Final Analysis Report

Dakota Rennels

Maryam Aslani

Abdul Khan

Introduction

A non-profit organization has hired our team to build a predictive model to enhance the cost-effectiveness of the organization's direct marketing campaign. Currently, the organization is focused on reaching out to prior donors in the hopes of generating revenue. The strategy being presently leveraged by the non-profit has a response rate of approximately 10% which makes the strategy not sustainable. This approach isn't sustainable because it's not cost effective to reach out to everyone as the overall return on investment is -$0.55.

By reaching out to our team, the non-profit is aiming to implement a classification model to refine the approach of their marketing campaign. The intent behind the classification model is to help the non-profit more accurately identify likely donors. The impact of such identification will be a more streamlined marketing campaign that homes in on previous donors that are more inclined to donate again thus increasing revenue and maximizing profits. Currently, our team is given historical data and 2007 scoring observations to develop models by utilizing prior donor information, donation status in the last campaign and the amount of their donation.

Furthermore, the non-profit would like to develop a model that can predict the expected amount a particular donor is inclined to make. The current average expected donation is $14.50, if the non-profit can focus their efforts on those previous donors who are expected to donate a larger amount, then the non-profit would be able to further maximize revenue and profits. While this ask may appear to be a separate ask then the classification model spoken to above, these two tasks are intrinsically tied together as they work to increase the over return on investment for the non-profits marketing campaign while also opening the door for our team to win future work.

Our initial analysis of the issues behind the non-profits direct marketing campaign are that the campaign, as it is today, isn't focused on a particular audience. The net they have cast is too large

and thus resulting in a negative return on investment. We believe that by homing in on a few metrics and their relationships with each other, we cast many smaller nets that can overall produce a greater return on investment and thus ultimately maximize profits. We want to take a step back, review our historical data, and pinpoint the 'hotspots' and factors that resulted in an individual's donation and more specifically identify donors that are able and willing to make larger donations thus further maximizing profits.

The rest of this report is organized as follows: the data understanding section provides the results for exploratory data analysis, the data preparation section documents the steps taken to prepare the data, the modeling section provides the details for classification and regression models developed and the best model selected, the evaluation section discusses the accuracy of models, the deployment section discusses actionable findings, and finally the conclusion section concludes our report and provides a summary of our findings.

Data Understanding

Initial analysis of the data was performed to identify the best target variable to answer the looming business question. There has been a dataset presented to our team to help build an efficient model, *nonprofit.xlsx*. Table 1 presents the summary statistics of the provided dataset. It was confirmed the data has 6002 observations and none of the variables had missing values. It was also important to note the minimum and maximum values of each variable. These statistics along with the data dictionary will help identify what data type each variable is. Using the data dictionary, the *donr* variable, representing the donation status, was identified as the target variable for classification and to answer the business question of who to target for potential donors. The *damt,* representing the contribution amount, has been identified as the target variable for regression models to predict

how much a donor may contribute. It is worth noting that the maximum value of donation in previous years was 27 dollars with an average of $14.50 across all 6002 records.

Table 1. Summary statistics

Variable	Missing	Count	Minimum	Maximum	Mean	Standard Deviation	Skewness
damt	0	6002	0	27	7.209	7.3612	0.11698
donr	0	6002	0	1	0.499	0.5	0.00467
gifa	0	6002	1.89	72.27	11.678	6.5281	1.74146
gifdol	0	6002	23	1974	115.8	86.538	6.09138
gifl	0	6002	3	642	22.981	29.3964	7.18035
gifr	0	6002	1	173	15.654	12.4246	2.67345
hv	0	6002	51	710	183.905	72.7705	1.4889
inc	0	6002	1	7	3.939	1.4019	-0.01495
incavg	0	6002	14	287	56.789	24.8335	1.85779
incmed	0	6002	3	287	43.949	24.6644	2.00492
kids	0	6002	0	5	1.584	1.4125	0.39406
lag	0	6002	1	34	6.319	3.6414	2.41056
low	0	6002	0	87	13.885	13.1046	1.35139
mdon	0	6002	5	40	18.789	5.5963	1.1176
npro	0	6002	2	164	61.354	30.3052	0.28319
ownd	0	6002	0	1	0.885	0.3196	-2.40714
sex	0	6002	0	1	0.608	0.4883	-0.44168
wlth	0	6002	0	9	7.023	2.331	-1.46091

Further exploratory data analysis was performed to determine the variables' correlation. Presented in figure 1, the highest positive correlations are highlighted in red color, while the highest negative correlations are presented in blue. While analyzing the correlation matrix, we focused on relationships of target variables *donr* and *damt*. There are some strong correlations between the specified variables that could play an important role in the modeling and analysis. Variables *donr* and *damt* share relatively positive correlation with *ownd* and *wlth* variables. Moreover, both variables *damt* and *donr* share a strong, negative linear correlation to the *kids* variable. Variables *lag*, *low*, and *mdon* have a relatively low, negative linear correlation with variables *donr* and *damt*.

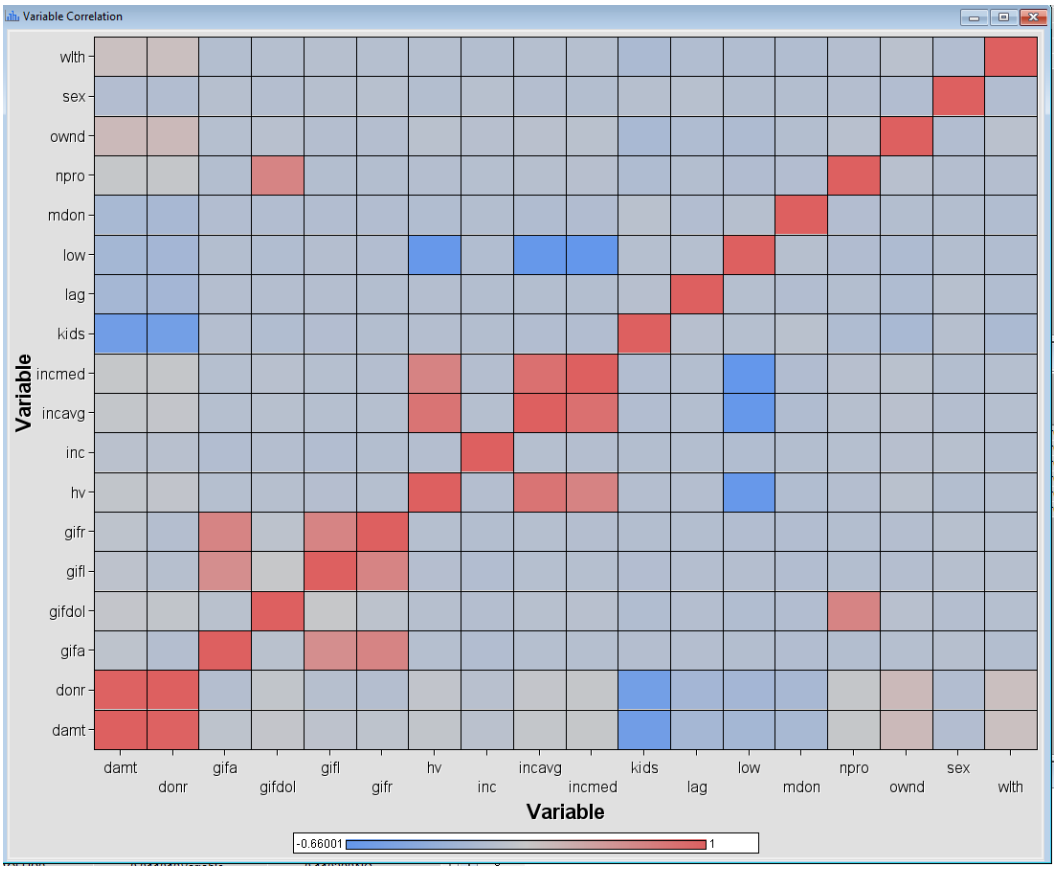
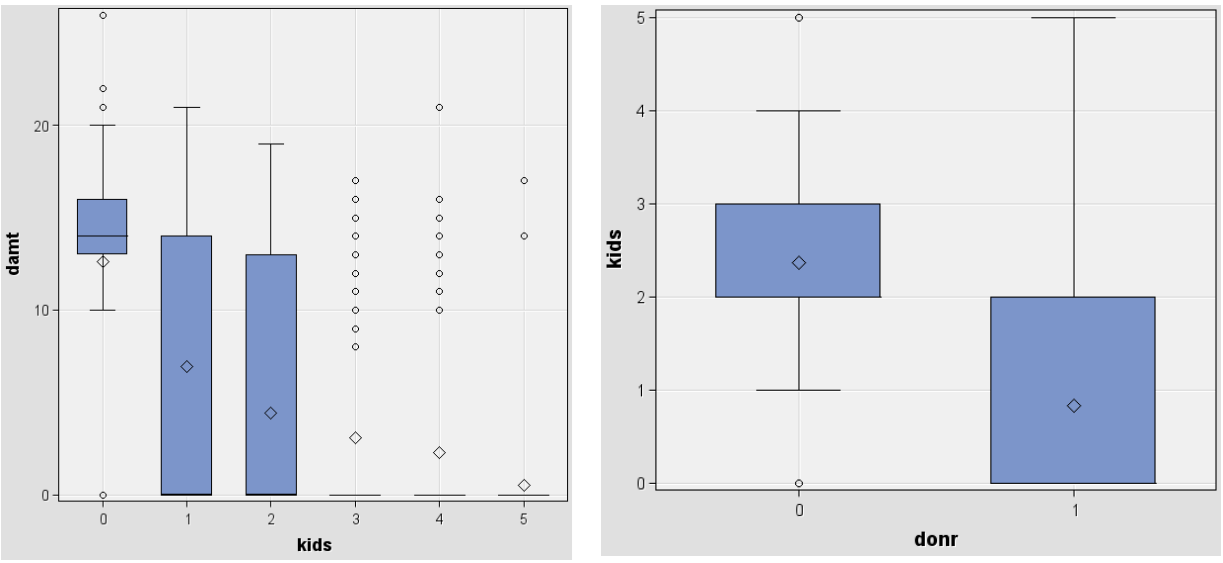


Figure 1. Correlation matrix for dependent and independent variables

Presented in figure 2 is the box plot for donation status and donation amount variables, where we can observe that they have both been affected by the number of kids.



a) Number of kids and donation amount

b) Donation status and number of kids

Figure 2. Exemplary box plot for targeted variables

With all exploratory data analysis conducted and correlation visualized and observed, we expect that a predictive model can be utilized to predict the donation status of individuals in the next campaign and provide a prediction of the contribution amount.

Data Preparation

For the data preparation phase, initially we focused on confirming the variables' data type and their corresponding role based on the provided data dictionary. We set both *donr* and *damt* as target variables with binary and interval types respectively. Where the models are focused on classification or regression task the corresponding relevant target variable is selected by rejecting the other. For classification models *damt* is set to reject to predict the *donr*, while *donr* are set to reject for regression model to prevent any data leaking in modeling steps. ID and region were set

to nominal, the *ownd* and sex variables set to binary, and the *inc* variable was set to ordinal, while remaining variables are set to interval.

Once the variables were assigned to their correct data type and target variables were identified, we partitioned the data for model training and testing purposes. With only 6002 observations in the dataset, a 70% training and 30% validation split was selected to confirm existence of enough data for accurate model training and evaluation.

Finally, where it is required by the modeling algorithm (e.g., KNN, Neural Network) in the data preparation phase, the input variables with interval data type were scaled to achievebetter results. Random sampling was performed; however, a seed is set for reproducibility purposes.

Modeling

This section provides the modeling details, including but not limited to the type and modeling algorithms used in this analysis. For predicting the possible individual contribution to the campaign, the first subsection provides the classification models details. Then, details for regression models used to predict the donation amount are discussed in the second subsection. Both these models will be used in conjunction for deployment and targeting individuals discussed in later sections.

Classification Models

For classification multiple models are devised to provide the most accurate prediction. Models utilized include KNN (with 18 different nearest neighbors), Neural Networks, Naïve Bayesian, and 4 different types of decision trees (including gradient boosting and random forest). Figure 3 presents the diagram of EDA and modeling within SAS Enterprise Miner.

The first cluster of nodes in the top left are the nodes used for initial exploratory data analysis. The data was then partitioned to train and test (for 2 splits of data in SAS validation plays as the test set) for modeling and evaluation. The top cluster of nodes are tree-based models (gradient boosting, random forest, and decision trees), followed by a Naïve Bayes model The middle group of nodes are the Neural Network models, followed by the AutoNeural Networks model that is connected straight to the partition. However, the specified Neural Networks require variables to be transformed which is achieved using the *transform node*. The KNN classification models is the cluster of models in the bottom. The model was run with multiple K-values to identify the optimal K-value for the KNN model. Model comparisons were performed independently for the KNN models, Neural Networks, and tree-based models, and Naïve Bayes model, respectively. Once each model comparison selected the best performing model of its group, an overall model comparison was used to identify the best model overall which then was used for scoring the model.

I

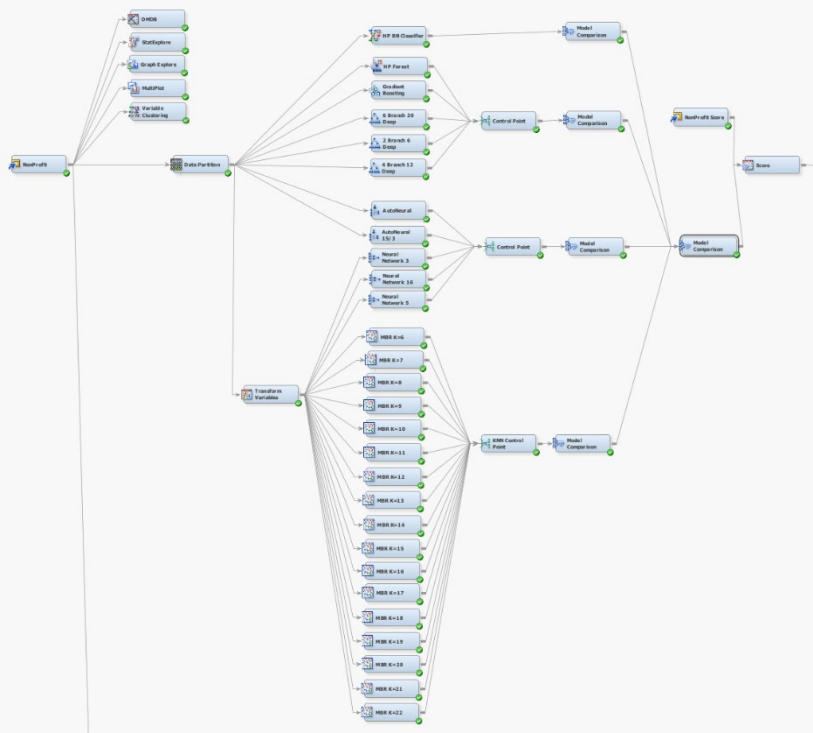


Figure 3. Classification models utilized in SAS Enterprise Miner

Model comparison presented in Figure 4 identifies the best model for classification of the target variable *donr* using the KNN models with different values of K as hyperparameter. The KNN node with K = 16 is the most accurate model of all the KNN models and will be used by model comparison to represent the KNN group in the overall model comparison

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	MBR26	MBR26	MBR K=16	donr	donr	0.197558
	MBR27	MBR27	MBR K=17	donr	donr	0.199778
	MBR23	MBR23	MBR K=13	donr	donr	0.199778
	MBR25	MBR25	MBR K=15	donr	donr	0.200333
	MBR31	MBR31	MBR K=12	donr	donr	0.200888
	MBR28	MBR28	MBR K=19	donr	donr	0.202553
	MBR34	MBR34	MBR K=22	donr	donr	0.203108
	MBR32	MBR32	MBR K=21	donr	donr	0.203108
	MBR33	MBR33	MBR K=20	donr	donr	0.203108
	MBR24	MBR24	MBR K=14	donr	donr	0.203663
	MBR19	MBR19	MBR K=18	donr	donr	0.204218
	MBR22	MBR22	MBR K=9	donr	donr	0.209212
	MBR29	MBR29	MBR K=10	donr	donr	0.210322
	MBR30	MBR30	MBR K=11	donr	donr	0.210877
	MBR36	MBR36	MBR K=6	donr	donr	0.211432
	MBR21	MBR21	MBR K=8	donr	donr	0.211987
	MBR20	MBR20	MBR K=7	donr	donr	0.215871

Figure 4. KNN with different K as hyperparameter

Running the tree-based models and comparison node, we identified the best model for classification of the target variable *donr* using tree-based models presented in Figure 5. The Gradient Boosting model performed the best and performed slightly better than the gradient boosting model. This means the Gradient Boosting model will be the tree-based model in the final overall model comparison.

Selected Model	Predecessor Node	M	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	Boost	...	Gradient Boosting	donr	donr	0.108768
	HPDMForest	...	HP Forest	donr	donr	0.112653
	Tree	...	6 Branch 20 Deep	donr	donr	0.130411
	Tree4	...	4 Branch 12 Deep	donr	donr	0.13263
	Tree3	...	2 Branch 6 Deep	donr	donr	0.145949

Figure 5. Tree based models

Finally, we ran the Neural Network nodes to identify the best model for classification of the target variable *donr* utilizing the Neural Network models presented in Figure 6. The AutoNeural Network with 15 maximum iterations and 3 hidden is the best model selected from the Neural Network models.

Selected Model ▼	Predecessor Node	M	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	AutoNeural4	...	AutoNeural 15/3	donr	donr	0.091565
	Neural7	...	Neural Network 16	donr	donr	0.096004
	Neural6	...	Neural Network 3	donr	donr	0.107103
	Neural8	...	Neural Network 5	donr	donr	0.118757
	AutoNeural3	...	AutoNeural	donr	donr	0.517758

Figure 6. Models utilizing Neural Networks

Followed by the identification of the best model in each group of models, the best model overall is evaluated using the model comparison node presented in Figure 7. The AutoNeural 15/3 model was selected as the best model. A more in-depth analysis between the results of the models will be explained in the evaluation section.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate
Y	MdlComp7	AutoNeural4	AutoNeural 15/3	donr	donr	0.091565
	MdlComp18	HPBNC2	HP BN Classifier	donr	donr	0.096559
	MdlComp4	Boost	Gradient Boosting	donr	donr	0.108768
	MdlComp6	MBR26	MBR K=16	donr	donr	0.197558

Figure 7. Best overall models for classification task

Regression Models

Regression analysis using multiple algorithms is conducted to predict the expected amount of donation by individuals selected by the best classification model. The algorithms used are presented in the lower branch of Figure 8 including polynomial linear regression, tree based regressors (gradient boosting, decision tree, and random forest), neural networks and auto neural.

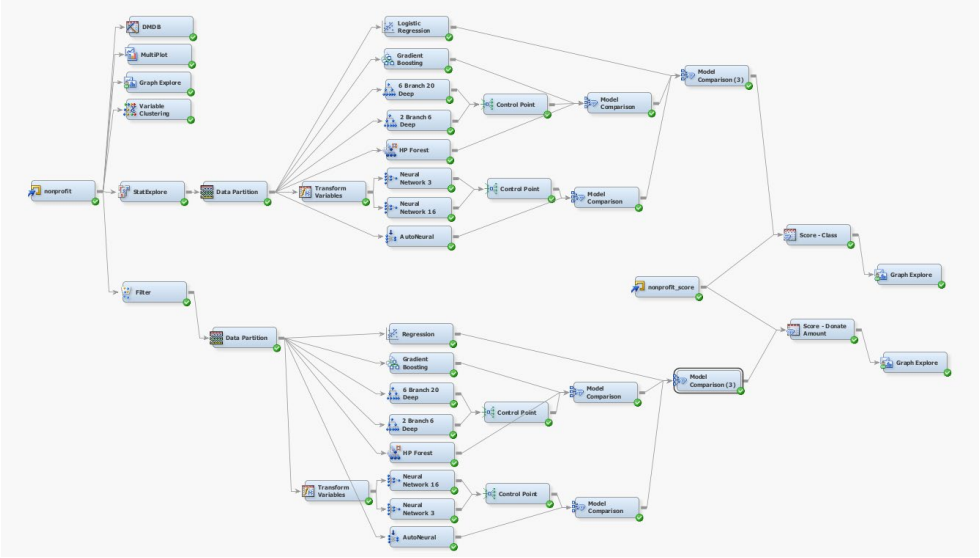


Figure 8. Regression models utilized to predict the donation amount

Following the same pattern, the regression models are evaluated first within each group and then are sent to model comparison nodes to find the best overall model. As presented in Figure 9, we can see that the neural network with 16 nodes in hidden layer is the best performing model presenting the lowest average squared error.

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error
Y	MdlComp4	Neural4	Neural Network 16	damt	damt	1.074959
	Reg2	Reg2	Regression	damt	damt	1.108359
	MdlComp5	Boost2	Gradient Boosting	damt	damt	1.201377

Figure 9. The best overall regression models

Evaluation

In this section we will first evaluate the classification models utilized in this analysis followed by the evaluation of the regression models.

Classification Models

It is important to address why we used the 70/30 data partitioning. The best model for the 40/30/30 data split generated (the Neural Network with 50 maximum iterations and 5 hidden units with backpropagation) provided the misclassification rate of 0.110617 sensitivity of 0.90, and specificity of 0.87, while the 70/30 data split generated a decently better model with a lower misclassification rate 0.091565, slightly higher sensitivity of 0.91 and specificity of 0.89. Table 2 and 3 provide the respective event classification and accuracy for the 70/30 and 40/30/30 data splits.

Table 2. Event Classification for different data splits:

Data Partition	**Model**	**Fales Negative (FN)**	**True Negative (TN)**	**False Positive (FP)**	**True Positive (TP)**
40/30/30	Neural Network 16	86	788	113	812
70/30	AutoNeural 15/3	73	811	92	826

Table 3. Accuracy with different data splits :

Data Partition	**Model**	**Misclassification Rate**	**Sensitivity [TP/(TP+FN)]**	**Specificity [TN/(TN+FP)]**

40/30/30	Neural Network 16	0.110617	0.9042	0.8746
70/30	AutoNeural 15/3	0.091565	0.9188	0.8981

Given the presented results and the better achievable accuracy using the 70/30 data splits, it is worth comparing the results for the models utilizing the same data split. Table 4 and 5 provide the respective event classification and accuracy for models utilizing the 70/30 data splits. Comparing the models, AutoNeural 15/3 has the lowest misclassification rate and a high sensitivity and specificity. It is worth noting that the Naïve Bayesian model has a relatively close misclassification rate with a higher sensitivity percentage but a lower specificity rate. Even though the misclassification rate is higher, it is worth testing this model as it might outperform the AutoNeural 15/3 model for maximizing profit. This is because the potential of adding one donor is significantly larger than falsely identifying a non-donor as a donor.

Table 4. Event Classification for top models and 70/30 data split

Model	**Fales Negative (FN)**	**True Negative (TN)**	**False Positive (FP)**	**True Positive (TP)**
AutoNeural 15/3	73	811	92	826
Naïve Bayes (HP BN Classifier)	50	779	124	849
Gradient Boosting	94	801	102	805
KNN (K = 16)	149	696	207	750

Table 5. Accuracy of top models with 70/30 data split

Model	Misclassification Rate	Sensitivity [TP/(TP+FN)]	Specificity [TN/(TN+FP)]
AutoNeural 15/3	0.091565	0.9188	0.8981
Naïve Bayes (HP BN Classifier)	0.096559	0.9444	0.8627
Gradient Boosting	0.108768	0.8954	0.8870
KNN (K = 16)	0.197558	0.8343	0.7708

Regression Models

Same 70/30 data splits are utilized for the regression analysis. Regression models are developed and trained on the training dataset to predict the *damt*, representing the donation amount, in the test set. We aim to utilize the trained regression model for selecting and contacting individuals who are expected to contribute a higher dollar amount first. We should note that we first filtered out individuals who did not have any contribution in the previous cycle and trained the model on those who contributed to the last campaign. Table 6 presents the results of the top three regression models utilized in this analysis and their corresponding average squared error and sum of squared errors. The neural Network with 16 neurons in the hidden layers presents the best overall accuracy and is utilized for the remainder of regression analysis conducted.

Table 6. Top three regression models

Model	Misclassification Rate	Sensitivity [TP/(TP+FN)]
Neural Network 16	1.07	857.81

Polynomial Regression	1.10	884.47
Gradient boosting Regressor	1.20	958.69

Deployment

The best models developed and evaluated in the modeling section are utilized here for predictive analysis of the score dataset provided. First, we applied the classification model to predict individuals who are expected to donate in this campaign. The model identifies 1045 individuals as those who are not expected to contribute and the remaining 955 as prospective contributors. It is suggested that we only contact the individuals who are predicted to contribute by the model to refrain from additional costs exposed to the nonprofit organization. Figure 10 presents the predicted number of individuals in each group.

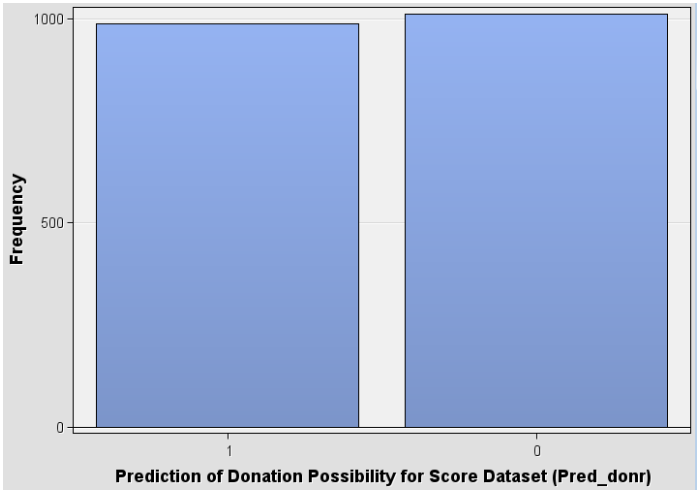


Figure 10. Count of individual predicted in each group

While contacting all the prospective contributors could be an approach to take, it is not suggested as the number of contributions is different. Those who are expected to contribute could only be contacted by considering a threshold that maximized the profit, while prioritization can also be considered based on the predicted amount of donation. Figure 11 presents the box plot of predicted contributions by the regression model. We can observe the median contribution of $14.30 with 75 percentile and maximum whisker of $15.30 and $18.40, respectively; however, targeting only this range could result in a potential profit loss. For overall max profitability, it would be valuable to target individuals whose expected contribution is at or above the minimum whisker value. This is because the loss of marketing to a non-donor would be $2.00 per flyer sent. This means even if five flyers were sent out and only one donor was targeted, the resulting donation would offset the cost of mailing the flyers and providing a net gain of donation.

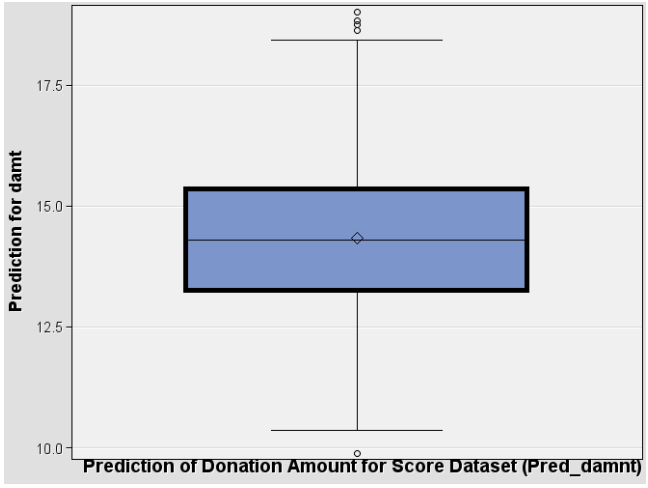


Figure 11. Predicted donations by participants in the score dataset

Conclusion

To recap, our team having been asked by the nonprofit organization to develop predictive models capable of identifying individuals who are expected to contribute to the next campaign and provide the predictive model capable of estimating the donation amount. In this report we developed models for both classification and regression analysis and evaluated the developed models based on their performance on the test dataset. The best models were then selected and applied to the score dataset provided by the nonprofit organization. Individuals who are expected to contribute to the next campaign are identified and expected donation amounts are predicted. The predicted donation amount can be utilized to develop a threshold in accordance with the advertisement costs to increase the profit achieved and lower the cost. Our results present accurate models developed and applied to the datasets provided.

One of the limitations worth mentioning for future modeling and predictions was the lack of location data. The only regional data provided was the five geographic regions in the *region* variable. Since flyers are being distributed, it makes sense that the nonprofit organization has the exact location information to pair with the IDs from the data supplied. More in-depth location data would have been valuable for insights into potentially grouping repeat donors by more than just regional location.