Assignment II (Individual)

November 4, 2017

The solutions to the assignment must be submitted by Sunday, 17 December, 11.55 PM. PLEASE SEND YOUR ASSIGNMENT AS A SINGLE ZIP (RAR) AS "MLDM_II_Name.ZIP" TO apds.mldm@gmail.com WITH THE SUBJECT LINE "MLDM_II".

The Problem

An insurance company started a new insurance policy this month. The company has launched a direct marketing campaign to expand its policy holder base. To this end the company sends one of its staff members to personally interact with potential customers who are likely to buy the insurance in an effort to convert her/him to a new policyholder.

On an average, interacting with (that is, probing) each person costs the company Rs. 1,000 including expenses toward transportation, brochure, salary of the staff member etc irrespective of whether the person probed would eventually buy the policy or not. The average profit made by the company per policy holder per year is Rs. 10,000. So, if the person becomes a policy holder after probing, the net profit to the company is Rs. 9,000 (=Rs. 10,000 - Rs.1,000). However, if she doesn't become a policyholder the net loss to the company is Rs. 1,000 (as the probing cost of Rs. 1,000 gets wasted!). Similarly, if a person who would really have turned into a policy holder is not probed then the company loses a potential policyholder and the net loss to the company is Rs. 9,000 (the opportunity cost of missing a policyholder). Needless to mention if a person who wouldn't have purchased the policy is not probed there is neither any profit nor any cost to the company as no probing cost is incurred and there is no opportunity cost either.

The company wants to use predictive analytics to predict whether a person would buy its insurance product. If the prediction is "Yes" the person will be probed and if the prediction is "No" the person will not be probed. It may be assumed that the awareness level of the prospective customers about the product is such that no person would buy the product unless probed. Needless to mention, a person may or may not buy the product even after probing.

Table 1 below summarizes the net profit to the company per prediction depending on the various possible scenarios of the prediction about a person her/his true decision regarding purchase of the insurance product.

Table 1: Profit to the company per prediction (negative value indicates loss)

		CUSTOMER DECISION	
		WOULD BUY (if probed)	WOULDN'T BUY (irrespective of probing)
PREDICTION about the CUSTOMER	YES – would buy (person gets probed)	9000	-1000
	NO – wouldn't buy (person is not probed)	-9000	0

The company wants to maximize its profit by increasing is policy holder base by probing potential customers with probing implications in tune with Table 1 above. Consequently, it's very important for the company to correctly take a decision regarding the individuals in the population who are to be probed.

To meet this goal the company wants to buy a classification model, which will predict whether a person, if probed, would buy the insurance policy or not based on the profile of the customer. The company has a huge database of customers who had been probed in the past for selling similar insurance policies. Some of them had bought those policies and others didn't despite coaxing them.

The company has randomly chosen a subset of 1000 records out of this huge database of records from the past and these 1000 records are being handed over to you, the data miner, for building classification models. The company has also randomly chosen 200 records from the remaining database (which are not being shared with you) for evaluating the performance of the models you will build.

Each record, representing a person, has 76 attributes of which the first 75 are predictor attributes and the last attribute, named OUTCOME, denotes whether the person had become a policyholder (OUTCOME=1) or not (OUTCOME=0) after (s)he was approached in the past for selling an insurance product.

Once the company receives the model, it will see how much profit the company would have made if it had applied the model on the 200 records to decide who among them were to be probed. Suppose, among those 200 people only 50 actually became policy holders (50 is not the true number, it's for illustration only). The following scenarios of four hypothetical models explain how the company evaluates the models by computing the profit in tune with Table 1. The profit is obtained by multiplying the corresponding elements of Table 1 and in the Table for the model.

Model 1: All records are predicted to be NO (that is, OUTCOME=0 for all 200 records).
Model 2: All records are predicted to be YES (that is, OUTCOME=1 for all 200 records)
Model 3: Some 110 records are predicted to be YES (of which 30 actually bought the product) and the remaining 90 are predicted to be NO (of which 20 would have been policy holders if probed!).
Model 4: The ideal model – only the 50 are predicted to be YES (and all of them bought the product) and remaining 150 were predicted NO (none of them would have become policyholders even if probed)

Table showing the performance of Model 1:

		TRUE OUTCOME		
		OUTCOME=1	OUTCOME=0	TOTAL
PREDICTED OUTCOME	OUTCOME=1	0	0	0
	OUTCOME=0	50	150	200
	TOTAL	50	150	200

Profit(Rs): 0(9000) + 50(-9000) + 0(-1000) + 150(0) = - 4,50,000, that is, loss of Rs. 4,50,000.

Table showing the performance of Model 2:

		TRUE OUTCOME		
		OUTCOME=1	OUTCOME=0	TOTAL
PREDICTED OUTCOME	OUTCOME=1	50	150	200
	OUTCOME=0	0	0	0
	TOTAL	50	150	200

Profit (Rs): 50(9000) + 0(-9000) + 150(-1000) + 0(0) = 300,000, that is, profit of 300,000

Table showing the performance of Model 3:

		TRUE OUTCOME		
		OUTCOME=1	OUTCOME=0	TOTAL
PREDICTED OUTCOME	OUTCOME=1	30	80	110
	OUTCOME=0	20	70	90
	TOTAL	50	150	200

Profit(Rs.): 30(9000) + 20(-9000) + 80(-1000) + 70(0) = 10,000

Table showing the performance of Model 4:

		TRUE OUTCOME		
		OUTCOME=1	OUTCOME=0	TOTAL
PREDICTED OUTCOME	OUTCOME=1	50	0	50
	OUTCOME=0	0	150	150
	TOTAL	50	150	200

Profit (Rs.): 50(9000) + 0(-9000) + 0(-1000) +150(0) = 4,50,000

Consequently, if the prediction of a model is applied on the dataset of 200 records from the past, a profit up to Rs. 4,50,000 is possible! The higher this computed profit the better the model is likely to be for the company.

This is how any model will be evaluated.

The Assignment:

The dataset of 1000 records and the Data dictionary that explains all 76 attributes in a record have been uploaded.

You should try two of classification methods discussed in class, namely, Decision Trees, and Artificial Neural Network, and thus build two classification models. For each method, try to make it as good as you can. Then present a comparison, both pros and cons, of the quality of the models developed as applied to the given data set and recommend the model, which you considered the best among the two you developed.

You can use XLMiner for the project.

Guidelines:

Step 1. Eliminate attributes, which you think will not make a positive contribution to the model. You can use your own qualitative arguments (some gut-feeling based subjective judgments would be fine at this stage provided you have reasons to convince us why you felt that way) in deciding which attributes you will retain to build the model. It's advisable that you try out other methods such as data visualization, regression analysis, principal component analysis etc.

Step 2. Then partition the 1000 strong dataset into three parts as mentioned below:

(a) Training set (of say, 650 records randomly selected),
(b) Validation set (of say, 200 records randomly selected), and
(c) Test set (the remaining 150 records).

Step 3. Develop your classification model using your Training set.

Step 4. Iteratively improve the model by fine tuning relevant model parameters after observing the performance of the model on Training set and Validation set but do not ever look into let alone use the Test set at this iterative stage. "Performance" refers to the computed profit on the data set based on the misclassification matrix of your model output and how the company would have computed the profit on it.

Step 5. Once you finalize the model, apply it only once on the Test set (but, do not change the model after that). Note the accuracy measure on the test set as the percentage of profit obtained with respect to the maximum obtainable profit (that is, if all records were classified correctly).

Step 6. Repeat Steps 3 to 5 using each of the two classification methods.

Submission format and Evaluation:

Evaluation of your assignment for grading purpose would have the following two components as mentioned in (a) and (b) below. Both files should be compressed as a single zipped file is to be mailed to **apds.mldm@gmail.com**. Name of this zipped file should be "***MLDM_II_Name.ZIP***"

(a) A 3-page report in pdf format containing the final attributes of the dataset you retained for modeling, the modeling method you chose, and the parameters of the model (attributes retained for modeling etc) with brief arguments in favour of your decisions on such matters.

The report must contain the misclassification matrix of each of the two models on your Training set, Validation set, and the Test set, respectively. In addition, you must report the profit computed on the Test set.

Comment on the method you considered the best.

The name of the report file should be "***Report_II_Name.pdf***. The report must contain your name and email address.

(b) Submit your models (include the Excel files; compress together as a single zipped file) naming it as "***Model_II_Name.ZIP***. Archive these two files into "***MLDM_II_Name.ZIP"*** and send it the email address **apds.mldm@gmail.com.**

(c) It's an individual assignment.

(d) It's a practical problem. In case of any doubt or lack of clarity in the problem statement or data dictionary, make assumptions that you consider reasonable and go ahead. Please be aware beyond this dataset and the data dictionary nothing will be supplied to you. That's how the practical world is – don't expect there will be somebody to answer questions regarding any doubt you may have.