MSIS 672- DATA MINING

FALL 2020

FINAL PROJECT

DIRECT-MAIL FUNDRAISING

GROUP MEMBERS:

ALISHA YOGANAND WARKE

AMIRA KAZI

ANKIT NASHINE

EXECUTIVE SUMMARY

Decision Tree and Multiple Regression models were used in order to predict TARGET_B and TARGET_D in FutureFundraising.csv. We used Decision Tree for TARGET_B, and Multiple Regression for TARGET_D. in order to assess the organization's purpose of predicting the future fundraising, the models underwent analysis that compared their outputs. The dataset used for building models was Fundraising.csv.

MAIN REPORT

I. INTRODUCTION:

The case Direct Mail Fundraising looks at the Veteran's organization which is one of the largest direct-mail fundraisers in the United States. The organization has been seeking an efficient way to help them predict their future fundraising to maximize their expected net profit. There are two sets of data that are available in order to conduct the analysis that will help the organization with their search

- TARGET_B (binary indicator)
- TARGET_D (donation amount)

The National Veterans' organization is looking for an efficient way to create a cost-effective method for their direct marketing campaign. The organization stands as one of the largest direct mail fundraisers in the United States with a database of over 13 million donors.

The organization's recent mailing records show that even though the organization has a large pool of donors, the overall response rate from them is only 5.1%. On average, the total donation from everyone who responded was $13. Potential donors receive a gift package in the mail that include gifts of personalized address labels and assortments of cards and envelopes. Each mail package costs about $0.68 to produce and be sent out. The organization is in attempts of developing a classification model to effectively capture donors, so that the expected profit is maximized. In order for the sample to have equal numbers of donors and non-donors weighted sampling is used, under-representing the non-responders.

The introduction and case narrative indicate that the core problem of the organization is it is struggling to follow up with the donors' contribution. An effective classification model is needed to maximize the expected net profit in the future. It needs to be determined which classification model is properly suited to the organization's needs and which method serves the best technique to predict the future fundraising for the veterans' organization.

2. TARGET_B

2. I) DATA EXPLORATION AND DATA PREPARATION:

For the data content, the datasets that were used for this case study were “Fundraising .csv” and “Future Fundraising.csv”. The file fundraising.csv contains 3120 records with 50% of donors (TARGET_B = 1) and 50% non-donors (TARGET_B = 0). The amount of donation is presented by TARGET_D and is also included but not used in this case for building a decision tree model for predicting TARGET_B.

Using sapply function, we checked the class of all the variables. We selected several variables such as “zipconvert_2”, “zipconvert_3”, “zipconvert_4”, “zipconvert_5”, “homeowner.dummy”, “INCOME”, “gender.dummy”, “WEALTH”, and “TARGET_B” and converted these into factors as their class was mentioned as integers. We removed the following columns “Row.Id”, “Row.Id”, and “TARGET_D”. TARGET_D is not significant as it is the donation amount. In order to bring them to a common scale, we normalized the following columns. “HV”, “Icemd”, “Icavg”, “IC15”, “NUMPROM”, “RAMNTALL”, “MAXRAMNT”, “LASTGIFT”, “totalmonths” “TIMELAG”, “AVGGIFT”. We normalized these because the range was quite high for some columns while others had quite low ranges.

2. II) DATA ANALYSIS:

In the process of data analysis, After normalization of both data frames, we ran partition for the data and divided the datasets, 60% going to training, and validation taking the remaining 40%. We built a Decision Tree model on training dataframe, and checked it’s performance on training dataset.

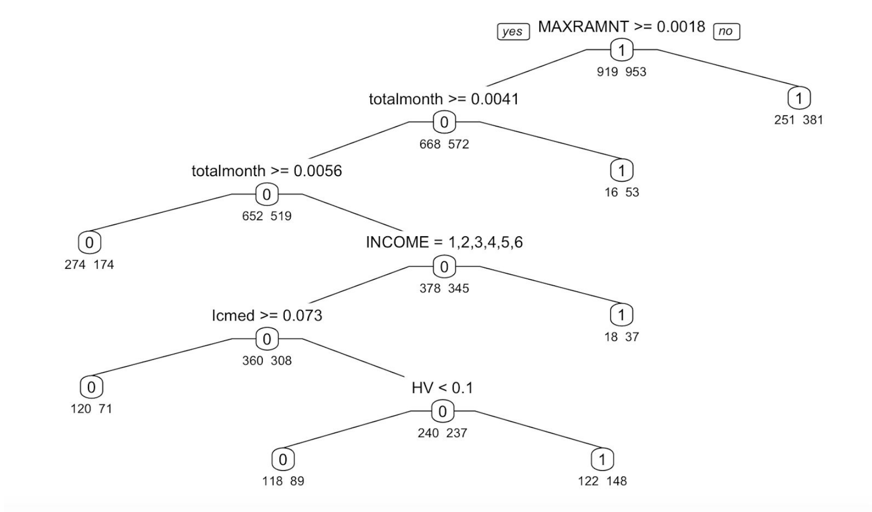


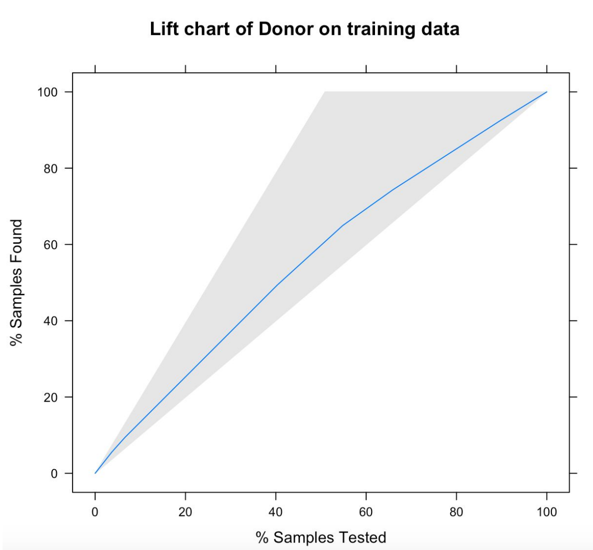
Figure1: Decision Tree model on Training dataset

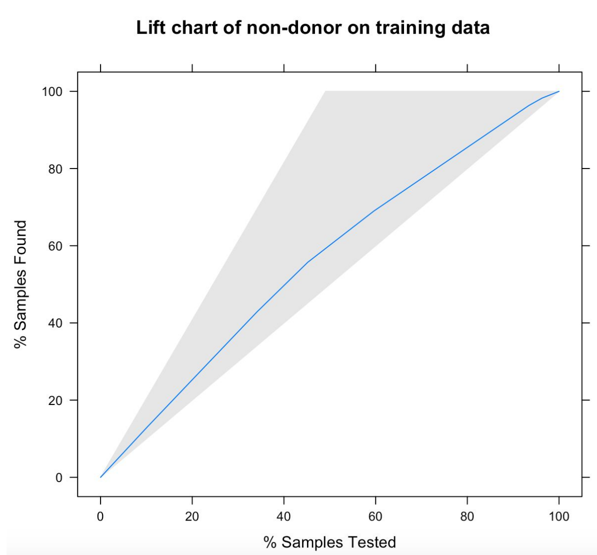
In the decision tree model, it can be seen we considered many variables but the most significant ones were “MAXRMNT”, “totalmonth”, “income”, “Icmed”, “HV”. We assessed the model’s performance on the training data set using a confusion matrix and lift chart. As found in the confusion matrix, our specificity on the training data set is 64.95% which is fairly good.

NOTE: We are considering specificity here because, positive class is 0 that means specificity represents class 1 here.

```
> confusionMatrix(class.tree.pred.train, as.factor(train.df$TARGET_B))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 512 334
         1 407 619

               Accuracy : 0.6042
                 95% CI : (0.5816, 0.6264)
    No Information Rate : 0.5091
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2069

 Mcnemar's Test P-Value : 0.008169

            Sensitivity : 0.5571
            Specificity : 0.6495
         Pos Pred Value : 0.6052
         Neg Pred Value : 0.6033
             Prevalence : 0.4909
         Detection Rate : 0.2735
   Detection Prevalence : 0.4519
      Balanced Accuracy : 0.6033

       'Positive' Class : 0
```

Lift chart of donors and non-donors on training data set is represented in the chart below:

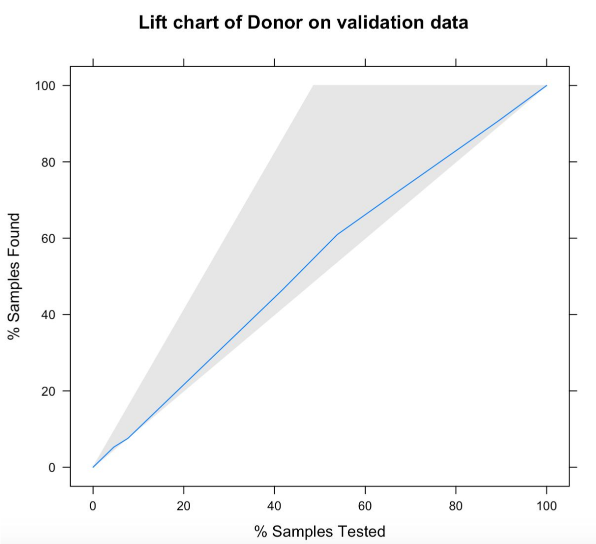


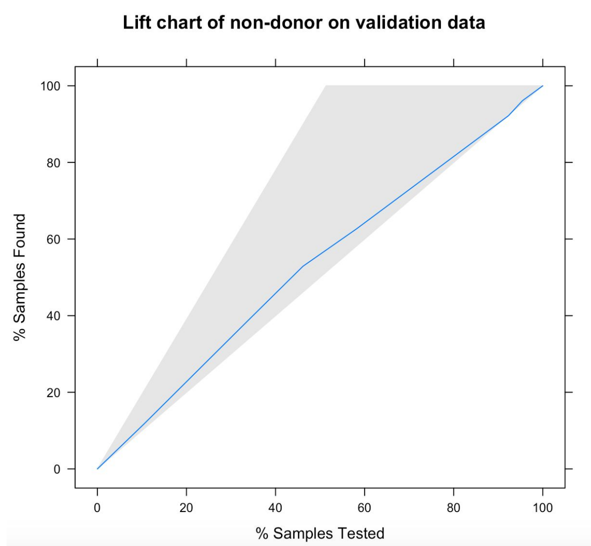


Moving forward, we assessed the model's performance on the validation data set as well, using a confusion matrix and lift chart. From the confusion matrix we found the specificity to be 60.96%, which is close to the specificity of the training data set (64.95%). Therefore, our decision tree model is good in making predictions on unseen data and does not have overfitting problem.

```
> confusionMatrix(class.tree.pred.valid, as.factor(valid.df$TARGET_B))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 339 237
         1 302 370

               Accuracy : 0.5681
                 95% CI : (0.5401, 0.5958)
    No Information Rate : 0.5136
    P-Value [Acc > NIR] : 6.41e-05

                  Kappa : 0.138

 Mcnemar's Test P-Value : 0.005839

            Sensitivity : 0.5289
            Specificity : 0.6096
         Pos Pred Value : 0.5885
         Neg Pred Value : 0.5506
             Prevalence : 0.5136
         Detection Rate : 0.2716
   Detection Prevalence : 0.4615
      Balanced Accuracy : 0.5692

       'Positive' Class : 0
```

Lift chart for donors and non-donors on validation data set is represented in the chart below:





By looking at the lift chart we can say that although our model is not perfect but the curve is above base-line, so it's better than the random benchmark(base-line).

We used this model to make predictions of TARGET_B on FutureFundraising. We created a data set with “donors” only, this tells us which records have donors. When we predict donations, we only select those records that have donors.

3. TARGET_D

3. I) DATA EXPLORATION AND DATA PREPARATION:

For the second part, we created a dataframe from “Fundraising .csv” by selecting only those records that have TARGET_B = 1. We used these records to build a linear regression model while target variable being TARGET_D and predictors as all the remaining variables except “Row.Id”, “Row.Id.” and “TARGET_B”(We dropped them in training and validation datasets). In the beginning we did not create dummy variables, this is because when we put as.factors they all act as dummy variables(diagram below) as shown in binary.

```
Call:
lm(formula = TARGET_D ~ ., data = train2.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.910  -3.273  -0.434   2.081 132.107 

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value            Pr(>|t|)    
(Intercept)       -7.72081    5.07033  -1.523              0.1282    
zipconvert_21      0.38899    1.03204   0.377              0.7063    
zipconvert_31     -0.17777    1.14944  -0.155              0.8771    
zipconvert_41      1.28406    1.09122   1.177              0.2397    
zipconvert_51           NA         NA      NA                  NA    
homeowner.dummy1  -0.67484    0.93211  -0.724              0.4693    
NUMCHLD            0.07533    1.13759   0.066              0.9472    
INCOME2           -2.98618    1.52701  -1.956              0.0509 .  
INCOME3           -2.71063    1.64988  -1.643              0.1008    
INCOME4           -1.34426    1.41799  -0.948              0.3434    
INCOME5           -0.44981    1.51244  -0.297              0.7662    
INCOME6           -1.40418    1.91306  -0.734              0.4632    
INCOME7           -2.96839    1.81208  -1.638              0.1018    
gender.dummy1     -0.55888    0.74365  -0.752              0.4526    
WEALTH1            2.97751    2.61208   1.140              0.2547    
WEALTH2           -2.24854    2.66348  -0.844              0.3988    
WEALTH3           -2.43625    2.66428  -0.914              0.3608    
WEALTH4           -1.09886    2.64284  -0.416              0.6777    
WEALTH5           -1.79506    2.64632  -0.678              0.4978    
WEALTH6           -1.36738    2.63201  -0.520              0.6036    
WEALTH7           -0.31218    2.60813  -0.120              0.9048    
WEALTH8            0.24112    2.29472   0.105              0.9163    
WEALTH9            1.55808    2.76650   0.563              0.5735    
HV                -4.92584    4.15355  -1.186              0.2360    
Icmed             30.15613   55.32763   0.545              0.5859    
Icavg            -11.58079   59.32509  -0.195              0.8453    
IC15             -87.11691  280.57130  -0.310              0.7563    
NUMPROM          310.80004  151.96552   2.045              0.0412 *  
RAMNTALL          -6.56262   24.43407  -0.269              0.7883    
MAXRAMNT        -126.51203   69.98716  -1.808              0.0711 .  
LASTGIFT        1648.85924  353.91942   4.659          0.00000376 ***
totalmonths     1091.50790  529.44559   2.062              0.0396 *  
TIMELAG          -26.62306  367.44729  -0.072              0.9423    
AVGGIFT         7176.75527  582.95459  12.311 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.865 on 747 degrees of freedom
Multiple R-squared:  0.5222,	Adjusted R-squared:  0.5017 
F-statistic: 25.51 on 32 and 747 DF,  p-value: < 0.00000000000000022
```

3. II) DATA ANALYSIS:

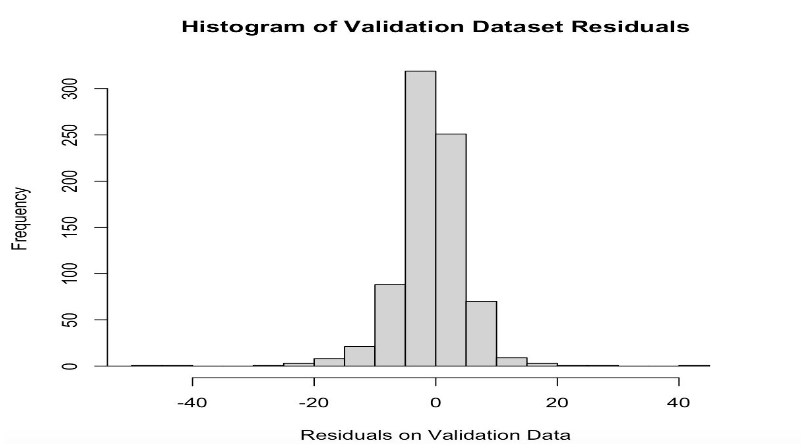
Moving on, we created training and validation sets splitting it by 50% equally.

Similar to the way we checked the decision trees performance on validation and training datasets, we are checking our linear regression model's performance on validation and training datasets.

We created a histogram based on Residuals of Training Data. Most of the errors lie around 0, this means most of our errors lie within 0. As all the errors are close to 0, we can conclude that our model is a good prediction. It's performance on the training data set is good.

Then we assessed the model's performance on validation data, and can now check the accuracy. The RMSE is found to be 7.31.

We again created a histogram based on Residuals of Validation Data.Most of the error residuals are around 0, so the model is reliable for the validation dataset too.



Using our model we next made predictions of donor amounts on TARGET_D for "FutureFundraiser" for all records where TARGET_B == 1 and add 0 for TARGET_B == 0 to show non-donors.

We now inserted the predicted values from the models we built to the original Dataset "FutureFundraising.df" with all attributes which we dropped during aur analysis and the original values before normalization, so the final outcome we see is the original dataset with all columns and TARGET_B indicating 1 for predicted donors and 0 for predicted non-donors along with TARGET_D showing the estimated donation amount from the predicted donors.

4. FINDINGS AND CONCLUSION

We found that in FutureFundraising our model classified 918 records as potential donors and our specificity on validation dataset was 0.6096 so out of these 900 predicted records, 60.96% are supposed to be donors

We can conclude that the veteran's should send mail packages to only those members that are classified as potential donors by our predictive model on future fundraising datasets, in order to improve the cost effectiveness of the direct marketing campaign. We can also send mail to those donors whose probability (valid.df$prob0) of being non-donor is between 0.5 and 0.6 as their probabilities of being non-donors (valid.df$prob0) was close to the probability of being donors (valid.df$prob1). This will be helpful in capturing those donors that our model misclassified as non-donors.

=====***=====