**IDS 572**

**Data Mining for Business**

**Assignment 2**

**Team Members:**

**Harish Visweswaraiya**

**Eugene Livshin**

1. **Data exploration: Import the data, and examine the different variables – distribution of values, mean and std deviation, range of values. What do you observe? What variable transformations do you make (and why)? Perform Principal Components Analysis (PCA) – which variables do you include for PCA (give your reason). The dataset has many variables – some (most?) of them may not be useful for our purpose. Your first task is to explore the data, determine missing values and how you might handle these, which variables you think need not be considered, which should be transformed, etc. This is a major task – and can take time, much more than the modeling step that comes next. You will find below a list of subset of variables that someone found useful. Which variables will you consider for modeling (and why)? Which attributes will you omit from the analyses and why.**

**How do you clean the data, handle missing values? What new attributes/values do you derive? How do you approach data reduction? What methods for data reduction do you try?**

Objective of this assignment is to identify the statistical model that will predict the list of potential donors from a wide dataset (large set of potential dependent variables) for a national veteran’s organization. The given input CSV file ‘pvaBalancedTrg’ contains 481 variables. Out of them, there are 2 dependent (Target) variables that denotes whether the person is a donor (TARGET\_B) & the dollar amount associated with the donation (TARGET\_D). Since, we are predicting only the potential donors in this assignment, we are ignoring TARGET\_D variable. Also, CONTROLN is the identifier variable that uniquely identifies a donor. Hence, we initially have 478 potential dependent variables.

We know that including such huge number of variables into a classification model can be a tedious process and will make the model as complicated as possible. So, the first step is to analyze them and eliminate by intuition, the set of variables that might not be related to predicting the donor. The next step is to perform data exploration tasks such as performing ‘data transformation’ & determine missing values to replace them with appropriate values. The last step is to perform Principal component analysis (PCA) by forming a group of similar variables, so as to reduce the list of dependent variables.

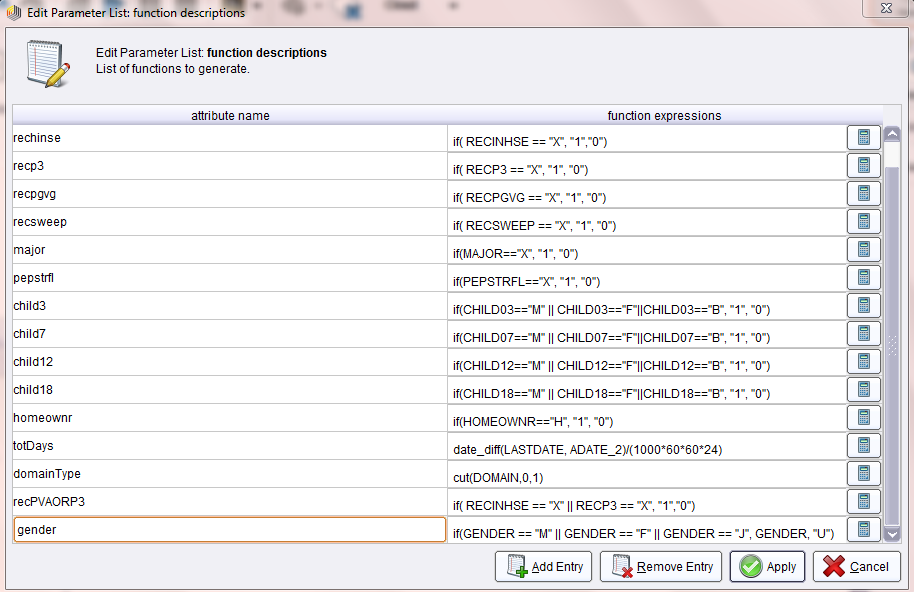
The following are some examples of reasons we used for excluding some variables:

* Exclude variables that have mostly missing values
* Exclude variables that are identifiers (source of the data field, clerical, etc.)
* Exclude variables where existing variable covers need (ex. exclude DOB because AGE variable exists)

The following operators were used for the data cleansing process:

* Generate Attributes – create new attributes by modifying the existing values based on required mathematical operations

For example, we combined In House File Flag & P3 File Flag into a single attribute and named them as recPVAORP3. Also, we typecast character values to binary values. Screenshot of different operations performed using ‘Generate Attribute’ operator is shown in Appendix Section 1.1



* Remove old attributes – Select operator with ‘invert selection’ option has been used to remove old attributes that are no longer required as they have been transformed to new attribute in the above operation
* Remove useless attributes - Select operator with ‘invert selection’ option has been used to remove old attributes that we thought are not useful for this scenario.

|  |  |
| --- | --- |
| **Attribute Name** | **Reason** |
| ODATEDW | Date of donor's first gift will not helpful in predicting the future. Last date can be helpful though |
| OSOURCE | Record source not important for target variable |
| TCODE | Sample dataset does not contain any Title code of interest like Prince, Princess,etc. So, it has been ignored |
| STATE/ZIP/MAILCODE | Doesn't help in predicting donor with geographic location. Might have been useful if it had included whether the donor is from same State/Zip as the organization |
| DOB | Since it is a redundant information with age also available, this has been ignored |
| MDMAUD | Only 1 value for code, might as well use MAJOR field instead |
| CLUSTER | Cluster captured from second DOMAIN byte |
| AGEFLAG | Exact age is not an area of concern for the target variable |
| NUMCHLD | Mostly missing values |
| WEALTH1 | Measures similar thing as income, missing values in 50% of dataset |
| DATASRCE | Again, Data Source is not of interest |
| SOLP3 / SOLIH | Number of solicitation limit not related to target variable |
| WEALTH2 | Values missing for half of the dataset |
| GEOCODE / LIFESRC | This also has something to do with the data source for Census data. Doesn't have a direct impact on target |
| ADI / DMA / MSA | Code field |
| PEC1-2 | Transportation data not relevant |
| TPE1-13 | Transportation data not relevant |
| ADATE\_2-24 | Date variables are not relevant for this model. Date Intervals might be relevant |
| RFA\_2-24 | Donor's recency frequency and amount status can not be used as an attribute to classify the donor |
| MAXADATE | Date variables are not relevant for this model. Date Intervals might be relevant |
| RDATE\_3-24 | Date variables are not relevant for this model. Date Intervals might be relevant |
| RAMNT\_3-24 | Dollar amount of the gift based on promotion not relevant to classify as a donor. Will be important to predict the best promotion but not relevant for this case. Also RAMNTALL captures the summarized information |
| MINRDATE / MAXRDATE / LASTDATE / FISTDATE / NEXTDATE | Date variables are not relevant for this model. Date Intervals might be relevant |
| MDMAUD\_R, F, A | Already have MDMAUD |

We now have removed some attributes as they are not useful for prediction of donor. Next task is to treat the missing values. Following operations are performed to replace the missing values

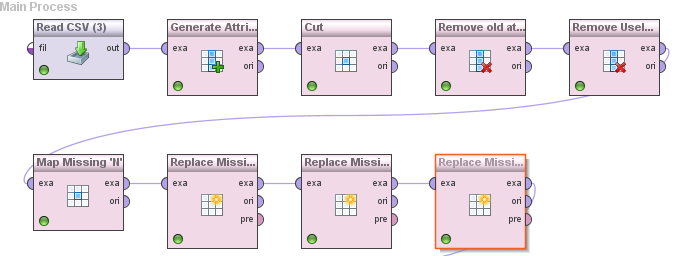
* Map Operator: Maps specified values of selected attributes to new values. This operator can be applied on both numerical and nominal attributes. Few variables have been purposefully left blank in the dataset to denote a ‘No’. They are represented by “?” in rapid miner. We have mapped this value to another value so that it is not counted as a missing value.

For example the “?” in gender variable was mapped to “U” which means Unknown Gender according to the data dictionary.

* Replace missing values: This operator can be used to replace missing values in an attribute. This can be applied to a single, selected subset of attributes or all the attributes.

We replaced all missing values of Numeric variables to -1 with certain exceptions like AGE & INCOME, for which we replaced with average of available data as that would be more sensible. For a character attribute like DOMAIN we replace the missing values by ‘M’ and for domainTYPE, a newly created variable in generated attributes operator we replace the missing value by 0.

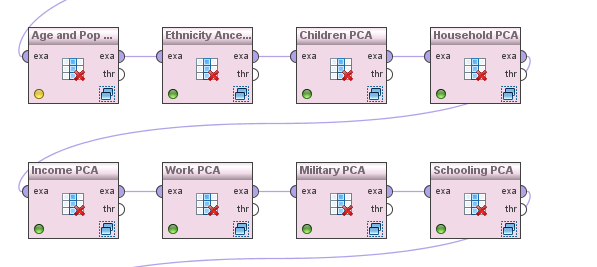
Screenshot of data cleansing & Data transformations performed in rapidminer is shown below:

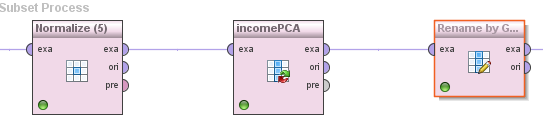


**Principal Components Analysis:**

Even after removing some of the useless attributes we find that our dataset has still got lot of variables, which can render a useless classification model. Hence we performed Principal Component Analysis on selected group of similar variables to further reduce the number of variables that can be used to build the model.

The objective of the PCA is to remove the overlap of information between numerical variables and result in a reduced set of numerical variables that contain most of the information.





**PCA Steps:**

The following operators are used to perform Principal Component Analysis.

* Work on Subset: Base operator required to select a subset (one or more attributes) of input variables and necessary operations are performed in the sub process on the selected subset
* Normalize Operator: This operator is used as the first step of subset to normalize the variable so as to remove the scale effect. After normalization the range of variable is reduced to 0-1.
* PCA Operator: This operator allows to choose the amount of variance the principal components has to explain from the selected subset of data. We have chosen ‘0.8’ for most cases.
* Rename Operator (Rename by Generic Names): After reducing the group of variables to reduced size of principal components we rename the newly generated principal components and remove the left out variables.

We defined 8 PCA on 8 sets of variables to analyze the impact of PCA and reduce the number of variables to improve prediction accuracy for neighborhood attributes. they are listed below:

**Age and Population PCA on Neighborhood attributes:**

There were 19 variables, which reflect demographics characteristics of donor age and population. PC’s are renamed as PCAgePop. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 3 components.

**Income PCA on Neighborhood attributes:**

There were 27 variables, which reflect demographics characteristics of donor income. PC’s are renamed as PCAincome. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 4 components.

**Ethnicity & Ancestry PCA on Neighborhood attributes:**

There were 45 variables, which reflect demographics characteristics of donor ethnicity & Ancestry. PC’s are renamed as PCAethnicity. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 7 components.

**Children PCA on Neighborhood attributes:**

There were 15 variables, which reflect demographics characteristics of donor children. PC’s are renamed as PCAethnicity. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 5 components.

**Household PCA on Neighborhood attributes:**

There were 93 variables, which reflect demographics characteristics of donor household characteristics. PC’s are renamed as PCAethnicity. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 8 components.

**Work & Schooling PCA on Neighborhood attributes:**

There were 46 variables, which reflect demographics characteristics of donor employment characteristics. PC’s are renamed as PCAwork. Similarly, there were 13 variables that reflected their schooling characteristics. They were renamed to PCAschooling. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 8 & 5 components respectively.

**Military PCA on Neighborhood attributes:**

There were 10 variables, which reflect demographics characteristics of donor’s military or Veteran service. PC’s are renamed as PCAmilitary. After performing a PCA with these variables that had a threshold of 80% cumulative variance the variables were reduced to 4 components.

After all these computations and PCA’s, we ended up with 54 regular attributes. It includes 2 special attributes (TARGET\_B & CONTROLN).

**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].**

**Consider the following classification techniques on the data:**

**• decision Trees (you can use J48, or any other suitable type of decision tree)**

**• logistic Regression**

**• k-nearest neighbors**

**• naïve-Bayes**

**• random forests**

**• SVM**

**What parameter values do you try for the different techniques, and what do you find to work best? How do you select the subset of variables to include in the model?**

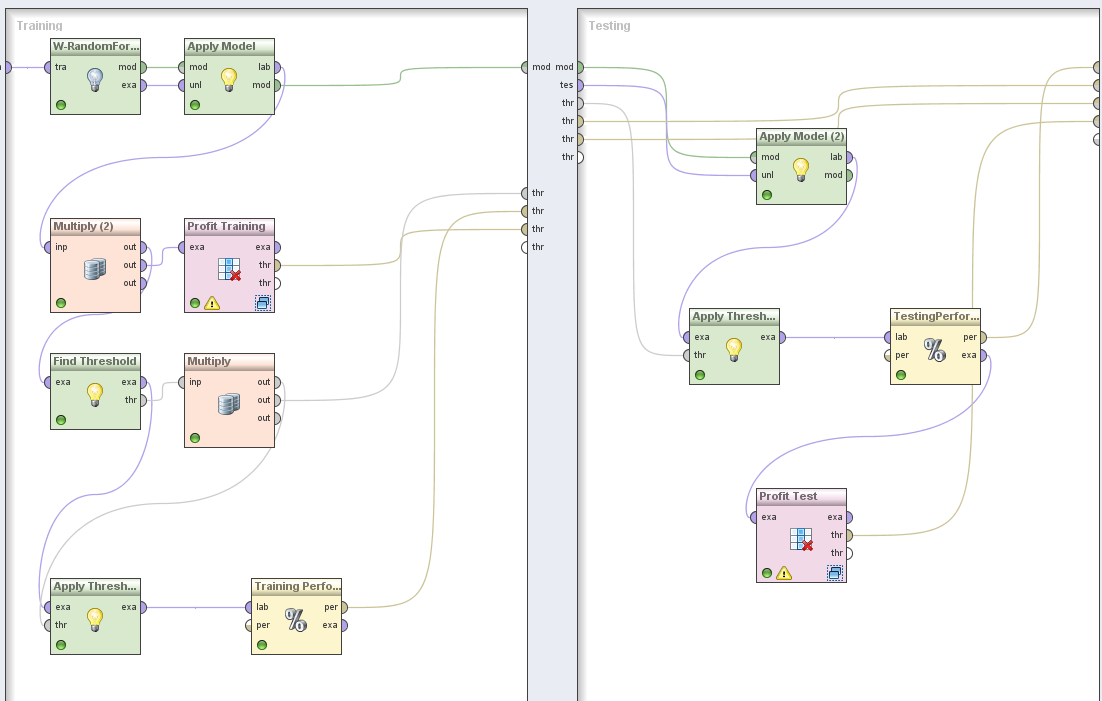
**What methods do you use to select variables that you feel should be included in the model(s)?**

**Does variable selection make a difference for the different models? Do you find PCA to be useful?**

**Provide a comparative evaluation of performance of your best models from each technique**

This section describes the modeling and rapid miner solution process that we went through to find our best model. We used a Split Validation Operator to setup the Training and Testing environment. As stated in the problem we used 60% of the data for training and 40% for validation. We also set the seed to ‘12345’ to ensure the same data set was being used on every run.

A screenshot of the internal rapid miner split validation process is shown below:



Our approach was to use our base set of attributes that we had identified with many different models and parameters. For each model type, when we perceived good results, we used that parameter configuration to test with different combinations of attributes and PCA attributes. A summary table testing results are shown in the companion Excel document.

The following summarizes the different models and accompanying parameters:

**Random Forests:**

**W-Random Forest** operator gave us better performance than Random Forest operator. Please refer to the Appendix Section that shows the performance of the best model in the training and the test data.

**What parameter values do you try for the different techniques, and what do you find to work best?**

|  |  |
| --- | --- |
| **Parameters of model** | **Significance of parameters** |
| I - number of trees in random forest | More the count, better was the testing performance as different set of sample data are considered for average prediction. We found '450' as the optimum number of trees to grow in our model |
| K - Number of features to consider | Setting 'k' to '5' increased both Training & Testing performance |
| S - Seeding | Default value '1' gave us the best model |
| depth | Maximum depth of '15' gave us the best model. |

**Run each method on a chosen subset of the variables - how do you select the subset of variables to include in the model? What methods do you use to select variables that you feel should be included in the model(s)? Does variable selection make a difference for the different models? Do you find PCA to be useful?**

We had a base list of 53 variables, excluding PCA, which was consistent (performance & accuracy) across different models.

**Significance of PCA**

We tried including several PCAs to study its impact and found that, using PCAs (Neighborhood Income, Age&Population, Household) decreased the performance. We tried reducing the depth of trees when we included the PCA components. Though it increased the performance, our best model was the base model.

**Naïve-Bayes Classification:**

Naïve-Bayes operator was used to design a classification model on the final set of variables. Please refer to the Appendix Section that shows the performance of the best model in the training and the test data.

|  |  |
| --- | --- |
| **Parameters of model** | **Significance of parameters** |
| Laplace correction | No difference in performance |

We understood Naïve-Bayes classification is data-dependent.

**Significance of PCA**

Including Income PCA & AgeandPop PCA increased the performance for Naïve-Bayes when compared to its model with final set of base attributes. So, our best model for Naive-Bayes is derived by including Income & AgeandPop PCs to our final set of attributes. It makes sense as Naive-Bayes is data-dependent.

**k-Nearest Neighbours Classification:**

**K-NN** operator was used on our final set of variables. Please refer to the Appendix Section that shows the performance of the best model in the training and the test data.

|  |  |
| --- | --- |
| **Parameters of model** | **Significance of parameters** |
| k - number of nearest neighbors | Increasing the k value from 1 to 5 increased the performance by 75%. Performance degraded for any value more than that. So, 5 was our optimum k value |

**Significance of PCA**

We included Income PCA & AgeandPop PCA individually and together with our final set of attributes. There was no significant improvement in performance because of addition of PCA components. So, our best model with k-NN classification model is built with just final set of attributes.

**Logistic Regression Classification:**

**W-Logistic** operator was used on our final set of attributes. Please refer to the Appendix Section that shows the performance of the best model in the training and the test data.

|  |  |
| --- | --- |
| Parameters of model | Significance of parameters |
| R - Set the ridge in the log-likelihood | Increasing the ridge parameter from default value of 1.0E-8 until 0 gave us consistently best results. Further increase in ridge degraded the performance . |
| M - Number of iterations | Default parameter of -1 was used as there were no significant difference otherwise |

**Significance of PCA**

We included Income PCA & AgeandPop PCA individually and together with our final base set of attributes. Inclusion of PCA components increased the performance of our classification model. Our best model under logistic regression was inclusion of Neighborhood Income PCA with our base set of variables.

**W-J48 (Decision Tree):**

**W-J48** Operator was applied on the base set of final attributes. Please refer to the Appendix Section that shows the performance of the best model in the training and the test data.

|  |  |
| --- | --- |
| **Parameters of model** | **Significance of parameters** |
| C - Confidence threshold for pruning | The is set to 0.25 which is used to classify attributes |
| M- Minimum instances per leaf | A value of 35 was given so as to avoid over fit of the training data. A smaller value of M can hamper accuracy in the testing data. |
| R- Use reduced error pruning | This option is enable to build a better model |
| B | This option is checked to enable binary splits. |

**Significance of PCA**

We included Income PCA & AgeandPop PCA individually and together with our base set of attributes. There was no significant improvement in performance because of addition of PCA components.

**SVM (Support Vector Machine)**

**Support Vector Machine** Operator was applied on the base set of final attributes. Please refer to the Appendix Section that shows the performance of the best model in the training and the test data.

Kernel Type - Several kernel types were tested: dot, radial, polynomial & Gaussian combination. For polynomial kernel type

|  |  |
| --- | --- |
| **Parameters of model** | **Significance of parameters** |
| C - Confidence threshold for pruning | The is set to 0.25 which is used to classify attributes |
| M- Minimum instances per leaf | A value of 35 was given so as to avoid over fit of the training data. A smaller value of M can hamper accuracy in the testing data. |
| R- Use reduced error pruning | This option is enable to build a better model |
| B | This option is checked to enable binary splits. |

**Significance of PCA**

We included Income PCA & AgeandPop PCA individually and together with our base set of attributes. There was no significant improvement in performance because of addition of PCA components.

**3**. **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? (Hint: given the actual response rate of 5.1%, how do you think the classification models will behave under simple sampling)?**

Response of Population dataset is 5.1% whereas in our sample dataset, we intentionally use 40% response and avoided random sampling. This is because ratio of donors is far less when compared to non-donors in the original population, so using the same ratio in the sample will lead to bias and model might provide faulty results. Therefore, we use weighted sampling to produce a training set with equal number of donors & non-donors.

**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.**

As we are using a weighted sampling, classification accuracy is not a good performance metric for our purposes of maximizing net profit. Initially selected weights will affect the metrics that are produced and as our goal is to maximize profit, the best model can be determined based on the cumulative net profit.

We follow the below mentioned process to determine the best model -

Recalculate the new net-profit based on the new cost. Adjusted values are computed as follows:

**Average Donation = $13, Actual Cost = $ 0.68. Therefore, net-profit per donor = $13 - $0.68 = $12.32**

|  |  |  |
| --- | --- | --- |
|  | **Donors** | **Non-donors** |
| **Actual** | 5.10% | 94.90% |
| **Weighted** | 39.47%(3912) | 60.53%(6000) |
| **Adjusted classification costs** | 12.32 \* (5.1/39.47) = $1.592 | -0.68 \* (94.9/60.53) = -$1.067 |

Thus the adjusted values will be used to compute cumulative maximum profit.

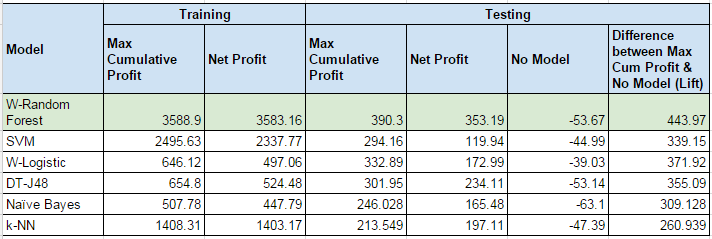
Similarly, the other option is to recompute the weighted confusion matrix and adjust the classification.

**4. Calculate Net Profit: For each method in Question 2 (choose the ‘best’ model for each method/technique, either with the full or reduced set of variables), calculate the lift of 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.)**

Maximum Cumulative Profit and Net Profit has been calculated for training and testing data set of each of the best model. Comparison chart between different models is shown below. **Opportunity Cost has not been considered for calculation of Net Profit**.

Also, we considered the corresponding ‘No Model’ values for the Max. cumulative profit values in the testing data for the respective models to calculate the Lift.

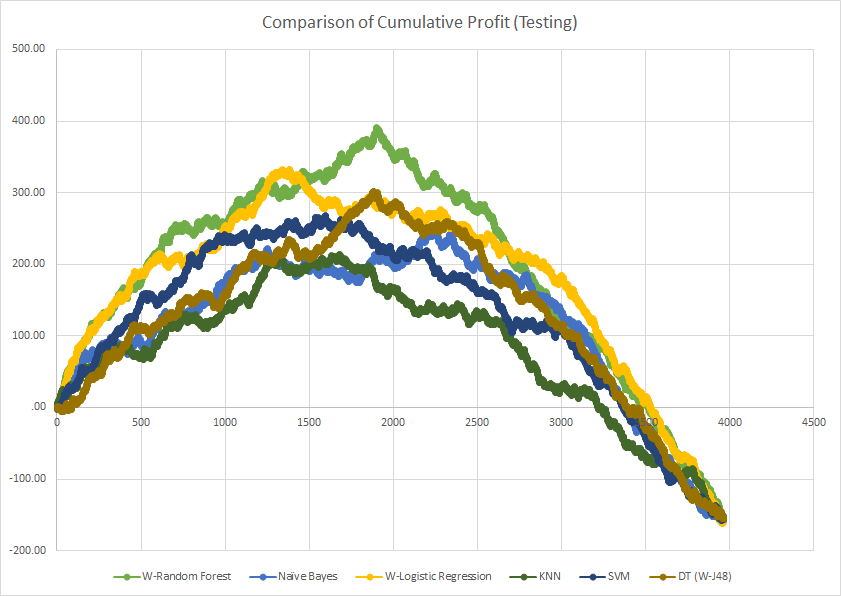
Summary table of the comparison between different models is shown below:



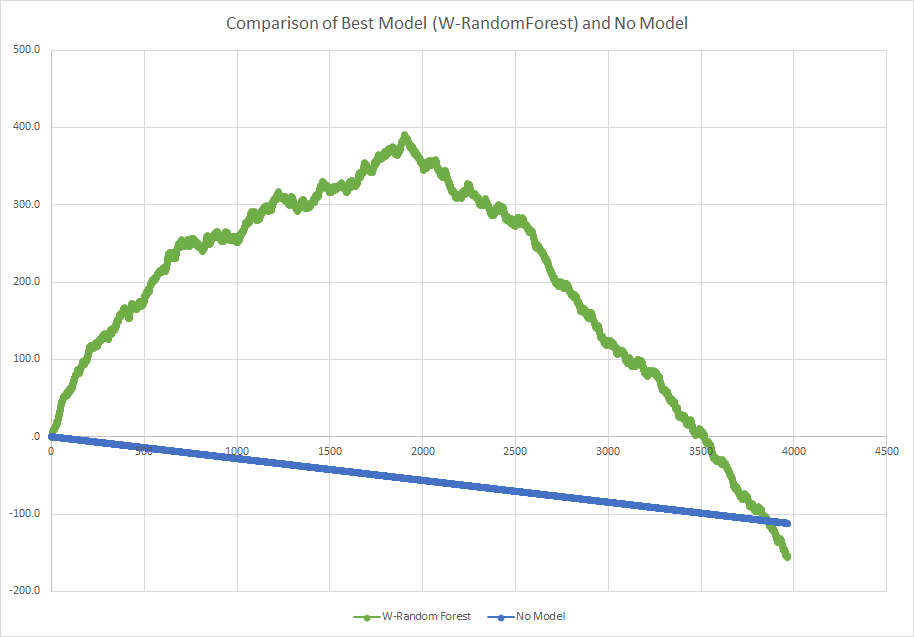
It could be seen that we obtained the maximum lift from W-Random Forest Classification Model for the Testing(Unseen) data.

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

Net profit lift curves for the validation set were plotted for the best classification models. We found out that W-Random forest clearly dominated with maximum cumulative profit of 390.3 as shown in the graph below:



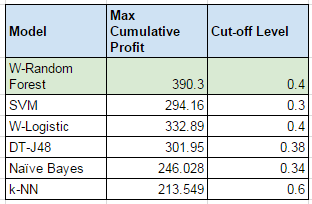
When compared with the no model for the validation set, we obtained a max. Lift of 443.97 for W-Random Forest.



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

Our criteria for the ‘best’ model is based on maximum cumulative profit.

The best performing models for each method is mentioned below and their respective cut-off points are also mentioned with them.



From the above table, we chose ‘W-Random Forest’ as the best model as it corresponded to the max. Cumulative profit value. Also, we obtained the max profit at ‘1902’ data point. Thus, our best model specifies to target 47.9 % (1902/3965 = 0.479) of individuals from the unseen data. We obtained a cut-off value of **0.4** for the best model.

**7. The file FutureFundraising.xls contains the attributes for future mailing candidates. Using your “best” model from Step 2 (#5), which of these candidates do you predict as donors and non-donors? List them in descending order of probability of being a donor. What cutoff do you use to predict donor/non-donor?**

Now that we have identified the best classification model to predict donors & non-donors, we use it to predict the donors from the unseen data(Validation) - FutureFundraising.xls

Please refer to the attached excel sheet

We have sorted the data in ‘descending’ order of confidence(1) of predicting as donor and non-donors. Our classification model predicted **2456 donors out of 20000** future mailing candidates. We used a cut-off value of **0.4** to predict donor/non-donors.

**Appendix**



**Section 1.1**

