DSA211 - Statistical Learning With R

Project Part 2 Q2 Report

Section Number: G1

Instructor: Professor Goh Jing Rong

Toh Jing Lin Cheryl	**01430750**	**cheryl.toh.2021**
Foo Chuan Wei	**01394670**	**cwfoo.2021**
Nguyen Hanh Trang	**01422031**	**htnguyen.2021**
Seah Li Ping Megan	**01395272**	**megan.seah.2019**
Sharafinaz Binte Shawal	**01443439**	**sharafinazs.2021**

1.0 Test Model

All R Outputs and Inputs may be found in the Appendix.

From our Project Part 2 Q1 Report, the best model we will use for the testing is **M5** as shown in Figure 1.

(Intercept)	Cards	GenderFemale	Income
7.339071e+02	-8.612623e+01	1.028961e+02	1.201578e-01
I(Income^2)	Rating:GenderFemale	Cards:GenderFemale	Rating:Income
3.749728e-04	1.749931e-01	6.591505e+01	1.014376e-05
Rating:I(Income^2)	GenderFemale:I(Income^2)	Rating:Cards:GenderFemale	Rating:Cards:I(Income^2)
4.319330e-08	-7.447831e-05	3.387013e-03	1.377175e-09
Rating:GenderFemale:I(Income^2)			
-2.260941e-08			

Figure 1: Best Model for lasso regression (**M5**)

2.0 Test Result

The test MSE value obtained was 201447.7.

Course Name: **DSA211-G1-Statistical Learning with R**
Instructor(s): _, **Dr GOH Jing Rong**
Student Name: _, **FOO CHUAN WEI**
Campus ID: **01394670**
Date/Time: **02 Feb 2024 14:23**

Name	Marks	Grade	Weight
Class Participation		A+	
Group Project		A+	
Assignment		A+	

APPENDIX

1.0 Input

```
#BEST MODEL FROM PROJECT PART TWO Q1
library(leaps)
library(glmnet)
RNGkind(sample.kind = "Rounding")
set.seed(123)
bank = read.csv("Bank2023P.csv", stringsAsFactors = TRUE)
attach(bank)

train <- sample(1:nrow(bank),800)
test <- -train

bank.train = bank[train,]
bank.test = bank[test,]

train.x = model.matrix(Balance~Rating*Cards*Gender + Rating*Income*Cards +
Rating*I(Income^2)*Cards + Rating*Income*Gender+Rating*I(Income^2)*Gender
+ Income*Cards*Gender + I(Income^2)*Cards*Gender, data = bank.train)
train.y = bank.train$Balance

test.x = model.matrix(Balance~Rating*Cards*Gender + Rating*Income*Cards +
Rating*I(Income^2)*Cards + Rating*Income*Gender+Rating*I(Income^2)*Gender
+ Income*Cards*Gender + I(Income^2)*Cards*Gender,bank.test)
test.y = bank.test$Balance

lasso.mod <- glmnet(train.x, train.y, alpha=1)
lassocv.out <- cv.glmnet(train.x, train.y, alpha=1)

lassolam <- lassocv.out$lambda.min
lassolam
```

```
lasso.pred <- predict(lasso.mod, s=lassolam, newx=test.x)
mean((lasso.pred-test.y)^2)

x = model.matrix(Balance~Rating*Cards*Gender + Rating*Income*Cards +
Rating*I(Income^2)*Cards + Rating*Income*Gender+Rating*I(Income^2)*Gender
+ Income*Cards*Gender + I(Income^2)*Cards*Gender, bank)
y = bank$Balance

out.lasso <- glmnet(x,y,alpha=1)
lasso.coef <- predict(out.lasso, type="coefficients", s=lassolam)[1:21,]

lasso.coef[lasso.coef!=0]
```

1.0 Output

```
[1] 4.710897
[1] 201781.3
                  (Intercept)                         Cards
GenderFemale
                 7.339071e+02                -8.612623e+01
1.028961e+02
                       Income                    I(Income^2)
Rating:GenderFemale
                 1.201578e-01                 3.749728e-04
1.749931e-01
          Cards:GenderFemale                Rating:Income
Rating:I(Income^2)
                 6.591505e+01                 1.014376e-05
4.319330e-08
     GenderFemale:I(Income^2)       Rating:Cards:GenderFemale
Rating:Cards:I(Income^2)
                -7.447831e-05                 3.387013e-03
1.377175e-09
Rating:GenderFemale:I(Income^2)
                -2.260941e-08
```

2.0 Input

```
RNGkind(sample.kind = "Rounding")
set.seed(123)

testbank <- read.csv("Bank2023testP.csv", stringsAsFactors = TRUE)
test.xx <- model.matrix(Balance~Rating*Cards*Gender + Rating*Income*Cards +
Rating*I(Income^2)*Cards + Rating*Income*Gender+Rating*I(Income^2)*Gender
+ Income*Cards*Gender + I(Income^2)*Cards*Gender, data=testbank)
test.yy <- testbank$Balance
yhatall <- predict(out.lasso, newx=test.xx, s=lassolam) #uses lowest lasso from
mean((test.yy-yhatall)^2)
```

2.0 Output

```
[1] 201447.7
```