

ADTA 5230

Final Project Report

Sathvik Chava

Sumanjali Banjara

Paulina Gomez Ibarra

May 5, 2025

1. Introduction/Business Understanding/Analytics Questions

The data we will be working with for this report is from a non-profit organization. They shared with us their goal of improving the cost-effectiveness of their direct marketing campaigns to previous donors and tasked us with creating predictive models to help them get to their desired outcomes. As we became familiar with this data set, we realized not only could we use it to help the non-profit organization with its cost-efficient goals, but it also provides many business opportunities to be assessed.

Based on their records, their average response rate is about 10% and out of those, they get an average donation of \$14.50 per donor. Their mailing costs are \$2.00 in production and sending, but they realized that the cost-effectiveness is low, as direct profits are at a loss of $-\$0.55$ per mailing. Due to this problem, we decided that we must create both classification and regression models to identify the best fitting models that would improve cost effectiveness for the non-profit organization's goals.

Given the information provided and the dataset on hand, we developed the following business questions to help guide our data analysis for both classification and regression models.

- **Classification:** What individuals are most likely to donate to the non-profit?
- **Regression/Prediction:** How much will each donor likely donate?

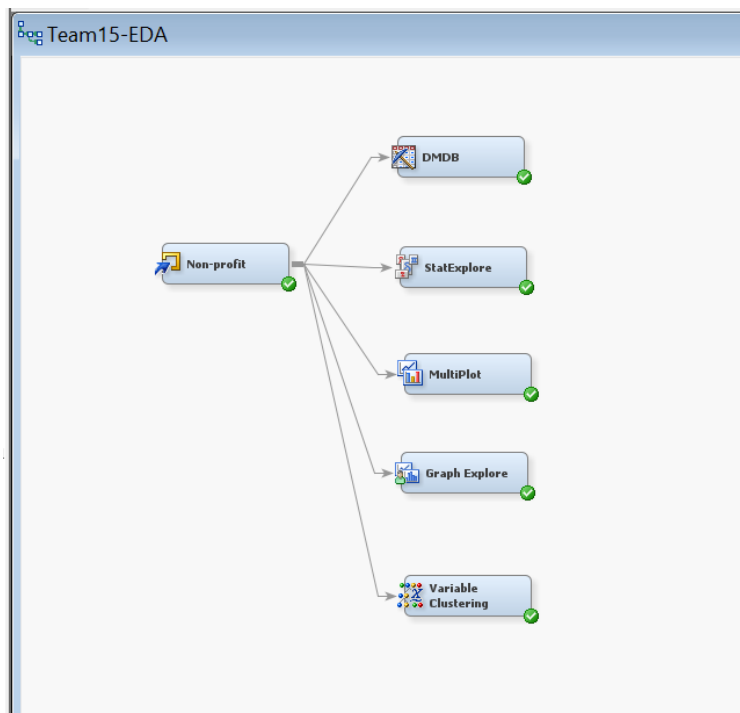
By answering these questions, the non-profit organization can focus its direct marketing strategies on individuals who are likely to donate and avoid the loss of mailing costs from non-donors. This would help maximize the returns of costs and reduce them to be used only when necessary. Furthermore, with the prediction model, estimating the approximate contribution from each donor based on previous data will allow the non-profit organization to know where to target their marketing efforts and maximize overall profits.

2. Data Understanding/EDA:

To understand the dataset and prepare it for further modelling, we conducted a comprehensive EDA using SAS enterprise Miner. Our goal was to identify trends, variable importance and data issues that could impact model performance. We have 2 key targets:

- **donr:** This variable represents the individual who responded positively to a past donation request. It is a binary variable where 1 represents an individual who donates and 0 represents the non-donor.
- **damt:** This variable represents the dollar amount donated by an individual. The average donation is about \$14.45, with a minimum of \$8 and a maximum of \$27. The distribution is a bit right skewed with relatively low variance, showing consistency in donation among the positive responders.

In terms of data exploration, we used multiple SA nodes – DMDB, StatExplore, Graph Explore, Variable Clustering and Multiplot to explore and visualize the data.



2.1 Descriptive Statistics - DMDB and StatExplore Node Results:

Interval Variable Summary Statistics

Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
damt	damt	0	6002	0.00	27.00	7.209	7.3612	0.11698	-1.8276
gifa	gifa	0	6002	1.89	72.27	11.678	6.5281	1.74146	5.9115
gifdol	gifdol	0	6002	23.00	1974.00	115.800	86.5380	6.09138	92.7570
gifl	gifl	0	6002	3.00	642.00	22.981	29.3964	7.18035	94.3679
gifr	gifr	0	6002	1.00	173.00	15.654	12.4246	2.67345	15.1894
hv	hv	0	6002	51.00	710.00	183.905	72.7705	1.48890	4.2074
incavg	incavg	0	6002	14.00	287.00	56.789	24.8335	1.85779	7.0285
incmed	incmed	0	6002	3.00	287.00	43.949	24.6644	2.00492	8.2734
kids	kids	0	6002	0.00	5.00	1.584	1.4125	0.39406	-0.7801
lag	lag	0	6002	1.00	34.00	6.319	3.6414	2.41056	8.4347
low	low	0	6002	0.00	87.00	13.885	13.1046	1.35139	1.8415
mdon	mdon	0	6002	5.00	40.00	18.789	5.5963	1.11760	2.3876
npro	npro	0	6002	2.00	164.00	61.354	30.3052	0.28319	-0.6321

Class Variable Summary Statistics

Variable	Label	Type	Number of Levels	Missing
donr	donr	N	2	0
inc	inc	N	7	0
ownd	ownd	N	2	0
region	region	C	5	0
sex	sex	N	2	0
wlth	wlth	N	10	0

As seen on the images above, the dataset contains 6002 records, and there are no missing values across variables. The variable *donr* is evenly split, making it easy for classification modeling. Interval variables such as *gifdol*, *gifl* and *gifr* are heavily skewed, which means we may need to apply transformation before modeling. Chi-Square analysis, showing *inc*, *region*, *ownd*, and *wlth* have strong relationships with *donr*, as the p-value is less than 0.0001. Variable worth plots, identifying *kids*, *incavg*, *icmed* and *hv* as top contributors for predicting *damt* and *sex* was found to be insignificant for donation.

2.2 Class Distribution – Graph Explore

Distribution of Class Target and Segment Variables
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Level	Frequency Count	Percent
TRAIN	donr	TARGET	0	3008	50.1166
TRAIN	donr	TARGET	1	2994	49.8834



The donor class (*donr*) is balanced with 50.1166% non-donors and 49.8834% donors.

Donation amount (*damt*) is skewed with majority donor contributing between \$10 and \$15.

2.3 Correlation and redundancy analysis

To detect multicollinearity, we used the variable clustering node which produces heatmap for correlation between variables. We found strong positive correlations were found among *gifa*, *gifdol*, *gifl* and *gifr*. A strong negative correlation was seen between *kids* and *damt* ($r = -0.55$), indicating that households with more children tend to donate less.



3. Data Preparation

The goal of data preparation was to make sure that the dataset was clean, standardized and optimized for both classification and regression modeling.

As they were no missing values found, the imputation was not necessary. Categorical variables were properly encoded using SAS DMDB to make sure the models could be interpreted correctly.

Binary variables were in 0/1 format which worked for classification models we planned to use.

We made data 2 parts- 70% of data for training and 30% for validation. As the data is relatively small, it is important to carefully divide the data especially when using complex models like neural networks and gradient boosting. That's why we chose a 70% training split – to make sure the models have enough data to learn and still leaving enough for validation to check their performance.

We used the transform node to standardize the variables to bring all data variables to the same scale which makes it easy for modeling and to avoid problems over fitting etc.

We selected the donr as target variable for classification and dmat as target variable for regression.

We rejected the dmat and id in classification and dnor and id in regression model.

4. Modeling

4.1 Classification Modeling

We have implemented several classification models to predict whether a prior donor would respond to mail. The prediction that the model is supposed to make is a binary classification, DONR=1 for donor and DNOR=0 for non-donor. Each model that we chose has a unique strengths and applicability.

- **Logistic Regression:** We have used 2 regression models, one with default settings and one with selection model set to “step wise”. We chose step wise because it enhances interpretability and generalization to unseen data.
- **Neural Networks:** We have tested 3 Neural Networks, model 1 is with all default settings, model 2 is with back propagation and 3 hidden units, and model 3 is with back propagation and 6 hidden units.
- **Gradient Boosting:** We have chosen to use default setting gradient boosting, as it captures non linearities.

4.2 Classification Model Result Analysis and Interpretation

The overall best model is Neural Network with default settings. It has the lowest validation misclassification rate of 0.10024, and lowest validation Average Squared Error of 0.068904. This model has the best balance between accuracy and calibration among all the models. While the neural network stands first, the second-best model is Regression with Stepwise. The Standard Regression also performs very close to the Stepwise Regression. Back propagation neural networks performed the worst, showing that increasing model complexity did not improve the accuracy.

4.2.1 Fit Statistics of Classification Models

Enterprise Miner - Project Group 15

Results - Node: Model Comparison Diagram: Classification

File Edit View Window

Output

148	Fit Statistics						
149	Model Selection based on Valid: Misclassification Rate (_VMI3C_)						
150							
151							
152							
153				Valid:	Train:	Train:	Valid:
154	Selected	Model	Model Description	Misclassification	Average	Misclassification	Average
155	Model	Node		Rate	Squared	Rate	Squared
156					Error		Error
157	Y	Neural	Neural Network	0.09212	0.07174	0.10024	0.068904
158		Reg2	Regression StepWise	0.10488	0.08092	0.11262	0.077365
159		Reg	Regression	0.10544	0.08069	0.11167	0.077526
160		Boost	Gradient Boosting	0.10877	0.09757	0.11690	0.094446
161		Neural2	Neural Network BP HU3	0.13041	0.09919	0.13976	0.094658
162		Neural3	Neural Network BP HU6	0.13152	0.10083	0.14500	0.095128
163							
164							
165							
166							

Based on the bias-variance, Neural Network shows good and stable performance across training and validation sets, this indicates good generalization.

4.3 Prediction Modeling

We have implemented several models for the task of predicting the “damt” variable which is donation amount for the donor. Each model was considered due to their unique strengths.

- **HP Forest:** The ensemble decision trees are designed to improve accuracy,
- **Neural Network:** We used a neural network with default settings and another with 6 hidden units.
- **Auto Neural:** We used two auto neural models with hidden units two and six.

4.4 Prediction Model Result Analysis and Interpretation

The Auto Neural model with 6 hidden units has the highest accuracy, it’s validation average square error is 16.9058. The Neural network can be used it deploy as it is accurate.

4.4.1 Fit Statistics of Predictive Models

Enterprise Miner - Project Group 15

Results - Node: Model Comparison Diagram: Prediction

File Edit View Window

Output

28	Fit Statistics					
29	Model Selection based on Valid: Average Squared Error (_VASE_)					
30				Valid:	Train:	
31				Average	Average	Train:
32				Squared	Squared	Misclassification
33	Selected			Error	Error	Rate
34	Model	Model Node	Model Description			
35						
36						
37	Y	AutoNeural2	AutoNeural hu=6	16.9058	16.3956	.
38		Neural	Neural Network	17.8964	18.0659	.
39		AutoNeural	AutoNeural hu=2	20.4680	20.5136	.
40		Neural2	Neural Network bp hu=6	20.5105	21.1285	.
41		HPDMForest	HP Forest	20.6516	19.0775	.
42						

5. Evaluation

This chapter evaluates models on two bases, statistics and profitability. Profitability can be defined as the net gain that can be obtained using this model. All evaluations in this chapter are based on validation data to ensure unbiased assessment.

5.1 Best Classification model based statistical metrics

The best model by misclassification rate is Neural Network with default settings. It has the lowest total errors. The below table is a ranking the best model bases on the validation misclassification.

Model Name	Validation Misclassification Rate
Neural Network (default)	0.0921
Stepwise Regression	0.1049
Regression	0.1054
Gradient Boosting	0.1088
Neural Net BP (3 Hidden Units)	0.1304
Neural Net BP (6 Hidden Units)	0.1315

We can see that the Neural Network with 0.0921 validation misclassification is the lowest making it the best performer in all the models.

5.1 Best Classification Model Based on Maximum Profitability

The best model by profitability is also Neural Network with default settings. The profitability can be calculated by using the formula “Net Profit=(TruePositive×14.50)−((TruePositive + FalsePositive)×2.00)” .

The below table shows the profitability of each model sorted by max profitability first.

Model Name	True Positive	False Positive	Profit (in \$)
Neural Network	830	97	10181.00
Stepwise Regression	815	105	9977.50
Regression	815	106	9975.50
Gradient boost	805	102	9858.50
Neural network BP hu=3	795	131	9675.50
Neural network BP hu=6	794	132	9661.00

We can observe that the best model in statistical metrics and the best model in maximum profitability is the same. The Neural Network is the most profitable model with \$10,181 profit.

5.3 Evolution of Predictive Models

The performance of predictive models can be assessed using the validation average squared error (ASE).

Model Description	Validation ASE
AutoNeural (6 Hidden Units)	16.91
Neural Network (default)	17.9
AutoNeural (2 Hidden Units)	20.47
Neural Net BP (6 Hidden Units)	20.51
HP Forest	20.65

We can say that the best predictive model is Auto Neural with 6 hidden units with the lowest ASE. It has a very minimal train-validation gap, which means that the model can generalize well. The runner up model is Neural Network with default settings, It also performs better than remaining models. By evaluating the models, we can conclude that the neural network is the best classification model and the Auto Neural with 6 hidden units is the best predictive model.

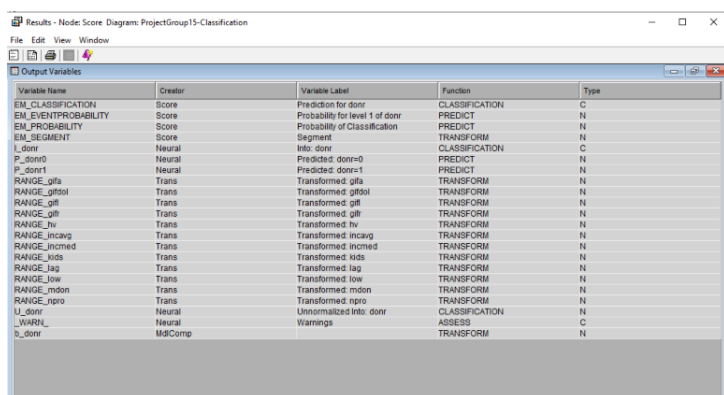
6. Deployment

Once we completed the modeling and evaluation, we updated our score data to our best fitting models for both classification and prediction models to make more accurate real-world decisions on our direct marketing campaigns. With the results from our score data in our models, we could get a better idea on how to answer our questions about what individuals are more likely to donate and the estimated amount of their donation, to better target these individuals who would maximize our profits and reduce our mailing costs.

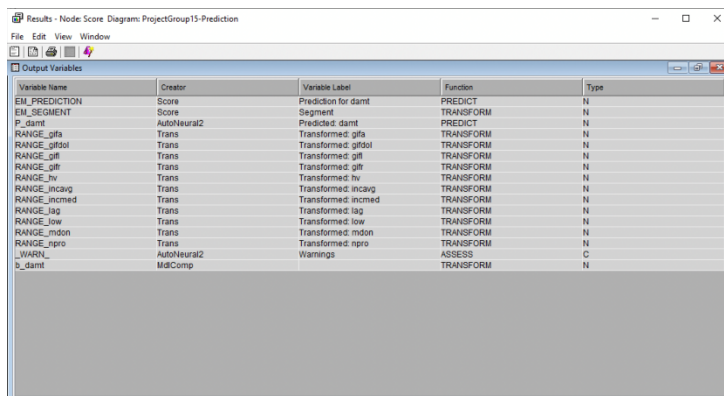
As mentioned above in our evaluation results, the best fitting model out of our classification models was the Neural Network model with default settings, as it had the lowest misclassification

rate. Out of the prediction/regression models, the Auto Neural model with 6 hidden units was the best fitting model, with the lowest average squared error (ASE).

In the deployment of the data, variables *donr* and *damt* were rejected from the analysis so that the score node could generate these variables. For classification model, the variable *p_donr1* and *p_donr0* were generated, with *p_donr1* classifying individuals likely to donate. For the prediction model, the variable *p_damt*, estimates the donation for each individual donor based on the Auto Neural model.



Variable Name	Creator	Variable Label	Function	Type
EM_CLASSIFICATION	Score	Prediction for donor	CLASSIFICATION	C
EM_EVENTPROBABILITY	Score	Probability for level 1 of donor	PREDICT	N
EM_PROBABILITY	Score	Probability of Classification	PREDICT	N
EM_SEGMENT	Score	Segment	TRANSFORM	N
L_donr	Neural	Is donor	CLASSIFICATION	C
P_donr0	Neural	Predicted donor=0	PREDICT	N
P_donr1	Neural	Predicted donor=1	PREDICT	N
RANGE_gfta	Trans	Transformed: gfta	TRANSFORM	N
RANGE_gftol	Trans	Transformed: gftol	TRANSFORM	N
RANGE_gft	Trans	Transformed: gft	TRANSFORM	N
RANGE_gft	Trans	Transformed: gft	TRANSFORM	N
RANGE_hv	Trans	Transformed: hv	TRANSFORM	N
RANGE_incavg	Trans	Transformed: incavg	TRANSFORM	N
RANGE_incmcd	Trans	Transformed: incmcd	TRANSFORM	N
RANGE_kids	Trans	Transformed: kids	TRANSFORM	N
RANGE_lag	Trans	Transformed: lag	TRANSFORM	N
RANGE_low	Trans	Transformed: low	TRANSFORM	N
RANGE_mdon	Trans	Transformed: mdon	TRANSFORM	N
RANGE_npro	Trans	Transformed: npro	TRANSFORM	N
U_donr	Neural	Unnormalized info: donor	CLASSIFICATION	N
_WARN	Neural	Warnings	ASSESS	C
b_donr	MidComp		TRANSFORM	N



Variable Name	Creator	Variable Label	Function	Type
EM_PREDICTION	Score	Prediction for damt	PREDICT	N
EM_SEGMENT	Score	Segment	TRANSFORM	N
P_damt	AutoNeural2	Predicted damt	PREDICT	N
RANGE_gfta	Trans	Transformed: gfta	TRANSFORM	N
RANGE_gftol	Trans	Transformed: gftol	TRANSFORM	N
RANGE_gft	Trans	Transformed: gft	TRANSFORM	N
RANGE_gft	Trans	Transformed: gft	TRANSFORM	N
RANGE_hv	Trans	Transformed: hv	TRANSFORM	N
RANGE_incavg	Trans	Transformed: incavg	TRANSFORM	N
RANGE_incmcd	Trans	Transformed: incmcd	TRANSFORM	N
RANGE_lag	Trans	Transformed: lag	TRANSFORM	N
RANGE_low	Trans	Transformed: low	TRANSFORM	N
RANGE_mdon	Trans	Transformed: mdon	TRANSFORM	N
RANGE_npro	Trans	Transformed: npro	TRANSFORM	N
_WARN	AutoNeural2	Warnings	ASSESS	C
b_damt	MidComp		TRANSFORM	N

The images above show that the variables generated with the score node and nonprofit score data were generated with the best fitting models: in the case of classification (*p_donr1* and *p_donr0*), Neural Network with default settings, and in the case of the prediction (*p_damt*), AutoNeural with 6 hidden units. This shows that the scoring was done correctly. To calculate the expected profit with this model using the variables generated from the score data.

- Expected Profit = $(pdonr1 \times pdamt) - \2.00

This formula can be used for each donor that is likely to donate (the variable is represented by the probability they will donate) and multiply it by the predicted donation amount they will make. These are multiplied and we subtract 2 from the result to account for the mailing costs incurred. Based on the expected profit, we can make the decision whether to send the mailing to the individual or not.

Using the mass marketing approach, we would target 2,007 individuals and mail them asking for a donation. Due to mailing costs incurred for all individuals (including those that do not answer our donation request), we would face higher costs that would bring down our profits. Our total expected profit would be \$401.39, with an average profit of \$0.20 per person. With our direct marketing approach, targeting only qualified individuals who would generate positive profits based on the profit formula, we would be targeting 508 individuals. This targeted marketing approach would generate an expected maximum profit of \$3129.60, with an average profit of \$6.16 per person. This shows the advantage of targeted marketing approaches to reduce costs and increase profits. This targeted strategy should be used for direct marketing purposes to get the highest profits in return.

7. Conclusion

In conclusion, using the nonprofit marketing data we created models for both classification and prediction. In terms of classification, we aimed to classify individuals between those who are likely to donate and those who aren't likely to donate. With our prediction model, we aimed to predict the average amount of donations made by each individual.

Our main goal was to identify the best fitting models that could help us predict these to lower the non-profit organization's costs and achieve the highest profits. This was obtained by many steps, starting with an in-depth EDA to get a better understanding of our variables and factors that could be contributing to those likely to donate compared to those who weren't likely. Then, we prepared our data to ensure that we had no missing data and ensured the appropriate division between train and validation data.

After that, we created three models for classification and three models for prediction with SAS Enterprise Miner, which with the Model Comparison Node, allowed us to identify the model of best fit for each (classification and prediction). With our models of best fit, Neural Network with default settings for Classification and AutoNeural with 6 hidden units for prediction, we uploaded the score data and updated the models with this data to unveil the best strategies for profit maximization.

The deployment of the score data allowed us to gain a better understanding of the differences in a mass marketing approach compared to a targeted marketing approach. The targeted marketing approach showed approximately 8 times more profit than the mass marketing approach. By targeting those individuals that with the estimated profit formula gave us a positive profit, we could ensure that the nonprofit organization wouldn't waste their mailing costs and could have a higher response rate with donations.

Ultimately, the nonprofit organization benefits from our models of best fit and can use these to estimate profits from everyone. With this information, they can create a targeted direct marketing strategy, where they will target fewer individuals, reduce mailing costs, and get a higher response rate, obtaining higher profits from the donations. This will increase their cost effectiveness and fulfill their strategic objectives, improving their direct marketing efforts.