**MIS419 Business Data Mining**

**Assignment 3**

**NAME: Fahad Khalid Alofani ID:437104121**

**NAME: Faisal Abdullah Almahmood ID:437104470**

**PVA Fundraising**

**Background**

A national veteran’s organization wishes to develop a data mining 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, we take a sample of this dataset 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**

The file pvaBalancedTrg.csv contains 7824 data points. The sample has been balanced to carry equal proportions of donors and non-donors, i.e. the data has 50% donors (TARGET−B = 1) and 50% non-donors (TARGET−B = 0). The amount of donation (TARGET−D) is also included but is not to be used in this case. The file contains 41 attributes.

**The assignment**

In this assignment, we examine the performance of different modeling techniques –logistic regression, decision trees, and k-nearest neighbor classifiers. Performance should be evaluated based on costs and benefits as given above; we want a model that maximizes profit.

**PART ONE**

**1. Data exploration**

Import the data, and examine the different variables. You are required to go through the following aspects and explain what you did in each:

* Data statistics: distribution of values, mean, standard deviation, range of values, frequencies, modes, etc.

A screenshot of a cell phone

Description automatically generated

NUMCHLD: MIN=1(Count:601), MAX=5(Count:6), AVERAGE=1.507, DEV= 0.789

Before filling out the missing values, we can see that the distribution is rightly skewed (Positive). We noticed that the minimum value is 1 and maximum is 5. So, it is possible that the missing value means that they have no children.

A close up of text on a white background

Description automatically generated

AGE: MIN= 3, MAX=98, AVERAGE=61.992, MODE=76(Count:187), DEV=16.255

Before filling out the missing values, we could say it is somehow normally distribution and there is no obvious skewness. We can replace the missing value with the average.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

GENDER: We noticed that the categories (J/A/C) are few, the sum of J/A/C are 40 so, they only represent 0.005 from the total. Therefore, we moved them to category U.

* Missing values: the techniques you used to handle missing values

NUMCHLD: There are 6876 missing values, based on the statistics above, we assumed that all the missing values are 0s and replaced them.

AGE: There are 1897 missing values, there are different ways to handle missing values, but we choose to replace the missing values with the average.

DOMAIN: There are 218 missing values, we decided to create another category for them and called it MV.

HOMEOWNR: There are 1851 missing values, since we don’t know if the missing value actually own a home. Therefore, we decided to put them all on the Unknown category.

INCOME: There are 1753 missing values, we decide to replace them with the median which is 4.

GENDER: There are 252 missing values, since we don’t know if the missing values are male or female, we decided to put them all on U category.

TIMELAG: There are 672 missing values, we assume that the values are missing because it has been a long time since his/her first gift (donation), so we replace them with the maximum.

WEALTH: There are 3646 missing values, we choose to take the median to replace the missing value which is 5.

NEXTDATE: There are 672 missing value, since the date is written as normal numbers and not as a date. Therefore, we removed the whole attribute.

* Data transformation: changing data from one type to another, rescaling, data mapping, and Principal Components Analysis (PCA); in other words, data reduction

We normalized the data by using range transformation method (MIN=0, MAX=1) to remove scale effect. Then we used PCA with variance threshold=0.8. There were 14 attributes (IC1 to IC14), The result was dropping 11 attributes and only use the first 3 (PC1 to PC3).

* Data visualization: visually examining the relations between variables and the target variable, i.e. scatterplots, histograms, bar charts, pie charts, etc.

1-The relationship between NUMCHILD and the people who donate is a reverse relationship. People who has no children are the most people who donate. The more children you have the less likely you will donate.A screenshot of a cell phone

Description automatically generated

2- INCOME: The people with low income (1-3) will most likely not donate, people with high income (4-7) will most likely donate. So, the higher your income the more likely you will donate.

A screenshot of a cell phone

Description automatically generated

* Variable selection: omitting variables, subsets, and list of final variables to consider for modeling.

There were 41 attributes before we did data transformation, now there’s 22 regular attributes with 2 special attributes. What we did was dropping:

* FISTDATE
* LASTDATE
* MAXRDATE
* MINRDATE
* NEXTDATE

because the dates were written as normal numbers and not as a date. Therefore, we dropped all the attributes.

Also, we did the PCA, there were 14 attributes (IC1 to IC14), The result was dropping 11 attributes and only use the first 3 (PC1 to PC3).

Also, we dropped Target\_D because we calculate it from Target\_B so we don’t need it.

**2. Modeling**

Partitioning - Partition the dataset into 60% training and 40% validation (set the seed to 12345). [A specified seed ensures that we obtain the same random partitioning every time we run it. With no specified seed, the system clock is typically used to set the seed, and a different partitioning can result in different runs].

A screenshot of a cell phone

Description automatically generated

Modeling – create decision tree models to predict whether or not an individual is going to donate for the fund raising. Try different models until you reach to one or two models that you think are best. Describe these best models, and state how you decided they are the best.

Model1:

A screenshot of a cell phone

Description automatically generated

Accuracy for testing: 56.68%

Accuracy for training: 57.82%

Model2:

A screenshot of a cell phone

Description automatically generated

Accuracy for testing: 56.52%

Accuracy for training: 60.06%

We decide to choose model 1, because it has higher testing and lower difference.

**PART TWO**

**3. Modeling (continued)**

Consider the following classification techniques on the data:

* Naïve Bayes
* *K*-nearest neighbors
* Logistic Regression
* Neural Network (if covered)
* Others that may apply

Be sure to test different parameter values for each method, as you see suitable. What parameter values do you try for the different techniques, and what do you find to work best? Provide a comparative evaluation of performance of your best models from each technique. Feel free to use the following template table to show performance comparisons:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model# | Technique | Parameters (Configurations) | Training Accuracy | Testing Accuracy | Difference |
| 1 | DT (1) | Criterion: Gine index  Max Depth:60  Confidence:0.005 | 57.82% | 56.68% | 1.14% |
| 2 | DT (2) | Criterion: Info Gain  Max Depth:30  Confidence:0.005 | 60.06% | 56.52% | 3.54% |
| 3 | RF (1) | Criterion: Accuracy  Max Depth:3  Number of trees:100 | 59.16% | 57.41% | 1.75% |
| 4 | RF (2) | Criterion: Info Gain  Max Depth:4  Number of trees:100 | 60.78% | 57.28% | 3.52% |
| 5 | KNN (1) | k= 57 | 58.22% | 56.45% | 1.77% |
| 6 | KNN (2) | k= 70 | 60.63% | 56.23% | 4.4% |
| 7 | NB | - | 54.07% | 52.30% | 1.77% |
| 8 | LR | Solver: Auto | 59.22% | 55.27% | 3.95% |

Best models:

**DT (1)** is better because it has higher testing and lower difference

**RF (1)** is better because it has higher testing and lower difference

**KNN (1)** is better because it has higher testing and lower difference

**NB** only one model

**LR** only one model

**4. Classification under asymmetric response and cost**

What is the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors? Why not use a simple random sample from the original dataset? Given the actual response rate of 5.1%, how will the classification models behave under simple sampling? In this case, is classification accuracy a good performance metric for our purposes of maximizing net profit? If not, how would you determine the best model? Explain your reasoning.

We used weighted sampling to get more accuracy, if we use simple random sample, we will get less accuracy and will ignore rare cases. The classification models under simple random sample will give us inaccurate model because it will ignore the rare cases. For this case (rare cases) the recall is the best performance metric and the accuracy is not a good performance metric, only good when data is balanced.

**5. Calculate Net Profit**

* For each method in Question 2 & 3 (choose the ‘best’ model for each method/technique), calculate the net profit for both the training and validation set based on the actual response rate (5.1%). Again, the expected donation, given that they are donors, is $13.00, and the total cost of each mailing is $0.68. (Hint: 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.)

Weighted Cost= (.949/.5) \* 0.68 = 1.29

Weighted Profit= (.051/.5) \* (13-0.68) = 1.26

Net profit = profit – cost

Net profit DT= 1417.5 – 1181.64 = **235.86**

Net profit RF= 1157.94 – 874.62 = **283.32**

Net profit KNN= 1195.74 – 963.63 = **232.11**

Net profit NB= 538.02 - 457.95 = **80.07**

Net profit LR=1050.84 – 863.01 = **187.83**

* Draw each model’s net profit lift curve for the validation set onto a single graph. Are there any models that dominate?

Yes, Random Forest dominate from point 500 to 1250. Also, it has the highest value at 288.

**6. Best Model**

* Summarize the performance of the ‘best’ model from each method, in terms of Net Profit from predicting donors in the validation dataset; at what cutoff is the best performance obtained? You may use the following table as template to assess the models:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classification Technique | Training Accuracy | Testing Accuracy | Net Profit | Cutoff Value |
| Decision Trees | 57.82% | 56.68% | 235.86 | 0.50466 |
| Random Forest | 59.16% | 57.41% | 283.32 | 0.51688 |
| K-Nearest Neighbor | 58.22% | 56.45% | 232.11 | 0.49139 |
| Naïve Bayes | 54.07% | 52.30% | 80.07 | 0.21371 |
| Logistic | 59.22% | 55.27% | 187.83 | 0.50233 |

* From your answers above, what do you think is the “best” model? (What criteria do you use to determine ‘best’?)

Random forest is the best model because it has the highest testing accuracy and it has the highest net profit. Also, if we want to reduce cost it is most profitable at low cost.

**7. Testing your final model**

The file FutureFundraising.xls contains the attributes for future mailing candidates. Using your “best” model (Q6), apply the model on the data and predict the class (donor or non-donor)? Don’t forget to use the same cutoff you determined in your best model to predict donor/non-donor?

Submit this file (in XLSX or XLS format), with your best model’s predictions (probabilities of being a donor). Make sure that the output file contains the following attributes:

* CONTROLN
* Confidence (1)
* Confidence (0)
* Target\_B (prediction)

[The FutureFundraising.xls data file does not contain values of the target variable – so you cannot really see how your model performs on this data. In evaluating your assignment, the instructor will determine how your ‘best’ model performs on this data. Note that part of each team’s evaluation will be based on how well you model performs relative to other models submitted by other teams.]

Submit the following:

* Your answers to the above questions in Word document
* One Excel document after applying the model in Question 7

**METADATA**

STATE State abbreviation (a nominal/symbolic field)

DOMAIN DOMAIN/Cluster code. A nominal or symbolic field.

could be broken down by bytes as explained below.

1st byte = Urbanicity level of the donor's neighborhood

U=Urban

C=City

S=Suburban

T=Town

R=Rural

2nd byte = Socio-Economic status of the neighborhood

1 = Highest SES

2 = Average SES

3 = Lowest SES (except for Urban communities, where

1 = Highest SES, 2 = Above average SES,

3 = Below average SES, 4 = Lowest SES.)

AGE Overlay Age

0 = missing

HOMEOWNR Home Owner Flag

H = Home owner

U = Unknown

NUMCHLD NUMBER OF CHILDREN

INCOME HOUSEHOLD INCOME

GENDER Gender

M = Male

F = Female

U = Unknown

J = Joint Account, unknown gender

WEALTH1 Wealth Rating

HIT Indicates total number of known times the donor has responded to a mail order offer other than PVA's.

IC1 Median Household Income in hundreds

IC2 Median Family Income in hundreds

IC3 Average Household Income in hundreds

IC4 Average Family Income in hundreds

IC5 Per Capita Income

IC6 Percent Households w/ Income < $15,000

IC7 Percent Households w/ Income $15,000 - $24,999

IC8 Percent Households w/ Income $25,000 - $34,999

IC9 Percent Households w/ Income $35,000 - $49,999

IC10 Percent Households w/ Income $50,000 - $74,999

IC11 Percent Households w/ Income $75,000 - $99,999

IC12 Percent Households w/ Income $100,000 - $124,999

IC13 Percent Households w/ Income $125,000 - $149,999

IC14 Percent Households w/ Income >= $150,000

CARDPROM Lifetime number of card promotions received to date. Card promotions are promotion type FS, GK, TK, SK, NK, XK, UF, UU.

NUMPROM Lifetime number of promotions received to date

RAMNTALL Dollar amount of lifetime gifts to date

NGIFTALL Number of lifetime gifts to date

CARDGIFT Number of lifetime gifts to card promotions to date

MINRAMNT Dollar amount of smallest gift to date

MINRDATE Date associated with the smallest gift to date

MAXRAMNT Dollar amount of largest gift to date

MAXRDATE Date associated with the largest gift to date

LASTGIFT Dollar amount of most recent gift

LASTDATE Date associated with the most recent gift

FISTDATE Date of first gift

NEXTDATE Date of second gift

TIMELAG Number of months between first and second gift

AVGGIFT Average dollar amount of gifts to date

CONTROLN Control number (unique record identifier)

TARGET\_B Target Variable: Binary Indicator for Response to 97NK Mailing

TARGET\_B Target Variable: Donation Amount (in $) associated with the Response to 97NK Mailing