LoanPredictions

Manish Sihag

March 9, 2017

```
train = read.csv("train.txt", sep = ",", header = T, na.strings = c("", NA))
test = read.csv("test.txt", sep = ",", header = T, na.strings = c("", NA))
```

```
summary(train)
```

```
##      Loan_ID       Gender    Married    Dependents        Education
##  LP001002:  1   Female:112   No  :213   0   :345   Graduate    :480
##  LP001003:  1   Male  :489   Yes :398   1   :102   Not Graduate:134
##  LP001005:  1   NA's  : 13   NA's:  3   2   :101
##  LP001006:  1                           3+  : 51
##  LP001008:  1                           NA's: 15
##  LP001011:  1
##  (Other) :608
##  Self_Employed ApplicantIncome CoapplicantIncome   LoanAmount
##  No  :500      Min.   :  150   Min.   :    0     Min.   :  9.0
##  Yes : 82      1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0
##  NA's: 32      Median : 3812   Median : 1188     Median :128.0
##                Mean   : 5403   Mean   : 1621     Mean   :146.4
##                3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:168.0
##                Max.   :81000   Max.   :41667     Max.   :700.0
##                                                  NA's   :22
##  Loan_Amount_Term Credit_History     Property_Area Loan_Status
##  Min.   : 12      Min.   :0.0000   Rural    :179   N:192
##  1st Qu.:360      1st Qu.:1.0000   Semiurban:233   Y:422
##  Median :360      Median :1.0000   Urban    :202
##  Mean   :342      Mean   :0.8422
##  3rd Qu.:360      3rd Qu.:1.0000
##  Max.   :480      Max.   :1.0000
##  NA's   :14       NA's   :50
```

```
str(train)
```

```
## 'data.frame':    614 obs. of  13 variables:
##  $ Loan_ID          : Factor w/ 614 levels "LP001002","LP001003",..: 1 2 3 4 5 6 7 8 9 10
 ...
##  $ Gender           : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Married          : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 2 2 2 2 2 ...
##  $ Dependents       : Factor w/ 4 levels "0","1","2","3+": 1 2 1 1 1 3 1 4 3 2 ...
##  $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 2 1 1 1
 ...
##  $ Self_Employed    : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 1 1 1 1 ...
##  $ ApplicantIncome  : int  5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...
##  $ CoapplicantIncome: num  0 1508 0 2358 0 ...
##  $ LoanAmount       : int  NA 128 66 120 141 267 95 158 168 349 ...
##  $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 360 360 ...
##  $ Credit_History   : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ Property_Area    : Factor w/ 3 levels "Rural","Semiurban",..: 3 1 3 3 3 3 3 2 3 2 ...
##  $ Loan_Status      : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 1 2 1 ...
```

```
train$Credit_History = as.factor(train$Credit_History)
summary(test)
```

```
##      Loan_ID       Gender    Married   Dependents        Education
##  LP001015:  1   Female: 70   No :134   0   :200   Graduate    :283
##  LP001022:  1   Male  :286   Yes:233   1   : 58   Not Graduate: 84
##  LP001031:  1   NA's  : 11             2   : 59
##  LP001035:  1                          3+  : 40
##  LP001051:  1                          NA's: 10
##  LP001054:  1
##  (Other) :361
##  Self_Employed ApplicantIncome CoapplicantIncome   LoanAmount
##  No  :307      Min.   :    0   Min.   :    0     Min.   : 28.0
##  Yes : 37      1st Qu.: 2864   1st Qu.:    0     1st Qu.:100.2
##  NA's: 23      Median : 3786   Median : 1025     Median :125.0
##                Mean   : 4806   Mean   : 1570     Mean   :136.1
##                3rd Qu.: 5060   3rd Qu.: 2430     3rd Qu.:158.0
##                Max.   :72529   Max.   :24000     Max.   :550.0
##                                                  NA's   :5
##  Loan_Amount_Term Credit_History     Property_Area
##  Min.   :  6.0    Min.   :0.0000   Rural    :111
##  1st Qu.:360.0    1st Qu.:1.0000   Semiurban:116
##  Median :360.0    Median :1.0000   Urban    :140
##  Mean   :342.5    Mean   :0.8254
##  3rd Qu.:360.0    3rd Qu.:1.0000
##  Max.   :480.0    Max.   :1.0000
##  NA's   :6        NA's   :29
```

```
str(test)
```

```
## 'data.frame':    367 obs. of  12 variables:
##  $ Loan_ID          : Factor w/ 367 levels "LP001015","LP001022",..: 1 2 3 4 5 6 7 8 9 10
 ...
##  $ Gender           : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 1 2 2 2 ...
##  $ Married          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 2 1 ...
##  $ Dependents       : Factor w/ 4 levels "0","1","2","3+": 1 2 3 3 1 1 2 3 3 1 ...
##  $ Education        : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 1 2 2 2 2 1 2
 ...
##  $ Self_Employed    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 NA 1 ...
##  $ ApplicantIncome  : int  5720 3076 5000 2340 3276 2165 2226 3881 13633 2400 ...
##  $ CoapplicantIncome: int  0 1500 1800 2546 0 3422 0 0 0 2400 ...
##  $ LoanAmount       : int  110 126 208 100 78 152 59 147 280 123 ...
##  $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 240 360 ...
##  $ Credit_History   : int  1 1 1 NA 1 1 1 0 1 1 ...
##  $ Property_Area    : Factor w/ 3 levels "Rural","Semiurban",..: 3 3 3 3 3 3 2 1 3 2 ...
```

```
test$Credit_History = as.factor(test$Credit_History)
train1 = train
train1$Loan_Status = NULL
total = rbind(train1,test)
total$Loan_ID=NULL
rm(train1)
summary(total)
```

```
##     Gender    Married    Dependents        Education   Self_Employed
##  Female:182   No  :347   0   :545   Graduate    :763   No  :807
##  Male  :775   Yes :631   1   :160   Not Graduate:218   Yes :119
##  NA's  : 24   NA's:  3   2   :160                      NA's: 55
##                          3+  : 91
##                          NA's: 25
##
##
##  ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term
##  Min.   :    0   Min.   :    0     Min.   :  9.0   Min.   :  6.0
##  1st Qu.: 2875   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360.0
##  Median : 3800   Median : 1110     Median :126.0   Median :360.0
##  Mean   : 5180   Mean   : 1602     Mean   :142.5   Mean   :342.2
##  3rd Qu.: 5516   3rd Qu.: 2365     3rd Qu.:162.0   3rd Qu.:360.0
##  Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480.0
##                                    NA's   :27      NA's   :20
##  Credit_History   Property_Area
##  0   :148       Rural    :290
##  1   :754       Semiurban:349
##  NA's: 79       Urban    :342
##
##
##
##
```

```
library(missForest)
imputed = missForest(total)
```

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!
```

```
summary(imputed)
```

```
##          Length Class      Mode
## ximp     11     data.frame list
## OOBerror  2     -none-     numeric
```

```
totalNew = imputed$ximp
summary(totalNew)
```

```
##     Gender    Married   Dependents        Education   Self_Employed
##  Female:191   No :349   0 :555     Graduate    :763   No :859
##  Male  :790   Yes:632   1 :162     Not Graduate:218   Yes:122
##                         2 :169
##                         3+: 95
##
##
##  ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term
##  Min.   :    0   Min.   :    0     Min.   :  9.0   Min.   :  6.0
##  1st Qu.: 2875   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360.0
##  Median : 3800   Median : 1110     Median :126.0   Median :360.0
##  Mean   : 5180   Mean   : 1602     Mean   :142.7   Mean   :342.1
##  3rd Qu.: 5516   3rd Qu.: 2365     3rd Qu.:162.0   3rd Qu.:360.0
##  Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480.0
##  Credit_History   Property_Area
##  0:153          Rural    :290
##  1:828          Semiurban:349
##                 Urban    :342
##
##
##
```

```
trainNew = totalNew[1:614,]
trainNew$Loan_Status = train$Loan_Status
testNew = totalNew[615:981,]
```

```
summary(trainNew)
```

```
##     Gender    Married   Dependents        Education   Self_Employed
##  Female:115   No :215   0 :350     Graduate    :480   No :530
##  Male  :499   Yes:399   1 :104     Not Graduate:134   Yes: 84
##                         2 :106
##                         3+: 54
##
##
##  ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term
##  Min.   :  150   Min.   :    0     Min.   :  9.0   Min.   : 12
##  1st Qu.: 2878   1st Qu.:    0     1st Qu.:100.0   1st Qu.:360
##  Median : 3812   Median : 1188     Median :128.0   Median :360
##  Mean   : 5403   Mean   : 1621     Mean   :146.3   Mean   :342
##  3rd Qu.: 5795   3rd Qu.: 2297     3rd Qu.:166.8   3rd Qu.:360
##  Max.   :81000   Max.   :41667     Max.   :700.0   Max.   :480
##  Credit_History   Property_Area Loan_Status
##  0: 93          Rural    :179   N:192
##  1:521          Semiurban:233   Y:422
##                 Urban    :202
##
##
##
```

```
summary(testNew)
```

```
##     Gender    Married   Dependents        Education   Self_Employed
##  Female: 76   No :134   0 :205     Graduate    :283   No :329
##  Male  :291   Yes:233   1 : 58     Not Graduate: 84   Yes: 38
##                         2 : 63
##                         3+: 41
##
##
##  ApplicantIncome CoapplicantIncome   LoanAmount    Loan_Amount_Term
##  Min.   :    0   Min.   :    0     Min.   : 28.0   Min.   :  6.0
##  1st Qu.: 2864   1st Qu.:    0     1st Qu.:100.5   1st Qu.:360.0
##  Median : 3786   Median : 1025     Median :125.0   Median :360.0
##  Mean   : 4806   Mean   : 1570     Mean   :136.8   Mean   :342.3
##  3rd Qu.: 5060   3rd Qu.: 2430     3rd Qu.:159.5   3rd Qu.:360.0
##  Max.   :72529   Max.   :24000     Max.   :550.0   Max.   :480.0
##  Credit_History   Property_Area
##  0: 60          Rural    :111
##  1:307          Semiurban:116
##                 Urban    :140
##
##
##
```

```
library(randomForest)
forestModel = randomForest(Loan_Status~., data = trainNew, ntree = 200)
pred = predict(forestModel, newdata = testNew)
prediction = as.matrix(test$Loan_ID,367,1)
colnames(prediction) = "Loan_ID"
prediction = as.data.frame(prediction)
prediction$Loan_Status = pred
write.csv(prediction, "submission.csv")
```