MAY 2023

BY: Gloria MWISENEZA

A picture containing graphical user interface

Description automatically generated

**The project comprises three predictive analytics models,**

**Logistic regression, Decision trees and the artificial neural networks.**

‘

PREDICTIVE ANALYTICS PROJECT

**Introduction**

This predictive analytics project is comprised of three models, regression analytics, decision trees and neural networks. Our analysis goal is to maximize the profit when sending out mails to likely to be donors. We are going to analyze the data from the most recent campaign run by Mid-West University Foundation and predict the expected gift amount from the donors. This project is going to follow the CRISP-DM process which is descriptively explained below in its six phases, including, business understanding, data understanding, Data preparation, model building, testing and evaluation and finally deployment.

Business understanding, Midwest University Foundation is looking to improve the cost effectiveness of their direct marketing campaigns to previous donors. The goal is to build a prediction model that can maximize the net expected profit when sending out mails to those who are likely to be donors.

# Data understanding, we are using data from previous donors, the data has their identity number, gender, their geographical region, whether they are homeowner or not, the number of children they have as well as their household income. The data also analyzes their previous gifts in terms of dollar amount as well as frequency.

Data preparation, when preparing the data, we check if there is any missing values using (is.na) and if so address it first, we are also making sure that all categorical variables are coded as factors before developing the model we check for multicollinearity in the data using vif function.

Model building, we are going to develop three models, logistic regression, decision trees and neural network model. The data is going to be partitioned at a ratio of 70-30% for training and testing respectively. The models will be predicting the DONR variable and using home, chld, hinc, wrat, avhv, incm, tgif, lgif, rgif, tdon, tlag, agif as predictors excluding DAMT.

Testing and evaluation, In the regression model we are going to look for the statistically significant model, for decision trees we are extracting the rules to better understand and evaluate the model and for neural networks we are testing for both 15 nodes and 50 nodes. We are going to compare the roc curves and the auc statistics from all the three models.

Deployment, when deploying the model we are going to compare the auc statistics from all three models and the model with the highest AUC statistics will be the model to use as the higher the auc the better is for the model if to predict classes as 0 or 1.

**Analysis and Results**

1st model is logistic regression, When preparing the data we first assessed any missing variables that may be present in the data using is code, [sum(is.na(MWUF))] and came to the conclusion that there were no missing values in the data set. The next step was to get all integer/categorical data as factors, and as numeric converted donr to factors using factor function [

MWUF$donr = factor (MWUF$donr)], and double checking it using “str” function [str(MWUF) ]. We also checked for multicollinearity using (vif (LogR1)>4) to find if there is any correlation among the variables and there was none.

In the modeling stage, we are predicting DONR variable using following variables: home, chld, hinc, wrat, avhv, incm, tgif, lgif, rgif, tdon, tlag, agif variables as predictors and excluding DAMT., as we preapare the data all the columns that will not be needed are dropped first as following [MWUF= MWUF[,-c(1,2,3,4,5,9,13,14,15,23)]. We are using 70-30% partitioning ratio for all models for train-test. For model construction and testing models’ predictive performance and for consistency we are using seed *123.*

*set. Seed (123)*

*train. Index= sample(c(1:dim(MWUF)[1]), dim(MWUF)[1]\*.7)*

*train = MWUF [train. Index, ]*

*test= MWUF [-train. Index, ]*

The first model we are building is the logistic regression (logR1) and we can see clearly what variables are statistically significant as we look at the summary of the model, by using 0.05 as our threshold value. The variables that happen to be statistically significant are intercept, home ownership, the number of children they have, their wealth rating, Median Family Income in potential donor's neighborhood in $ thousands.

*,* dollar amount of lifetime gifts to date*, n*umber of months since last donation *and finally the* number of months between first and second gift. The following code were used to build the logR1 model,

*LogR1= glm(donr~., data=train, family = "binomial")*

*options(scipen=999)*

*summary (LogR1)*

*A screenshot of a computer

Description automatically generated with medium confidence*

*We predict the model using the predict function and we exclude the variable in the 13th (donr) column as that is what is we are predicting. Plotting the roc curve is the next step and to do that we need to install the proc, rpart and rpart packages as well as reading the respective libraries in R, here is the roc curve obtained, as well as the area under curve we also called the confusion matrix for the model and all the relevant statics including its precision, sensitivity and f1score.*

*cv = table(test$donr, LogR1.pred>0.5)*

*A picture containing text, receipt, font, white

Description automatically generated*

*#Accuracy of LogR1, Precision and sensitivity and F1 score*

*sum(diag(cv))/sum(cv)*

*# Precision (Positive Predictive Value)=TP/(TP+FP)*

*precisionR1 = cv[2, 2]/(cv[2, 2] + cv[2, 1])*

*#Calculating Hit Rate (TPR) = TP/(TP+FN)*

*sensitivityR1= cv[2, 2]/(cv[2, 2] + cv[1, 2])*

*f1scoreR1= 2\*precisionR1\*sensitivityR1/(precisionR1+sensitivityR1)*

*f1scoreR1*

*install.packages(c("rpart", "party", "partykit", "caret", "pROC") ,repos = "http://cran.us.r-project.org")*

*library(rpart)*

*library(party)*

*library(partykit)*

*library(caret)*

*library(pROC)*

*library(rpart.plot)*

*test\_probR1 = predict(LogR1, newdata = test[, -13], type = "response")*

*test\_rocR1 = roc(test$donr~ test\_probR1, plot = TRUE, print.auc = TRUE)*

*par(pty="s")*

*test\_rocR1 = roc(test$donr ~ test\_probR1, plot = TRUE, print.auc = TRUE, legacy.axes=TRUE, percent=TRUE, xlab="False Positive Rate", ylab="True Positive Rate", lwd=4)*

*auc(test\_rocR1)*

*Chart

Description automatically generated*

*2nd model is a decision tree, the variables used in this decision tree model are ,* home, chld, hinc, wrat, avhv, incm, tgif, lgif, rgif, tdon, tlag, agif variables. We use rpart package to build the decision tree DT and predict function to make predictions, rpart.rules function help in calling out the model rules,

The rules are as the following,

A picture containing text, font, screenshot, white

Description automatically generated

Decision tree model

A picture containing text, diagram, sketch, line

Description automatically generated

We found the models confusion matric, accuracy and other statistics ,

*A close-up of numbers

Description automatically generated with low confidence*

*DT model auc statistic is 83.45%*

*Roc curve for the decision tree model DT*

*A picture containing text, diagram, line, plot

Description automatically generated*

*3rd model is the neural network model, we first explore the model using 15 nodes and after we use 50 nodes. For the neural network model, we use the package called nnet.*

*Neural network with 15 hidden nodes (ANN1) analysis is as follows, confusion matrix*

A close up of numbers

Description automatically generated with low confidence

Auc for ANN1 is 88%

A picture containing line, diagram, drawing, sketch

Description automatically generated

Roc curve for the ANN1 with 15 hidden nodes

A picture containing text, diagram, line, plot

Description automatically generated

Lastly, we build a neural network model (ANN2) with 50 hidden nodes, down is its confusion matrix , auc and plots

A close-up of numbers

Description automatically generated with low confidence

A picture containing sketch, line, drawing, art

Description automatically generated

Roc curve for the ANN2 with 50 hidden nodes

A picture containing text, diagram, line, plot

Description automatically generated

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **MODEL** | **ACCURACY** | **PPV** | **AUC** | **F1- Score** |
| LogR1 | 80.56% | 82.76% | 89.32% | 80.66% |
| DT | 83.34% | 79.57% | 83.5% | 83.92% |
| ANN1 | 79.06% | 75.27% | 88.60% | 79.95% |
| ANN2 | 79.28% | 79.97% | 85.0% | 77.47% |

**Conclusion**

Upon analyzing all the three predictive models, logistic regression, decision trees and artificial neural network, I concluded that I would recommend using regression analysis. I based my decision mainly to the model’s auc, which is higher compared to other models and this indicate that logistic regression has the capability of making better prediction of the classes of the outcome in classes of 1 or 0 than other models.

As I would recommend using logistic regression model, I am going to use the model to make predictions using the new data set MWUF\_new.csv to identify who would make donations

Prednew = predict(LogR1,MWUFnew[,-13],type= "response")

prednew.df = as.data.frame(Prednew)

prednew.df= ifelse(prednew.df>0.5, "donor", "non-donor")

prednew.df\_count = table(prednew.df)

prednew.df\_count

Using the LogR1 on MWUFnew data we find that 613 people would likely be donors and 1394 would be non-donors.



By having the above information, Mid-West University Foundation will be able to maximize the expected profit as they will send out mails to likely to be donors only cutting off on resources that would have been spent sending mails to everyone including the non-donors expecting 105 responses.

Specifically, sending mails only to likely be donors looks like this,

Average mailing donations $14.50

Mailing costs $2.00

Profit Mid-West University Foundation will get when Sending only to 613donors, (14.50-2.00)\*613 = $7662.5

Versus Profit Mid-West University Foundation will get when sending to everyone (613+1394) = 2007donors and non-donors expecting only 10% to be donors

(14.50-2.00) \*2007\*0.10= $2508.7

By employing predictions models especially logistic regression Mid-West University Foundation will earn profit of about more than $5000 (7662.5-2508.7) = $ 5153.8 as it will be targeting just donors.

# DATA DICTIONARY

* ID number
* REG1, REG2, REG3, REG4: Region (There are five geographic regions; only four are needed for analysis since if a potential donor falls into none of the four he or she must be in the other region. Inclusion of all five indicator variables would be redundant and cause some modeling techniques to fail. A “1” indicates the potential donor belongs to this region.)
* HOME: (1 = homeowner, 0 = not a homeowner)
* CHLD: Number of children
* HINC: Household income (7 ***categories***)
* GENF: Gender (0 = Male, 1 = Female)
* WRAT: Wealth Rating (Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and 0 being the lowest.)
* AVHV: Average Home Value in potential donor's neighborhood in $ thousands
* INCM: Median Family Income in potential donor's neighborhood in $ thousands
* INCA: Average Family Income in potential donor's neighborhood in $ thousands
* PLOW: Percent categorized as “low income” in potential donor's neighborhood
* NPRO: Lifetime number of promotions received to date
* TGIF: Dollar amount of lifetime gifts to date
* LGIF: Dollar amount of largest gift to date
* RGIF: Dollar amount of most recent gift
* TDON: Number of months since last donation
* TLAG: Number of months between first and second gift
* AGIF: Average dollar amount of gifts to date
* DONR: Classification Response Variable (1 = Donor, 0 = Non-donor)

DAMT: Prediction Resp

* onse Variable (Donation Amount in $).
* *Incm ,* Median Family Income in potential donor's neighborhood in $ thousands
* *tgif ,* Dollar amount of lifetime gifts to date

*tdon,* Number of months since last donation

*t lag ,*Number of months between first and second gift