Stat 412

Yolanda Jin

24/11/2021

NA data (Income NA), Minority, Majority Groups

- Added column for NA_indicator
- split groups to minority vs majority group
- Q1: do we want NA to be a separate group? based on EDA, might or might not do so

```
# Read Credit Scoring Data Training Set
cs_train = read.csv("cs-training.csv")
cs_train[1:10,]
```

```
##     X SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines age
## 1   1                1                              0.7661266  45
## 2   2                0                              0.9571510  40
## 3   3                0                              0.6581801  38
## 4   4                0                              0.2338098  30
## 5   5                0                              0.9072394  49
## 6   6                0                              0.2131787  74
## 7   7                0                              0.3056825  57
## 8   8                0                              0.7544636  39
## 9   9                0                              0.1169506  27
## 10 10                0                              0.1891691  57
##    NumberOfTime30.59DaysPastDueNotWorse    DebtRatio MonthlyIncome
## 1                                     2 8.029821e-01          9120
## 2                                     0 1.218762e-01          2600
## 3                                     1 8.511338e-02          3042
## 4                                     0 3.604968e-02          3300
## 5                                     1 2.492570e-02         63588
## 6                                     0 3.756070e-01          3500
## 7                                     0 5.710000e+03            NA
## 8                                     0 2.099400e-01          3500
## 9                                     0 4.600000e+01            NA
## 10                                    0 6.062909e-01         23684
##    NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
## 1                               13                       0
## 2                                4                       0
## 3                                2                       1
## 4                                5                       0
## 5                                7                       0
## 6                                3                       0
## 7                                8                       0
```

```
## 8                                8                       0
## 9                                2                       0
## 10                               9                       0
##    NumberRealEstateLoansOrLines NumberOfTime60.89DaysPastDueNotWorse
## 1                             6                                    0
## 2                             0                                    0
## 3                             0                                    0
## 4                             0                                    0
## 5                             1                                    0
## 6                             1                                    0
## 7                             3                                    0
## 8                             0                                    0
## 9                             0                                    0
## 10                            4                                    0
##    NumberOfDependents
## 1                   2
## 2                   1
## 3                   0
## 4                   0
## 5                   0
## 6                   1
## 7                   0
## 8                   0
## 9                  NA
## 10                  2
```

```
is.na(cs_train[7,]$MonthlyIncome)
```

```
## [1] TRUE
```

```
## 0. NA Monthly Income

# Add col to indicate whether monthly Income is NA or not
cs_train$NA_Indicator <- 0  # Set all NA_Indicator to zero
cs_train$NA_Indicator[is.na(cs_train$MonthlyIncome)] <- 1  # Change NA income indexes to 1
head(cs_train[cs_train$NA_Indicator==1,])  # Check the ones indicated as NA
```

```
##     X SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines age
## 7   7                0                           0.30568247  57
## 9   9                0                           0.11695064  27
## 17 17                0                           0.06108612  78
## 33 33                0                           0.08341801  62
## 42 42                0                           0.07289757  81
## 53 53                0                           0.99999990  62
##    NumberOfTime30.59DaysPastDueNotWorse DebtRatio MonthlyIncome
## 7                                    0      5710            NA
## 9                                    0        46            NA
## 17                                   0      2058            NA
## 33                                   0       977            NA
## 42                                   0        75            NA
## 53                                   0         0            NA
##    NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
## 7                                8                       0
```

```
## 9                                2                        0
## 17                              10                        0
## 33                               6                        0
## 42                               7                        0
## 53                               1                        0
##    NumberRealEstateLoansOrLines NumberOfTime60.89DaysPastDueNotWorse
## 7                             3                                     0
## 9                             0                                     0
## 17                            2                                     0
## 33                            1                                     0
## 42                            0                                     0
## 53                            0                                     0
##    NumberOfDependents NA_Indicator
## 7                   0            1
## 9                  NA            1
## 17                  0            1
## 33                  0            1
## 42                  0            1
## 53                  0            1
```

```
# If we want to split NA to another group
#cs_train_NA <- cs_train[cs_train$MonthlyIncome==NA,]
#head(cs_train_NA)



## 1. Separate minority data vs majority data vs NA data total 150k
cs_train_min <- cs_train[cs_train$SeriousDlqin2yrs==1,]  # 10,026 obs
cs_train_maj <- cs_train[cs_train$SeriousDlqin2yrs==0,]  # 139,974 obs
```

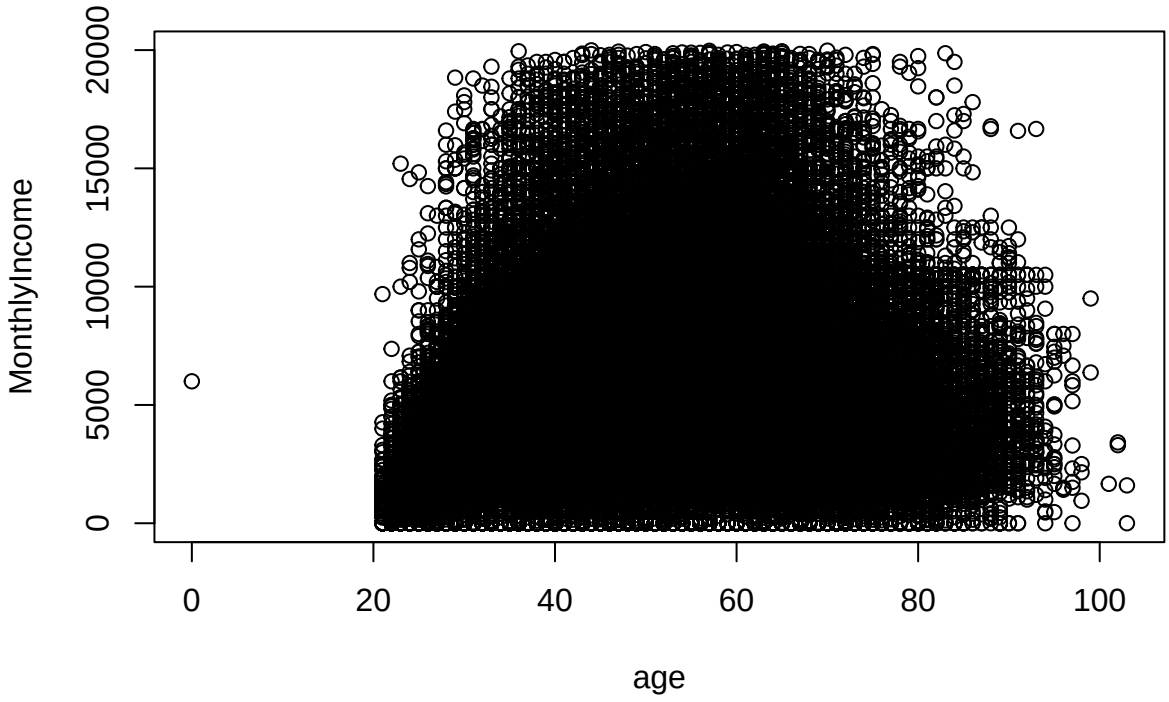
EDA

- monthly income = 0
- 30k monthly income = NA
-

Dependent = NA 4k

- age 0 remove - only 1 point
- age very old
- group by age group etc
- 13 - 101 or older (maybe cut at 100)
- 80 or more

```
#plot(SeriousDlqin2yrs ~ age, data = cs_train_maj)
cs_train_maj_g1 <- cs_train_maj[cs_train_maj$MonthlyIncome<500000,] # less than 500k monthly income
cs_train_maj_g1 <- cs_train_maj[cs_train_maj$MonthlyIncome<20000,] # less than 20k monthly income
plot(MonthlyIncome ~ age, data = cs_train_maj_g1)
```



Question/Goal

- Comparing our model performance against paper's model
- Most important factors

Ensemble Learning

- Lasso Ensemble Algorithm
- Aggregating base learner: Weighted base=learner

Balancing Data

- Use clustering and pick one sub-group from majority
  - can try Age/Income
  - try Income/debt ratio
- Use bagging algorithm to create more minority data
- Use NA indicator 0 or 1 (maybe use mean/median)

```
# Cluster Grouping Majority Data (Try to cluster data by age/monthly income)
set.seed(1) # for reproducibility

head(cs_train_maj)
```

```
##   X SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines age
## 2 2                0                            0.9571510  40
## 3 3                0                            0.6581801  38
## 4 4                0                            0.2338098  30
## 5 5                0                            0.9072394  49
## 6 6                0                            0.2131787  74
## 7 7                0                            0.3056825  57
##   NumberOfTime30.59DaysPastDueNotWorse    DebtRatio MonthlyIncome
## 2                                    0 1.218762e-01          2600
## 3                                    1 8.511338e-02          3042
## 4                                    0 3.604968e-02          3300
## 5                                    1 2.492570e-02         63588
## 6                                    0 3.756070e-01          3500
## 7                                    0 5.710000e+03            NA
##   NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
## 2                               4                       0
## 3                               2                       1
## 4                               5                       0
## 5                               7                       0
## 6                               3                       0
## 7                               8                       0
##   NumberRealEstateLoansOrLines NumberOfTime60.89DaysPastDueNotWorse
## 2                            0                                    0
## 3                            0                                    0
## 4                            0                                    0
## 5                            1                                    0
## 6                            1                                    0
## 7                            3                                    0
##   NumberOfDependents NA_Indicator
## 2                  1            0
## 3                  0            0
## 4                  0            0
## 5                  0            0
## 6                  1            0
## 7                  0            1
```

```
# Test with smaller group based on monthly income range
cs_train_maj_g1 <- cs_train_maj[cs_train_maj$NA_Indicator==0,] # Filter out NA monthly income first
cs_train_maj_g1 <- cs_train_maj_g1[cs_train_maj_g1$MonthlyIncome<20000,]  #38035 obs
head(cs_train_maj_g1)
```

```
##     X SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines age
## 2   2                0                            0.9571510  40
## 3   3                0                            0.6581801  38
## 4   4                0                            0.2338098  30
## 6   6                0                            0.2131787  74
## 8   8                0                            0.7544636  39
```

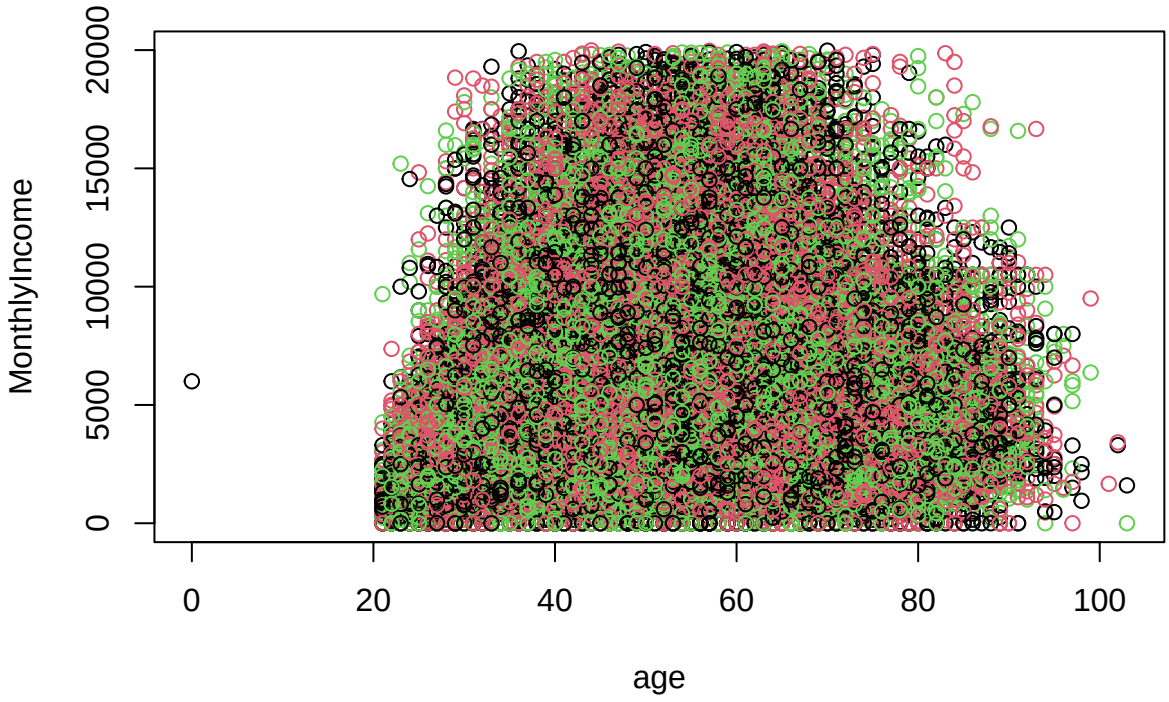
```
## 11 11                0                            0.6442260  30
##    NumberOfTime30.59DaysPastDueNotWorse  DebtRatio MonthlyIncome
## 2                                     0 0.12187620          2600
## 3                                     1 0.08511338          3042
## 4                                     0 0.03604968          3300
## 6                                     0 0.37560697          3500
## 8                                     0 0.20994002          3500
## 11                                    0 0.30947621          2500
##    NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
## 2                                4                       0
## 3                                2                       1
## 4                                5                       0
## 6                                3                       0
## 8                                8                       0
## 11                               5                       0
##    NumberRealEstateLoansOrLines NumberOfTime60.89DaysPastDueNotWorse
## 2                             0                                     0
## 3                             0                                     0
## 4                             0                                     0
## 6                             1                                     0
## 8                             0                                     0
## 11                            0                                     0
##    NumberOfDependents NA_Indicator
## 2                   1            0
## 3                   0            0
## 4                   0            0
## 6                   1            0
## 8                   0            0
## 11                  0            0
```

```
# Create a dat with the two predictors of interest
dat <- cs_train_maj_g1[,c(4,7)]  # Age and MonthlyIncome
head(dat)
```

```
##    age MonthlyIncome
## 2   40          2600
## 3   38          3042
## 4   30          3300
## 6   74          3500
## 8   39          3500
## 11  30          2500
```

```
n_maj <- nrow(dat) # get number of rows

# Initial assignments to three groups that will need to update
assignments <- factor(sample(c(1,2,3), n_maj, replace = TRUE))
#plot(dat, col=assignments, xlim = c(0,110), asp=1)
plot(dat, col=assignments)
```

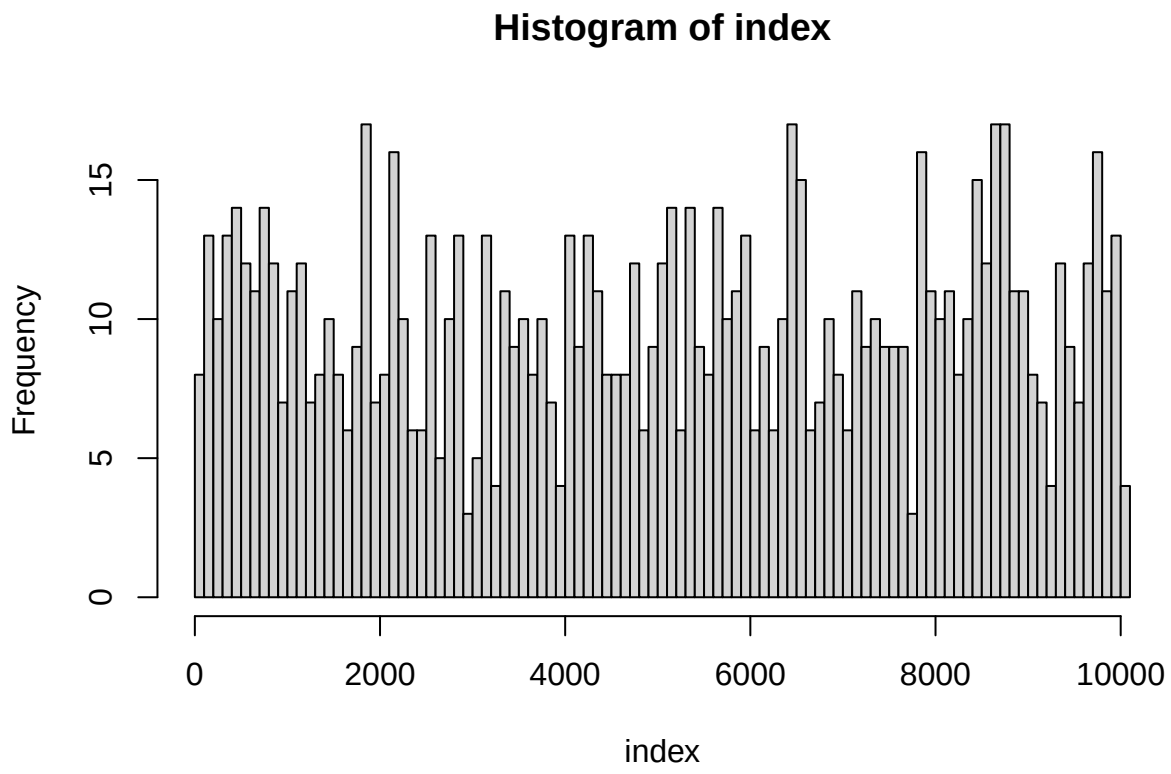



```
# Boostrapping Minority Data
set.seed(1) # for reproducibility
# set number of minority data to reproduce
n_add <- 1000

n_min <- nrow(cs_train_min)
n_min
```

```
## [1] 10026
```

```
index <- sample(n_min, n_add, replace = TRUE)
#plot(density(index),  main="")  # show density curve of the index we randomized
hist(index, breaks = 100)
```




```
min(index)
```

```
## [1] 27
```

```
max(index)
```

```
## [1] 10014
```

```
length(index)
```

```
## [1] 1000
```

```
# We add the additional data for future analysis
cs_train_min_add <- cs_train_min[index,]
head(cs_train_min_add)
```

```
##             X SeriousDlqin2yrs RevolvingUtilizationOfUnsecuredLines age
## 15807   15807                1                            0.99999990  46
## 120332 120332                1                            0.04713019  45
## 71375   71375                1                            0.71069704  52
## 145782 145782                1                            0.99999990  52
## 127189 127189                1                            1.25614754  56
## 60636   60636                1                            0.41318696  35
```

```
##        NumberOfTime30.59DaysPastDueNotWorse    DebtRatio MonthlyIncome
## 15807                                     0    0.6327286          2700
## 120332                                    0    0.5847736          7000
## 71375                                     0    1.1807549          3761
## 145782                                    0    0.0000000          3200
## 127189                                    3    0.1938939          5600
## 60636                                     1 1203.0000000            NA
##        NumberOfOpenCreditLinesAndLoans NumberOfTimes90DaysLate
## 15807                                3                       1
## 120332                               8                       0
## 71375                               11                       0
## 145782                               0                       0
## 127189                               7                       2
## 60636                               10                       0
##        NumberRealEstateLoansOrLines NumberOfTime60.89DaysPastDueNotWorse
## 15807                             1                                    1
## 120332                            2                                    0
## 71375                             2                                    0
## 145782                            0                                    0
## 127189                            0                                    1
## 60636                             0                                    0
##        NumberOfDependents NA_Indicator
## 15807                   0            0
## 120332                  0            0
## 71375                   2            0
## 145782                  2            0
## 127189                  0            0
## 60636                   0            1
```

Which models to try

- Logistic regression (compare the different link functions)
- look maybe merge Lasso with logistic regression
- RF

Evaluating Model/Comparing results

- AUC