MIDTERM PROJECT

Hope INC.

Renukalaxmi Dudhe

Richard A. Chaifetz School of Business,
Saint Louis University

ITM6400-02 – Applied Business Analytics

Abhimanyu Gupta

Hope Inc. is a non-profit organization that majorly depends on donations. So, with the given data set for 2018, I have used Logistic Regression for binary classification technique to create the lists accordingly to the requirement of Hope Inc.

Steps I used to create prediction-

Step 1- Merged the data sets of 2018 and 2019 taking the 2018 data set as the training data.

Step 2 - Added one column for giving binary classification to the amount_sum column as 0 and 1 according to the requirement.

Step 3 – I have made changes by adding multiple rows in the sheet to train the model properly and to get fair predictions.

Step 4- After the creation of the data set, uploaded the file in Statsbuddy tool to generate the RCode by taking the appropriate dependent variable and independent variables.

Step 3- Run the code in Rcloud to get the results.

1.List of donors who are most likely to donate at least $200

RCODE-

```
dataset <- read.csv("DONOR200.csv")

nrow(dataset)

ncol(dataset)

head(dataset)

trainingsize <- as.integer((997-279)*(1-0/100))

validationsize <- (997-279)-trainingsize

testingsize <- nrow(dataset)-(997-279)

training <- head(dataset,trainingsize)

validation <- tail(head(dataset,trainingsize+validationsize),validationsize)

testing <- tail(dataset,nrow(dataset)-(trainingsize+validationsize))

if(validationsize==0) {validation <- training}

nrow(training)

nrow(validation)

nrow(testing)

model <- glm(Atleast200~attended_event+age+yrs_since_grad,data=training,family="binomial")

summary(model)
```

```
validation$predictedLR <- predict(model,validation,type="response")
predictedclass <- ifelse(validation$predictedLR > 0.5,1,0)
confusionmatrix <- table(factor(predictedclass,levels=0:1),factor(validation$Atleast200,levels=0:1))
confusionmatrix
accuracy <- (1 - mean(predictedclass != validation$Atleast200,na.rm=TRUE))*100
paste(round(accuracy,2),"%",sep="")
predictedLR <- predict(model,testing,type="response")
predictedLR <- ifelse(predictedLR > 0.5,1,0)
testing <- cbind(testing,predictedLR)
write.csv(testing,"DONOR200_Prediction.csv",row.names=FALSE)
```

Percentage Accuracy – 66.57%

Confusion Matrix-

	0	1
0	330	158
1	82	148

- I have used the dependent variable as 'Atleast200' which was having values 1 for amount_sum >= 200 and 0 for amount_sum <200. Independent variables used are attended_event, age and yrs_since_grad.
- As per the results generated through the logistic regression technique, 80 individuals are inclined to donate atleast $200 to Hope Inc. in 2019.

Donating people – 80

Nondonating – 199

I have attached the Excel sheets for both the Data used and predictions of the list of donors who are likely to donate a sum of atleast $200.

2. List of donors who are most likely to donate at least $400

RCODE-

```
dataset <- read.csv("DONOR400.csv")
nrow(dataset)
ncol(dataset)
head(dataset)
```

```
trainingsize <- as.integer((1048-279)*(1-0/100))
validationsize <- (1048-279)-trainingsize
testingsize <- nrow(dataset)-(1048-279)
training <- head(dataset,trainingsize)
validation <- tail(head(dataset,trainingsize+validationsize),validationsize)
testing <- tail(dataset,nrow(dataset)-(trainingsize+validationsize))
if(validationsize==0) {validation <- training}
nrow(training)
nrow(validation)
nrow(testing)
model <- glm(Atleast400~attended_event+age+yrs_since_grad,data=training,family="binomial")
summary(model)
validation$predictedLR <- predict(model,validation,type="response")
predictedclass <- ifelse(validation$predictedLR > 0.5,1,0)
confusionmatrix <- table(factor(predictedclass,levels=0:1),factor(validation$Atleast400,levels=0:1))
confusionmatrix
accuracy <- (1 - mean(predictedclass != validation$Atleast400,na.rm=TRUE))*100
paste(round(accuracy,2),"%",sep="")
predictedLR <- predict(model,testing,type="response")
predictedLR <- ifelse(predictedLR > 0.5,1,0)
testing <- cbind(testing,predictedLR)
write.csv(testing,"DONOR400_Prediction.csv",row.names=FALSE)
```

Percentage Accuracy- "79.97%"

Confusion Matrix

```
    0   1
0 352  74
1  80 263
```

- I have used the dependent variable as ‘Atleast400’ which was having values 1 for amount_sum >= 400 and 0 for amount_sum <400. Independent variables used are attended_event, age and yrs_since_grad.
- As per the results generated through the logistic regression technique, 87 individuals are inclined to donate atleast $400 to Hope Inc. in 2019.

Donating people – 87

Nondonating – 192

I have attached the Excel sheets for both the Data used and predictions of the list of donors who are likely to donate a sum of atleast $400.

3.List of donors who are most likely to donate at least $800

RCODE-

```
dataset <- read.csv("DONOR800.csv")

nrow(dataset)

ncol(dataset)

head(dataset)

trainingsize <- as.integer((1045-279)*(1-0/100))

validationsize <- (1045-279)-trainingsize

testingsize <- nrow(dataset)-(1045-279)

training <- head(dataset,trainingsize)

validation <- tail(head(dataset,trainingsize+validationsize),validationsize)

testing <- tail(dataset,nrow(dataset)-(trainingsize+validationsize))

if(validationsize==0) {validation <- training}

nrow(training)

nrow(validation)

nrow(testing)

model <- glm(Atleast800~gender+marital_status+has_children+spouse_in_db+attended_event+age+yrs_since_grad,data=training,family="binomial")

summary(model)

validation$predictedLR <- predict(model,validation,type="response")

predictedclass <- ifelse(validation$predictedLR > 0.5,1,0)

confusionmatrix <- table(factor(predictedclass,levels=0:1),factor(validation$Atleast800,levels=0:1))

confusionmatrix

accuracy <- (1 - mean(predictedclass != validation$Atleast800,na.rm=TRUE))*100

paste(round(accuracy,2),"%",sep="")
```

```
predictedLR <- predict(model,testing,type="response")

predictedLR <- ifelse(predictedLR > 0.5,1,0)

testing <- cbind(testing,predictedLR)

write.csv(testing,"DONOR800_Prediction.csv",row.names=FALSE)
```

Percentage Accuracy- "86.68%"

Confusion Matrix

```
     0   1
0  384  24
1  78 280
```

- I have used the dependent variable as 'Atleast800' which was having values 1 for amount_sum >= 800 and 0 for amount_sum<800. Independent variables used are gender, marital_status, has_children, spouse_in_db, attended_event, age and yrs_since_grad.
- As per the results generated through the logistic regression technique,62 individuals are inclined to donate atleast $800 to Hope Inc. in 2019.

Donating people – 62

Nondonating – 217

I have attached the Excel sheets for both the Data used and predictions of the list of donors who are likely to donate a sum of atleast $800.

Another Approach-

I tried predicting the data without merging the two files. Firstly, I trained the existing 2018 data and used that model to make predictions for 2019 donors. Below is the Rcode which I used and tried writing the code to read the new data file.

```
dataset <- read.csv("DONOR800.csv")

nrow(dataset)

ncol(dataset)

head(dataset)

trainingsize <- as.integer((500-440)*(1-0/100))

validationsize <- (500-440)-trainingsize

testingsize <- nrow(dataset)-(500-440)

training <- head(dataset,trainingsize)
```

```
validation <- tail(head(dataset,trainingsize+validationsize),validationsize)

testing <- tail(dataset,nrow(dataset)-(trainingsize+validationsize))

if(validationsize==0) {validation <- training}

nrow(training)

nrow(validation)

nrow(testing)

model <- glm(Atleast800~gender+marital_status+has_children+spouse_in_db+attended_event+age+yrs_since_grad,data=training,family="binomial")

summary(model)

validation$predictedLR <- predict(model,validation,type="response")

predictedclass <- ifelse(validation$predictedLR > 0.5,1,0)

confusionmatrix <- table(factor(predictedclass,levels=0:1),factor(validation$Atleast800,levels=0:1))

confusionmatrix

accuracy <- (1 - mean(predictedclass != validation$Atleast800,na.rm=TRUE))*100

paste(round(accuracy,2),"%",sep="")

predictedLR <- predict(model,testing,type="response")

predictedLR <- ifelse(predictedLR > 0.5,1,0)

testing <- cbind(testing,predictedLR)

write.csv(testing,"DONOR800_Prediction.csv",row.names=FALSE)

> d <- read.csv("Donor 2019data.csv")

> p <- predict(model,d,type="response")

> p <- ifelse(p > 0.5,1,0)

> d <- cbind(d,p)

> View(d)

> write.csv(d,"DONOR800_Prediction2019.csv",row.names=FALSE)
```

Here d = 2019 data set, p = prediction of the 2019 data set

Accuracy- 91%
I have attached the DONOR800_Prediction2019 Excel sheet for predictions of the list of donors who are likely to donate a sum of atleast $800
Donating people – 100, Nondonating – 179