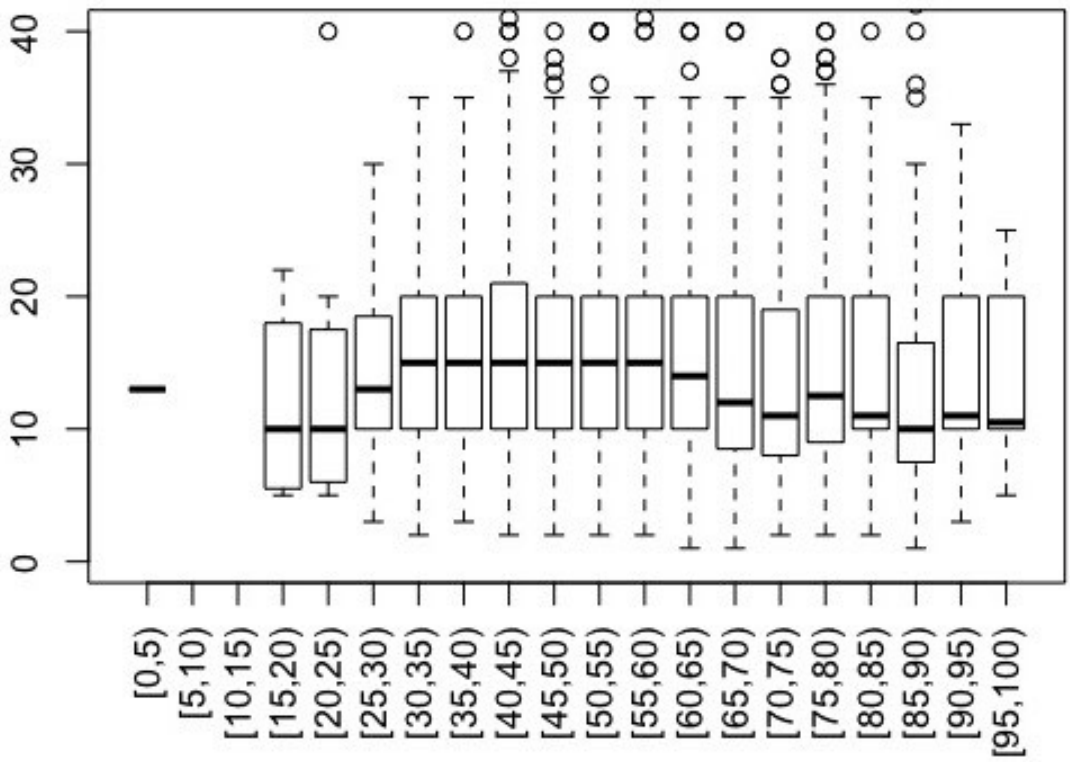
Data exploration

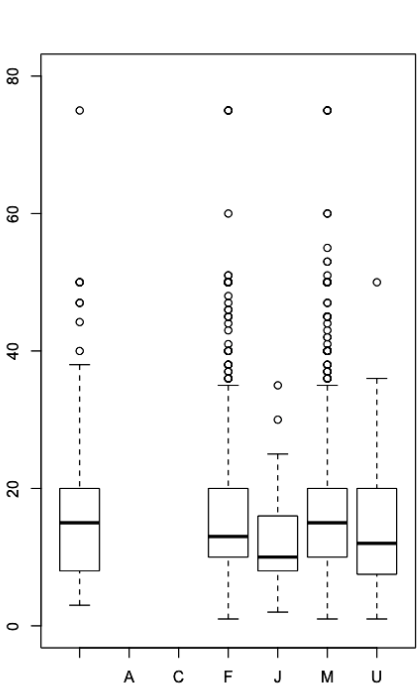
We first have a look at summary status of the data and the distribution of numeric variables. Here is an example for WEALTH1 and HIT variables. Other plots are available in the appendix. From this graphs we can see that HIT distribution have some values separated from the majority of `HIT`. A further checking shows that they are all of values 240 or 241.

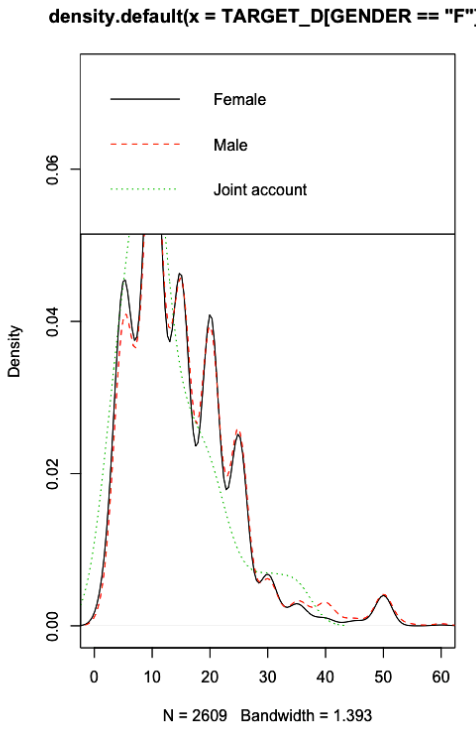




We then check the distribution of donation in various age groups. It is shown that people aged 30 to 60 are of higher median donation amount than others. It makes sense because they are the working force. Below we check the distribution of donation amount for different genders. The results show that the donation amount from joint account (“J”) is less than male (“M”) or female (“F”).





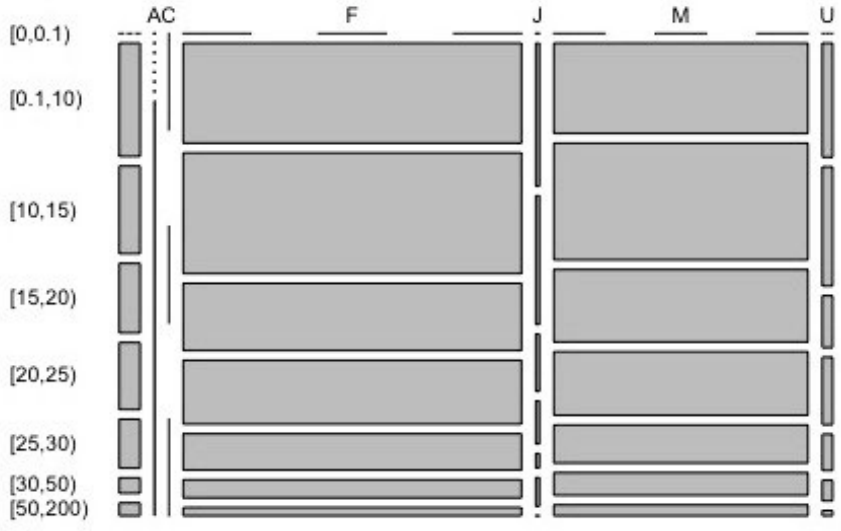


Next step is to check the correlation between the target variable and other numeric variables. The correlation between each pair of variables is computed using all complete pairs of observations on those variables, so that the resulting values will not be NA when there are missing values in the data. For categorical variables, we checked their association with chi-square test, below is given example of GENDER

(X-squared=NaN, df=42, p-value=NA), other variables can be seen in Appendix. These results are shown in the next tables.

	TARGET_D	TARGET_B	LASTGIFT	RAMNTALL	AVGGIFT	MAXRAMNT
TARGET_D	1	0,774	0,062	0,045	0,044	0,039
	INCOME	CLUSTER2	NUMPRM12	WEALTH1	MINRAMNT	LASTDATE
TARGET_D	0,032	0,029	0,025	0,025	0,020	0,019
	NUMPROM	CLUSTER	CARDPM12	NUMCHLD	CONTROLN	CARDPROM
TARGET_D	0,017	0,017	0,016	0,015	0,013	0,011
	FISTDATE	ODATEDW	HIT	CARDGIFT	NGIFTALL	MAXADATE
TARGET_D	0,008	0,007	0,007	0,006	0,005	0,004
	TIMELAG	DOB	HPHONE_D	AGE	RFA_2F	
TARGET_D	0,004	0,003	0,002	0,002	0,001	

GENDER



After this checks we will exclude some variables, below is final set.

DEMOGRAPHICS

STATE
RECINHSE
MDMAUD
DOMAIN
AGE
HOMEOWNR
NUMCHILD
INCOME
GENDER
HIT
PETS
COWNERS

HISTORY INFORMATION

PERSTRFL
CARDPROM
NUMPROM
CARDPM12
NUMPRM12
RAMNTALL
NGIFTALL
CARDGIFT
MINRAMNT
MAXRAMNT
LASTGIFT
TIMELAG
AVGGIFT

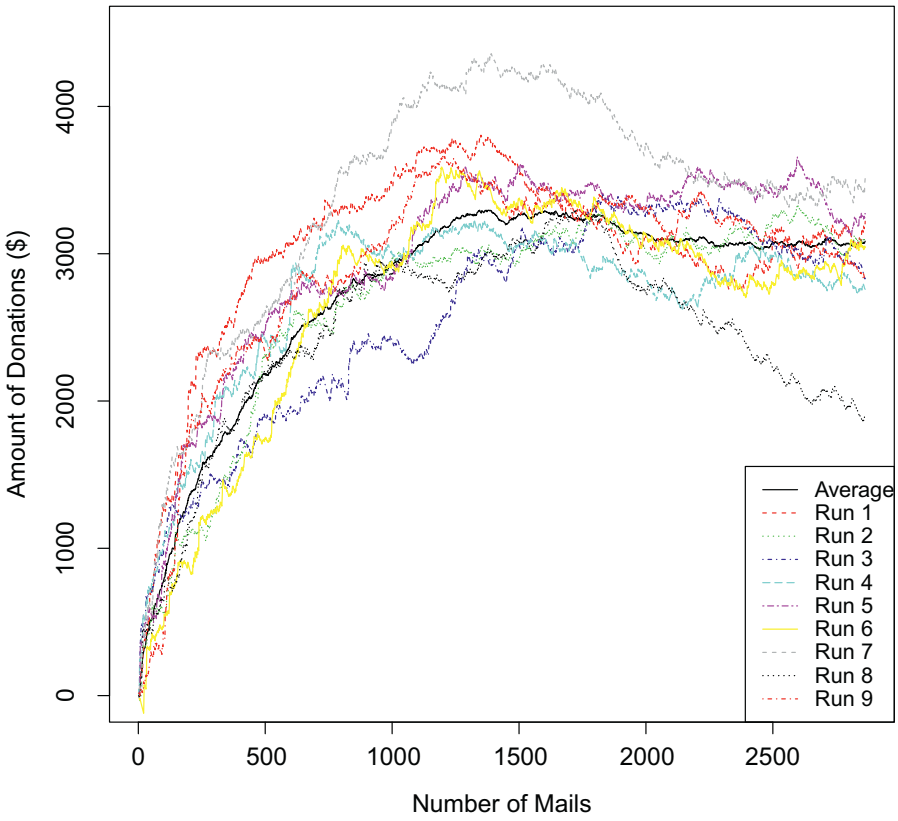
OTHERS

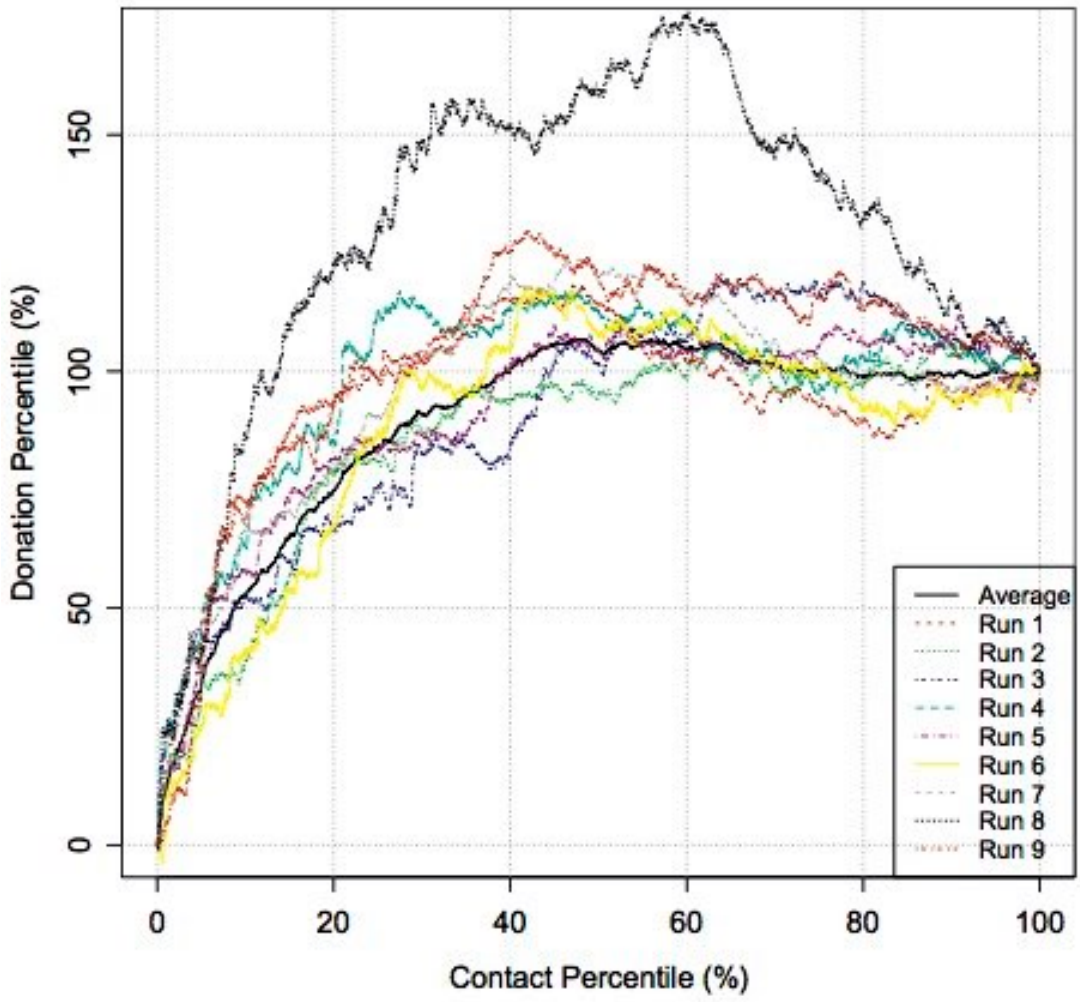
HPHONE_D
RFA_2F
RFA_2A
MDMAUD_R
CLUSTER_2
GEOCODE2
MDMAUD_F

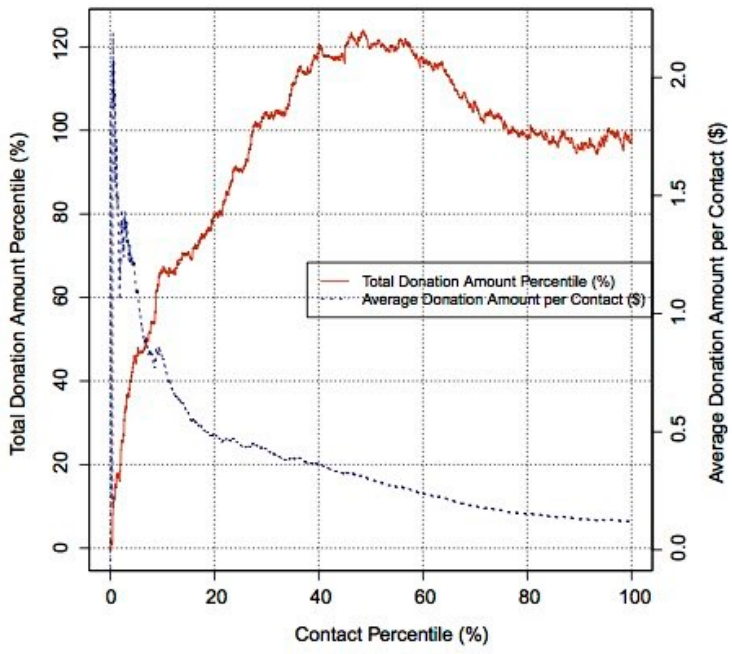
Building Decision Trees

In order to predict donation amounts we will use Decision Trees, they require setting following parameters: MinSplit, MinBusket, MaxSurrogate, and MaxDepth, to control the training of decision trees. MinSplit is the minimum number of instances in a node in order to be considered for splitting, MinBusket sets the minimum number of instances in a terminal node, MaxSurrogate stands for the number of surrogate splits to evaluate, and MaxDepth controls the maximum depth of the tree. We will use training dataset and divide it into two parts: training data (70%) and test data (30%), and the parameters for training decision trees. The MinSplit or MinBusket can be set to be of the same scale as 1/100 of training data.

With a decision tree model, the customers are ranked in descending order based on the predicted amount that they would donate. We plot the result of every run with the code below and the results are shown below. In the figures the black solid line illustrates the average performance of all nine runs, while the other lines are the performance of individual runs. The two figures show that run 7 produced the best result.

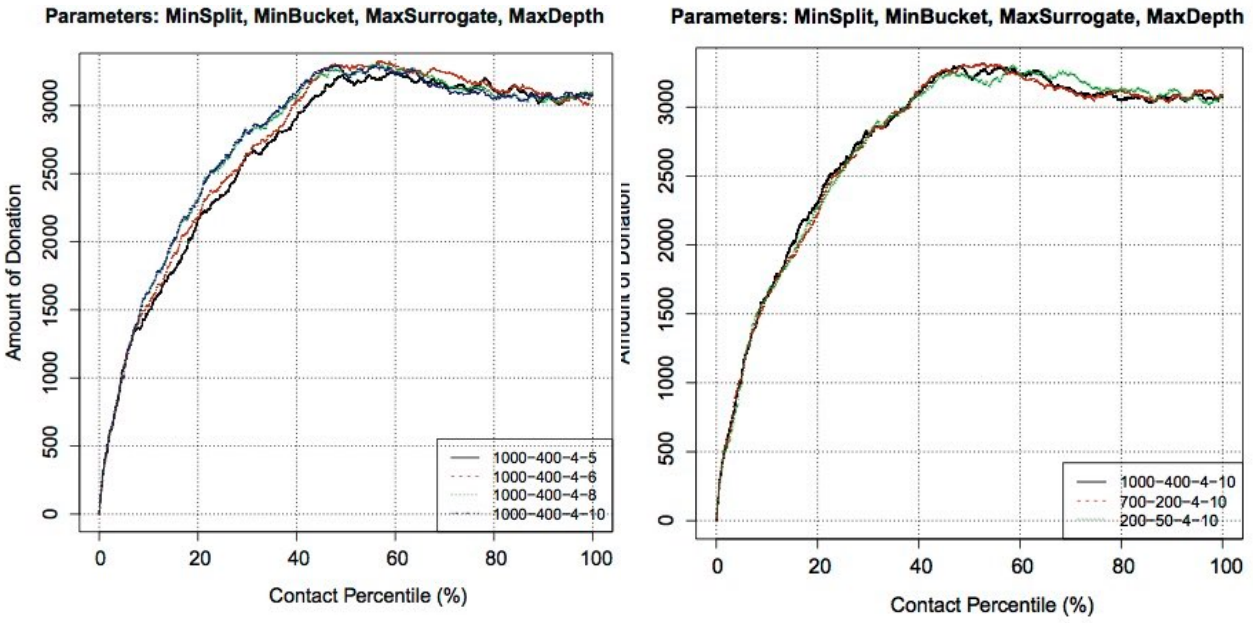






The next plot shows average result of the above nine runs, where the red solid line shows the percentage of donation amount collected and the blue dotted line shows the average donation amount by the customers contacted. The average donation amount per customer contacted is high in the left of the chart and then decreases when more customers are contacted. Therefore, the model is effective in capturing in its top-ranked list of the customers who would make big donations.

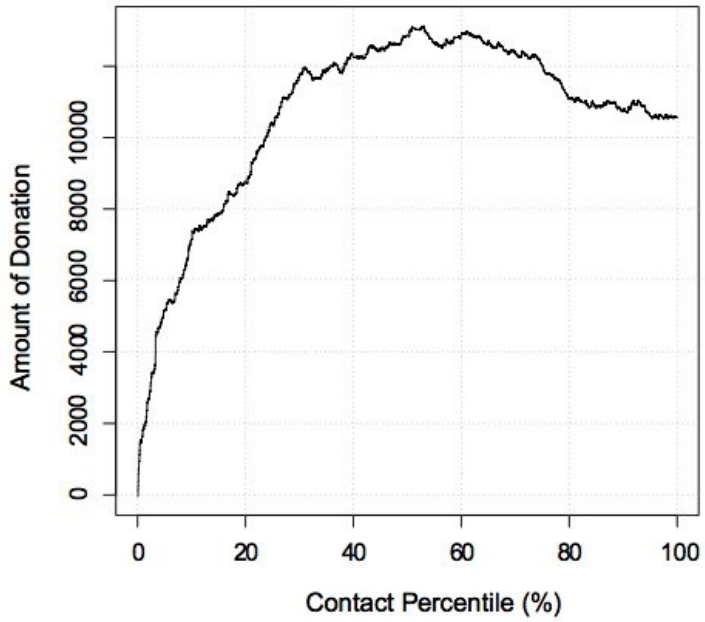
In this decision trees we set parameters randomly, now lets try different sets and choose the best one. The average results of running each setting nine times are given in next figures. The labels in the legend show the values of MinSplit, MinBusket, MaxSurrogate, and MaxDepth used in the six sets of parameters. For example, with the first setting "1000–400–4–5", MinSplit is set to 1000, MinBusket is 400, MaxSurrogate is 4, and MaxDepth is 5. Three different values are tested for MinSplit, which are 1000, 700, and 200. The corresponding values for MinBusket are 400, 200, and 50. The MaxDepth is also tried with four values: 5, 6, 8, and 10. The MaxSurrogate is set to 4 in all experiments



Results for testing trees with different parameters are shown above, where the horizontal axis represents the percentage of (ranked) customers contacted and the vertical axis shows the amount of donations that could be collected. A model is expected to collect more donations with the same number of contacts. We can see that results with depth 8 and 10 are better than depth 5 and 6. Also three different sets of minimum bucket size and minimum split size have very similar results. We choose "1000–400–4–10" to produce the final model, because it is less likely to overfit than other models with smaller minimum bucket sizes and split sizes.

Prediction

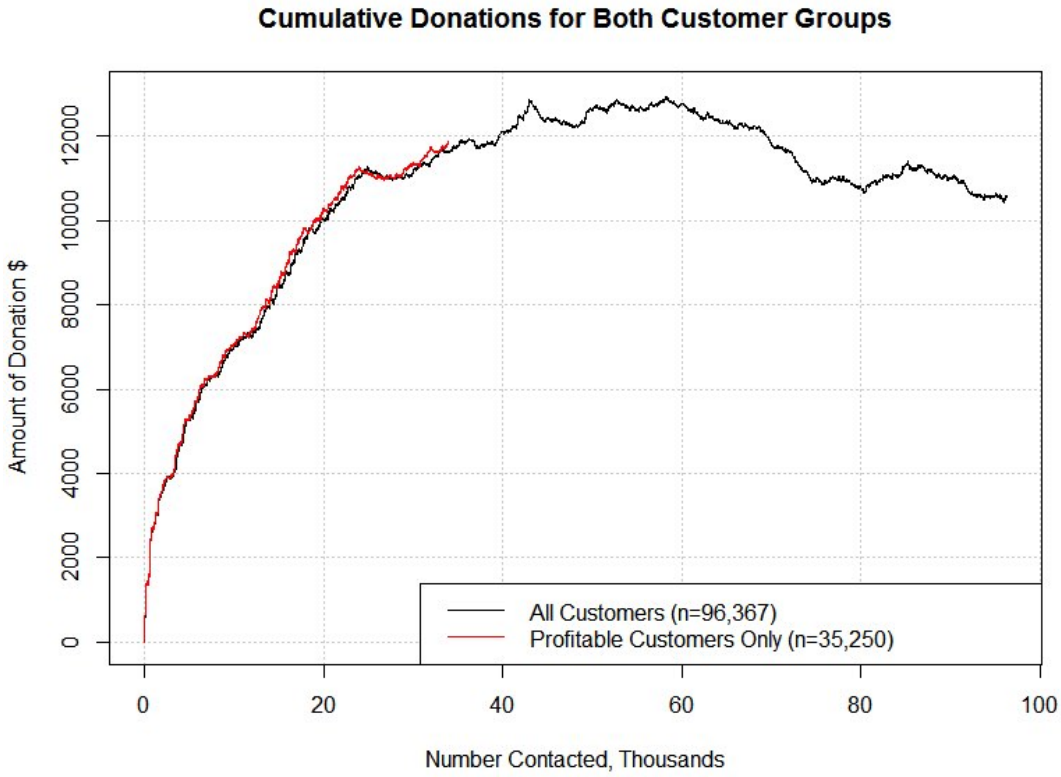
Finally, we will use the selected tree to score the validation dataset "validation.txt". People with a predicted donation amount greater than 0.68$, which is the cost of contact, would be mailed for donation purpose. The evaluation criterion is the total amount of donations deducted by the total cost of mail.



The above result shows that the model would produce a profit of $13,087. How can we check, whether it is good or not? - let's compare our profit with situation, if we will send letter to every customer in the validation dataset and send to no one. Recall the donation profit for all customers is just $10.560. How about only mailing the orders to those predicted to donate more than the $0.68 cost? Creating an index of those predictions more than the cost, we can subset the donation values, and arrive at $11857,77.

	N	Min	Mean	Std	Max	Sum
Response	96367	0	5,1 %	21,9 %	1	4,873
Donation_amount	96367	0 $	0,79 $	4,73 $	500,00 $	76,090 $
Profit - $0.68	96367	-0,68 $	0,11 $	4,73 $	499,32 $	10,560 $

Now we turn to the plot with both cumulative donation lines coexist, below. Instead of using customer percentile in the x-axis, we use the customer number in thousands to compare the number of customers mailed. With the all customer plot a familiar sight, note the red line restricting donation amounts to customers who donated more than the cost. The red line mainly follows the all customers line, but begins to overtake the regular line in cumulative donations around 10k customers. It stops just under $12k after 30k customers. The black line includes all customers, regardless of how they donated, so it extends farther than the selective red line. Here you can compare the y-axis total donation end points for both customer groups, with the red line ending higher then the black line.



Discussions

Decision tree is a good method to predict total donation amount, as was shown in this work. Therefore we always need to look, how can we improve the results, what are the weak points of introduced model. There were two target variables - "TARGET_B" which indicated yes or no to donation status, and "TARGET_D" which described the amount of donation. A next step would be to create a two-step model to predict who would donate, "TARGET_B", and then of those who would donate, how much ("TARGET_D"). In this case we may reduce mailings to customers with small donations amount and focus on optimal customer selection. Also, if this data is ongoing, and we will work with information about donations for last 20-30 years our simple model with hand-selection of decision trees may cost too much memory and performance speed. In order to solve this problem, we may use random forest.