**Background:**

A national veterans’ organization wishes to develop a predictive model to improve the cost-effectiveness of their direct marketing campaign. The organization, with its in-house database of over 13 million donors, is one of the largest direct-mail fundraisers in the United States. According to their recent mailing records, the overall response rate is 5.1%. Out of those who responded (donated), the average donation is $ 13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $ 0.68 to produce and send. Using these facts, sample of this dataset is taken to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling is used, under-representing the non-responders so that the sample has equal numbers of donors and non-donors.

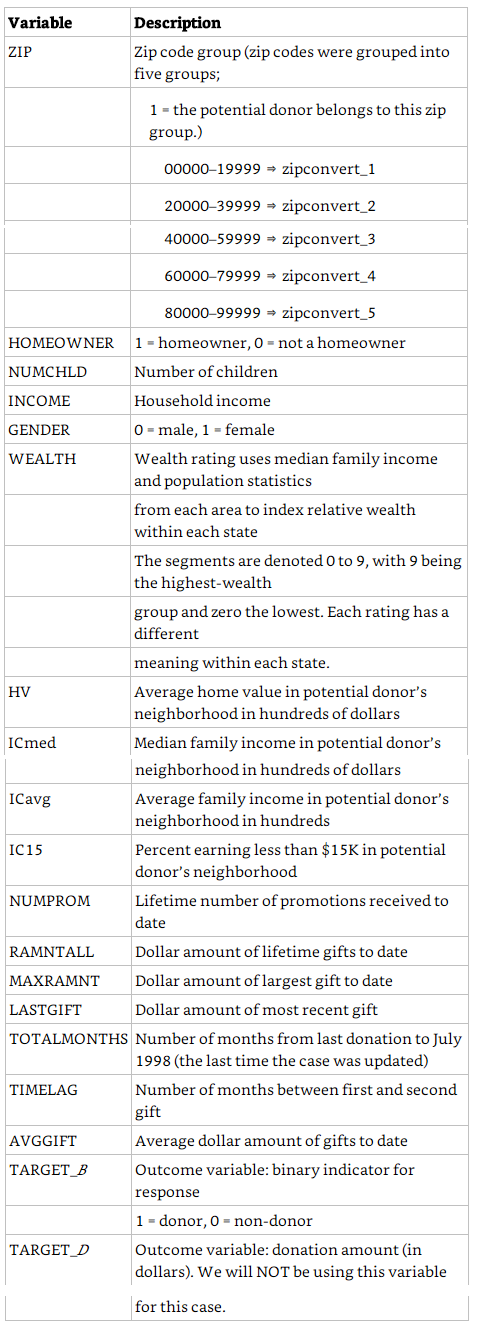
**Data sets:**

* [**Fundraising.csv**](https://www.dropbox.com/s/oujgo0kwhsxho9q/Fundraising.csv?dl=1) (used for model building)
* [**FutureFundraising.csv**](https://www.dropbox.com/s/ls5fooq94rcecqp/FutureFundraising.csv?dl=1) (used for testing)

**Data**

The file Fundraising.csv contains 3120 records with 50% donors (TARGET\_B = 1) and 50% non-donors (TARGET\_B = 0). The amount of donation (TARGET\_D) is also included but is not used in this case. The descriptions for the 22 variables (including two target variables) are listed

**Description of Variables for the Fundraising Dataset**

****

**Method**

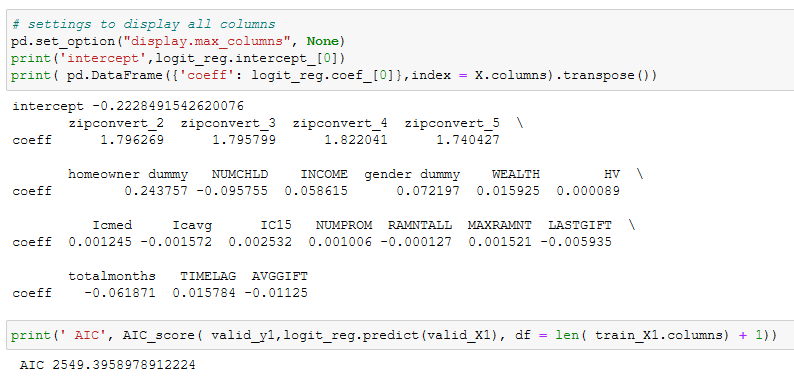
**Step 1: Data preparation:** Partitioned the dataset into 60% training and 40% validation (use random\_state=1).

**Step 2: Model Building:** Steps to build, evaluate, and choose a model.

1. *Select ion of classification tool and parameters.* TARGET\_D is not chosen in the analysis.

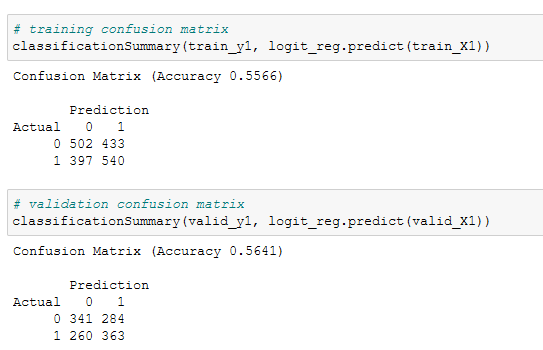
* The 2 classification method models chosen are logistic Regression and Neural Nets
* Reason for choosing :
  + Logistic regression is that it is easier to implement, interpret, and very efficient to train and is less inclined to over-fitting.
  + Neural Nets is because it can handle large amount of data sets, has the ability to implicitly detect complex nonlinear relationships between dependent and independent variables and can detect all possible interactions between predictor variables.
* The predictor variables chosen are
  + ['zipconvert\_2','zipconvert\_3','zipconvert\_4','zipconvert\_5','homeowner dummy','NUMCHLD','INCOME','gender dummy','WEALTH', 'HV','Icmed','Icavg','IC15','NUMPROM','RAMNTALL','MAXRAMNT','LASTGIFT','totalmonths','TIMELAG','AVGGIFT']

**Output results for the Logistic regression Method**



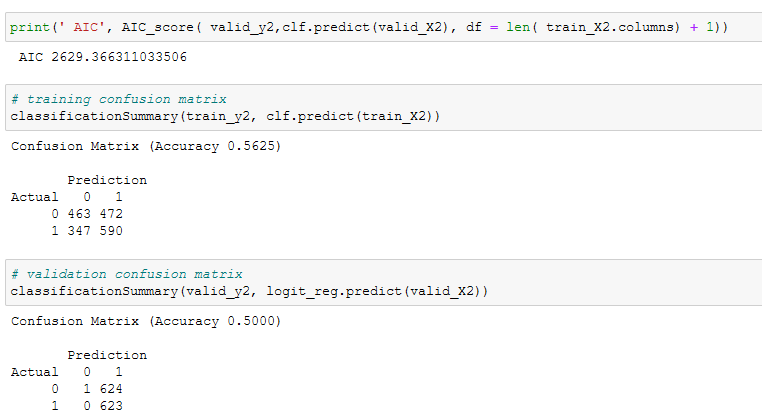
The positive coefficients for the dummy variables 'zipconvert\_2','zipconvert\_3','zipconvert\_4','zipconvert\_5','homeowner dummy','gender dummy' having any of these predictors met (all marked by 1 in the dummy variables) are associated with higher probabilities of being a doner

For the rest of the continuous predictors, positive coefficients in INCOME, WEALTH, HV, Icmed,Ic15,Numprom,Maxramnt, timelag indicate that a higher value on these predictors is associated with a higher probability of being a doner. Similarly, negative coefficients in NUMCHLD, Icavg, RAMNTALL, LASTGIFT, totalmonths indicate that a higher value on that predictor is associated with a lower probability of auction being competitive



The predictions are 56.41 % accurate while the training set is 55.66% . The model is thus not doing a great job!

**Output results for the Neural Net Method**



The predictions are 50 % accurate while the training set is 56.25% . The model is thus not doing a great job either!

(2) *Classification under asymmetric response and cost.*

* The Donor and Non donor classes are present in very unequal proportions with the donor class only forming 5.1% of the population. Weighted sampling is used, to under-represent the non-responders so that the sample has equal numbers of donors and non-donors. By boosting the proportion with the donors. it will improve the performance of the classifiers chosen as we are comparing equal sizes. A simple random sampling may produce too few of the donor class to yield useful information about what distinguishes them from the non donor class and can also result in bad/untrust-worthy results.

1. *Calculating net profit.* To calculate estimated net profit, we will need to undo the effects of the weighted sampling and calculate the net profit that would reflect the actual response distribution of 5.1% donors and 94.9% non-donors. To do this, each row's net profit is divided by the oversampling weights applicable to the actual status of that row. The oversampling weight for actual donors is 50%/5.1% = 9.8. The oversampling weight for actual non-donors is 50%/94.9% = 0.53.

* The expected donation, given that they are donors, is $13.00, and the total cost of each mailing is $0.68
* The net profit is thus $13.00 - $0.68 = $12.32
* The net cost = -$0.68
* Donor which is 5.1% has become 50%. Oversampling weight for actual donors is 50%/5.1% = 9.8
* Non-donors whis is 94.9% has become 50%. Oversampling weight for actual non-donors is 50%/94.9% = 0.53

**Logistic Regression**

**Training Confusion matrix, reweighted**

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| Actual 0 | 502/0.53 = | 433/0.53 |
| Actual 1 | 397/9.8 | 540/9.8 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Logistic Training - Weighted Confusion Matrix** | | | |
|  | Predicted 0 | Predicted 1 | Total |
| Actual 0 | 947 | 817 | 1764 |
| Actual 1 | 41 | 55 | 96 |
| Total | 988 | 872 | 1860 |

Cumulative gains of net profit is calculated based on all those predicted being mailed and all those predicted 0 not being mailed.

**The cumulative gains of net profit for training set= $12.32 (55) -$0.68 (817) = $** **122.04**

**Validation Confusion matrix, reweighted**

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| Actual 0 | 341/0.53 = | 284/0.53 |
| Actual 1 | 2607/9.8 | 363/9.8 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Logistic Validation - Weighted Confusion Matrix** | | | |
|  | Predicted 0 | Predicted 1 | Total |
| Actual 0 | 643 | 536 | 1179 |
| Actual 1 | 27 | 37 | 64 |
| Total | 670 | 573 | 1243 |

The adjusted misclassification rate is (27 + 536)/ 1243 = 45.3%. The model ends up classifying (536 + 37)/ 1243 = 46.10% of the records as 1’ s, when we assume 5.1% responders.

**The cumulative gains of net profit for validation set= $12.32 (37) -$0.68 (536) = $** **91.36**

**Neural nets**

**Training Confusion matrix, reweighted**

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| Actual 0 | 463/0.53 = | 472/0.53 |
| Actual 1 | 347/9.8 | 590/9.8 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Neural Training - Weighted Confusion Matrix** | | | |
|  | Predicted 0 | Predicted 1 | Total |
| Actual 0 | 874 | 891 | 1765 |
| Actual 1 | 35 | 60 | 95 |
| Total | 909 | 951 | 1860 |

**The cumulative gains of net profit for training set= $12.32 (60) -$0.68 (891) = $** **133.32**

**Validation Confusion matrix, reweighted**

|  |  |  |
| --- | --- | --- |
|  | Predicted 0 | Predicted 1 |
| Actual 0 | 1/0.53 = | 624/0.53 |
| Actual 1 | 0/9.8 | 623/9.8 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Neural Validation - Weighted Confusion Matrix** | | | |
|  | Predicted 0 | Predicted 1 | Total |
| Actual 0 | 2 | 1177 | 1179 |
| Actual 1 | 0 | 64 | 64 |
| Total | 2 | 1241 | 1243 |

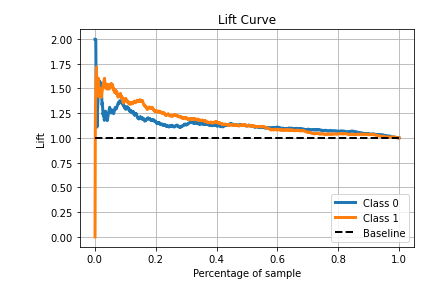
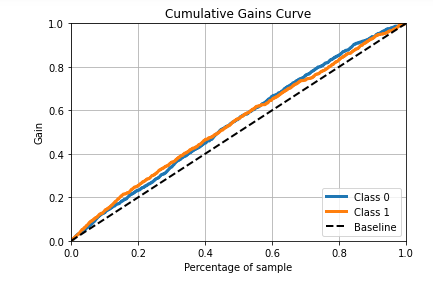
The adjusted misclassification rate is (1177 + 0)/ 1243 = 94.7%. The model ends up classifying (1177 + 64)/ 1243 = 99.8% of the records as 1’ s, when we assume 5.1% responders.

**The cumulative gains of net profit for training set= $12.32 (64) -$0.68 (1177) = $** **-11.88**

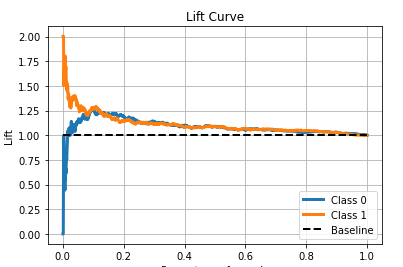
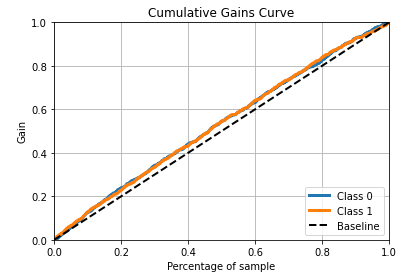
(4) *Cumulative gains curves.*

**Refer to Jupiter Notebook**

**Logistic Regression**

****

**Neural nets**

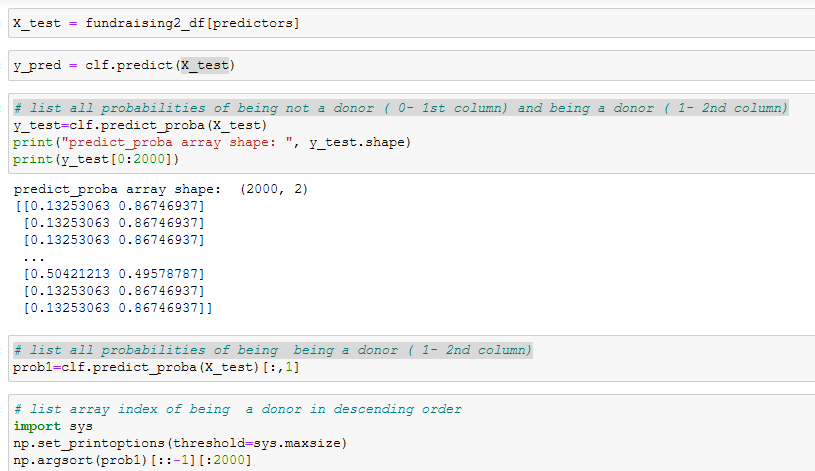
****

**The performance of the class 1 is almost twice as better as class 0 in the logistic Regression model. The neural net model performance to Class 1 and Class 0 are almost the same. It wrongly classifies all class 0 as class 1 , causing the company to send mails and incur a loss in the process.**

(5) *Best model*

**The Logistic regression is definitely a better model than Neural nets achieving a net profit target of $ 122.04 in the validation set**

**Step 3: Testing.** The file *FutureFundraising.csv* contains the attributes for future mailing candidates.



**From the logistic regression lift curve, we can observe that the lift for the donors is almost 2 at the beginning and drops to 1.1 at 0.2. So it would be advisable to go upto 0.2x2000 = 400 far down in the list for the mailing campaign.**