Logistic_Regression_without_cross-validation.R

jas

Fri Sep 09 18:48:51 2016

```
drive <- "D:/R/Analytics vidhya/Hackathon/Loan Prediction"
setwd(drive)
train <- read.csv("TrainD.csv", header = TRUE, stringsAsFactors = FALSE)
test <- read.csv("TestD.csv", header = TRUE, stringsAsFactors = FALSE)

head(train)
```

```
##    Loan_ID ApplicantIncome CoapplicantIncome LoanAmount Loan_Status Gender
## 1 LP001002            5849                0   4.959423           1      1
## 2 LP001003            4583             1508   4.852030           0      1
## 3 LP001005            3000                0   4.189655           1      1
## 4 LP001006            2583             2358   4.787492           1      1
## 5 LP001008            6000                0   4.948760           1      1
## 6 LP001011            5417             4196   5.587249           1      1
##   Married Dependents0 Dependents1 Dependents2 Dependents3. Education
## 1       0           1           0           0           0         1
## 2       1           0           1           0           0         1
## 3       1           1           0           0           0         1
## 4       1           1           0           0           0         0
## 5       0           1           0           0           0         1
## 6       1           0           0           1           0         1
##   Self_Employed Credit_History Property_AreaRural Property_AreaSemiurban
## 1             0              1                   0                      0
## 2             0              1                   1                      0
## 3             1              1                   0                      0
## 4             0              1                   0                      0
## 5             0              1                   0                      0
## 6             1              1                   0                      0
##   Property_AreaUrban TotalIncome ApplicantIncome_Zero
## 1                  1    8.674026                    0
## 2                  0    8.714568                    0
## 3                  1    8.006368                    0
## 4                  1    8.505323                    0
## 5                  1    8.699515                    0
## 6                  1    9.170872                    0
##   DebtRatio_Mainapplicant DebtRatio_TotalIncome CoappIncGApplinc
## 1              0.02436511            0.02436511                0
## 2              0.02792930            0.02101461                0
## 3              0.02200000            0.02200000                0
## 4              0.04645761            0.02428658                0
## 5              0.02350000            0.02350000                0
## 6              0.04928927            0.02777489                0
```

```
##   Loan_Amount_Term10.12.mnths Loan_Amount_Term112.120.mnths
## 1                           0                             0
## 2                           0                             0
## 3                           0                             0
## 4                           0                             0
## 5                           0                             0
## 6                           0                             0
##   Loan_Amount_Term1120.240.mnths Loan_Amount_Term1240.360.mnths
## 1                              0                              1
## 2                              0                              1
## 3                              0                              1
## 4                              0                              1
## 5                              0                              1
## 6                              0                              1
##   Loan_Amount_Term1360..Mnths Gender_Married Married_Dep0 Married_Dep1
## 1                           0              0            0            0
## 2                           0              1            0            1
## 3                           0              1            1            0
## 4                           0              1            1            0
## 5                           0              0            0            0
## 6                           0              1            0            0
##   Married_Dep2 Married_Dep3
## 1            0            0
## 2            0            0
## 3            0            0
## 4            0            0
## 5            0            0
## 6            1            0

######USing all parameters for prediction#######
logistic <- glm(Loan_Status ~ ., data=train[,!colnames(train) %in%
c("Loan_ID")],
                family='binomial')
summary(logistic)

##
## Call:
## glm(formula = Loan_Status ~ ., family = "binomial", data = train[,
##     !colnames(train) %in% c("Loan_ID")])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2762  -0.3805   0.5052   0.6793   2.5280
##
## Coefficients: (5 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   3.181e+00  5.699e+00   0.558  0.57673
## ApplicantIncome              -4.079e-06  3.956e-05  -0.103  0.91788
## CoapplicantIncome            -7.477e-05  6.455e-05  -1.158  0.24671
## LoanAmount                    8.610e-01  6.566e-01   1.311  0.18977
```

```
## Gender                           -2.394e-01  3.732e-01  -0.641  0.52124
## Married                          -4.839e-01  1.411e+00  -0.343  0.73168
## Dependents0                      -9.349e-01  1.251e+00  -0.747  0.45486
## Dependents1                      -1.581e+00  1.310e+00  -1.206  0.22764
## Dependents2                       3.801e-01  1.637e+00   0.232  0.81638
## Dependents3.                             NA         NA      NA       NA
## Education                         3.277e-01  2.707e-01   1.210  0.22611
## Self_Employed                     6.284e-02  3.229e-01   0.195  0.84570
## Credit_History                    3.956e+00  4.107e-01   9.631  < 2e-16 ***
## Property_AreaRural               -2.164e-01  2.704e-01  -0.800  0.42362
## Property_AreaSemiurban            7.574e-01  2.781e-01   2.723  0.00646 **
## Property_AreaUrban                       NA         NA      NA       NA
## TotalIncome                      -1.047e+00  8.768e-01  -1.194  0.23258
## ApplicantIncome_Zero                     NA         NA      NA       NA
## DebtRatio_Mainapplicant          -1.157e+00  2.491e+00  -0.465  0.64218
## DebtRatio_TotalIncome            -6.939e+01  3.173e+01  -2.187  0.02873 *
## CoappIncGApplinc                  3.132e-01  4.122e-01   0.760  0.44727
## Loan_Amount_Term10.12.mnths       1.327e+01  5.354e+02   0.025  0.98022
## Loan_Amount_Term112.120.mnths     7.493e-01  9.485e-01   0.790  0.42954
## Loan_Amount_Term1120.240.mnths    1.522e+00  7.754e-01   1.963  0.04961 *
## Loan_Amount_Term1240.360.mnths    1.320e+00  6.429e-01   2.053  0.04008 *
## Loan_Amount_Term1360..Mnths              NA         NA      NA       NA
## Gender_Married                    2.250e-01  6.487e-01   0.347  0.72865
## Married_Dep0                      8.338e-01  1.346e+00   0.620  0.53556
## Married_Dep1                      1.223e+00  1.413e+00   0.865  0.38688
## Married_Dep2                     -2.710e-01  1.727e+00  -0.157  0.87535
## Married_Dep3                             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 762.89  on 613  degrees of freedom
## Residual deviance: 542.42  on 588  degrees of freedom
## AIC: 594.42
##
## Number of Fisher Scoring iterations: 12

pred = predict(logistic, newdata=train,type="response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

pred <- as.integer(ifelse(pred>"0.499",1,0))
str(pred)

##  int [1:614] 1 1 1 1 1 1 1 0 1 1 ...

library(caret)

## Warning: package 'caret' was built under R version 3.2.5
```

```
## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.2.5

confusionMatrix(data=pred, train$Loan_Status)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  93  12
##          1  99 410
##
##                Accuracy : 0.8192
##                  95% CI : (0.7865, 0.8489)
##     No Information Rate : 0.6873
##     P-Value [Acc > NIR] : 9.197e-14
##
##                   Kappa : 0.5202
##  Mcnemar's Test P-Value : 3.275e-16
##
##             Sensitivity : 0.4844
##             Specificity : 0.9716
##          Pos Pred Value : 0.8857
##          Neg Pred Value : 0.8055
##              Prevalence : 0.3127
##          Detection Rate : 0.1515
##    Detection Prevalence : 0.1710
##       Balanced Accuracy : 0.7280
##
##        'Positive' Class : 0
##

######USing relevant selected parameters for prediction#######
logistic1 <- glm(Loan_Status ~
Credit_History+Property_AreaSemiurban+DebtRatio_TotalIncome+
                 Loan_Amount_Term1120.240.mnths+Loan_Amount_Term1240.360.mnths
, data=train[,!colnames(train) %in% c("Loan_ID")],
                 family='binomial')
summary(logistic1)

##
## Call:
## glm(formula = Loan_Status ~ Credit_History + Property_AreaSemiurban +
##     DebtRatio_TotalIncome + Loan_Amount_Term1120.240.mnths +
##     Loan_Amount_Term1240.360.mnths, family = "binomial", data = train[,
##     !colnames(train) %in% c("Loan_ID")])
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.2505  -0.3848   0.5397   0.7364   2.3693
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -2.9470     0.6584  -4.476 7.61e-06 ***
## Credit_History                    3.7912     0.3935   9.635  < 2e-16 ***
## Property_AreaSemiurban            0.8270     0.2345   3.527  0.00042 ***
## DebtRatio_TotalIncome           -24.9974    11.5929  -2.156  0.03106 *
## Loan_Amount_Term1120.240.mnths    0.9774     0.5931   1.648  0.09934 .
## Loan_Amount_Term1240.360.mnths    0.8742     0.4547   1.922  0.05456 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 762.89  on 613  degrees of freedom
## Residual deviance: 565.83  on 608  degrees of freedom
## AIC: 577.83
##
## Number of Fisher Scoring iterations: 5
```

```
pred1 = predict(logistic1, newdata=train,type="response")
pred1 <- as.integer(ifelse(pred>"0.499",1,0))
str(pred1)
```

```
##  int [1:614] 1 1 1 1 1 1 1 0 1 1 ...
```

```
confusionMatrix(data=pred1, train$Loan_Status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  93  12
##          1  99 410
##
##                Accuracy : 0.8192
##                  95% CI : (0.7865, 0.8489)
##     No Information Rate : 0.6873
##     P-Value [Acc > NIR] : 9.197e-14
##
##                   Kappa : 0.5202
##  Mcnemar's Test P-Value : 3.275e-16
##
##             Sensitivity : 0.4844
##             Specificity : 0.9716
##          Pos Pred Value : 0.8857
##          Neg Pred Value : 0.8055
##              Prevalence : 0.3127
##          Detection Rate : 0.1515
```

```
##    Detection Prevalence : 0.1710
##       Balanced Accuracy : 0.7280
##
##        'Positive' Class : 0
##

####The accuracy doesnt improve######
##Accuracy at 0.8192##
##LB Accuracy is 0.76##

#Predict Output
predicted= predict(logistic,test,type = "response")

## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading

head(predicted)

##         1         2         3         4         5         6
## 0.8577128 0.7662559 0.8066765 0.8921870 0.7043595 0.7848411

predicted <- as.integer(ifelse(predicted>"0.499",1,0))
head(predicted)

## [1] 1 1 1 1 1 1

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.2.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

test1 <- test %>% select (Loan_ID)
comb <- data.frame(test1,predicted)
write.csv(comb,"12345.csv",row.names = FALSE)
```