**ADTA 5230**

**Data Analytics 2**

**Develop predictive models to improve the cost-effectiveness of non-profit Organization.**

**Group- 13**

**Sai Vivek Reddy Gurram - 11547150**

**Sai Chaitanya Kota -11547151**

**Introduction:**

This project is to help a non-profit organization to increase the outcomes of the donations.

Their most current figures show that the average total response rate is around 10%. The average amount from individuals who reply to the mailing and donate is $14.50. Each mailing includes a free gift of custom address labels and costs $2.00 to make and distribute. Since the estimated profit from each letter is 14.50 x 10% - 2 = -0.55, mailing everyone is not cost-effective.

ABOUT DATASET:

Here, we have 2 data sets as non-profit and non-profit score data sets which contains 6003 and 2008 records respectively.

Overall, It has 24 variables in this datasets and consists of:

ID Number, TER1, TER2, TER3, TER4 (Regions), OWND (homeowner), KIDS (no of children), SEX (gender), WLTH (wealth rating), HV (avg home value), INCMED (median family income), INCAVG (average family income), LOW (percent of low income), NPRO(no of promotions), GIFDOL(dollar amount), GIFL(dollar amount of largest), GIFR (dollar amount of most recent gift), MDON (no of months), LAG(duration between first and second gift), GIFA (average dollar amount of gifts to date), DONR(Classification Response), DAMT (Prediction Response), VALID (Indicator of validation data) .

Classification Variables:

Overall, there are 24 variables in this dataset. In classification DAMT is rejected and DONR is Target variables

Table

Description automatically generated

Regression Variables:

In Regression Variables DONR is rejected and DAMT is Target.

Table

Description automatically generated

There is a combination of categorical and numerical variables in these all-predicted variables.

Partitioning Data Set with cluster method w.r.t valid variable in nonprofit.xlsx:

Table

Description automatically generatedText

Description automatically generated with medium confidence

Missing Values in Data: There are no missing values in the data

Diagram

Description automatically generated

Transform Variables: interval variable values have a minimum of 0 and a maximum of 1 after the transformation.

Table

Description automatically generated

**Logistic Regression**: It estimates the probability of an event occurring and the outcomes are bounded between 0 and 1.

Logit(pi) = 1/(1+ exp(-pi))

Graphical user interface, application, Teams

Description automatically generated

Fit statistics of logistic regression

Table

Description automatically generated

In logistic regression the misclassification rate is 0.11 for training dataset and 0.12 for Validation dataset.

Confusion matrix of logistic regression model for training and validation data

Table

Description automatically generated

ACCURACY= (TP+TN)/(TP+TN+FP+FN)

For training= (1795+1742)/(1795+1742+247+200) == 0.887

For validation = (908+863)/ (908+863+156+91) == 0.877

**Neural network:**

An artificial intelligence technique called a neural network instructs computers to analyze data in a manner like the way the human brain does.

* Optimization

In Neural network we use the back propagation method with maximum iterations of 50

Graphical user interface, text

Description automatically generated

* Network

Graphical user interface, text

Description automatically generated

An Architecture is used in constructing the network.

In Neural Network the architecture used is Multilayer Perceptron with 3 hidden units.

**Fit statistics:**

In Neural Network the misclassification rate is 0.09 for Training and 0.11 for Validation dataset

Table

Description automatically generated

Classification table (or) Confusion matrix in neural networks for train and valid data

A screenshot of a computer

Description automatically generated with low confidence

Table

Description automatically generated

**ACCURACY=** (TP+TN)/(TP+TN+FP+FN)

For Train = (1842+1767)/ (1842+1767+222+153) = 0.906

For Validation = (918+878)/ (918+878+141+81) = 0.889

**Decision Tree with multiple branches and depths:**

A Decision Tree is a particular sort of probability that enables you to choose a kind of process

Graphical user interface, text, application, chat or text message

Description automatically generated

Connected to ensemble to give the best model from the decision tree.

Graphical user interface, diagram

Description automatically generated

Ensemble method is nothing but obtaining the best decision tree model within several number of decision tree’s.

In this model we have used bagging to obtain the best decision tree model

Fit statistics for ensemble technique

Table

Description automatically generated

Misclassification rate of Training is 0.12 and for validation it is 0.12.

**Confusion matrix / classification table**

Table

Description automatically generated

KNN model with k neighbors = 4,5,6,7,8.

Fit statistics:

Model selection based on mis-classification rate

Table

Description automatically generated

Here overall where KNN with 8 neighbors had the least misclassification rate and it is the best among all the other options

Event classification table

Table

Description automatically generated

Model comparison of all classification models based on misclassification

Fit statistics

Table

Description automatically generated

Here, the Neural Network model has the least misclassification rate and the best model for classification.

**Prediction**:

Diagram

Description automatically generated

Uploaded excel file as score classification

 set the role as score. For score classification

Score module predicts the outcome of DONR based on best classification model.

The best classification model is Neural Network, the below predictions are based on Neural network model.

The snapshot of predicted results is

Table

Description automatically generated

**Overall classification model for this project**

Diagram

Description automatically generated

Regression Model:

For predicting the DAMT variable we constructed 4 models and predicted with the best performed model.

1. Linear Regression Model:

It is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables.

Table

Description automatically generated

1. Auto Neural:

Train vs Valid MSE plot

Graphical user interface, application, table

Description automatically generated

1. Gradient Boosting:

Plot of Average squared error of training and validation data

Chart, line chart

Description automatically generated

1. Decision Tree:

4 decision trees with ensemble technique

Table

Description automatically generated with medium confidence

After comparing all the models in model comparison block, we predicted the DAMT variable.

Graphical user interface, diagram, text, application, chat or text message

Description automatically generated

The best performing model is Ensemble, all the predictions are based on ensemble model

Table

Description automatically generated

The predicted results are

Table

Description automatically generated