LoanPaybackPrediction

May

May 29, 2018

library(car)

## Loading required package: carData

library(tidyverse)

## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.0.0 v purrr 0.2.5  
## v tibble 1.4.2 v dplyr 0.7.6  
## v tidyr 0.8.1 v stringr 1.3.1  
## v readr 1.1.1 v forcats 0.3.0

## -- Conflicts ------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::recode() masks car::recode()  
## x purrr::some() masks car::some()

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(GGally)

##   
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':  
##   
## nasa

library(ROCR)

## Warning: package 'ROCR' was built under R version 3.5.3

## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.5.3

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(ggplot2)  
library(RcmdrMisc)

## Loading required package: sandwich

library(caTools)  
loans <- read.csv("C:/Stuff/JobApplication/PD/loan\_data.csv")  
summary(loans)

## credit.policy purpose int.rate   
## Min. :0.000 all\_other :2331 Min. :0.0600   
## 1st Qu.:1.000 credit\_card :1262 1st Qu.:0.1039   
## Median :1.000 debt\_consolidation:3957 Median :0.1221   
## Mean :0.805 educational : 343 Mean :0.1226   
## 3rd Qu.:1.000 home\_improvement : 629 3rd Qu.:0.1407   
## Max. :1.000 major\_purchase : 437 Max. :0.2164   
## small\_business : 619   
## installment log.annual.inc dti fico   
## Min. : 15.67 Min. : 7.548 Min. : 0.000 Min. :612.0   
## 1st Qu.:163.77 1st Qu.:10.558 1st Qu.: 7.213 1st Qu.:682.0   
## Median :268.95 Median :10.929 Median :12.665 Median :707.0   
## Mean :319.09 Mean :10.932 Mean :12.607 Mean :710.8   
## 3rd Qu.:432.76 3rd Qu.:11.291 3rd Qu.:17.950 3rd Qu.:737.0   
## Max. :940.14 Max. :14.528 Max. :29.960 Max. :827.0   
##   
## days.with.cr.line months.with.cr.line long\_term revol.bal   
## 3,750 : 63 Min. : 5.97 - : 36 Min. : 0   
## 4,080 : 58 1st Qu.: 94.00 1.00 :9542 1st Qu.: 3187   
## 3,180 : 57 Median :138.00 Median : 8596   
## 3,630 : 56 Mean :152.03 Mean : 16914   
## 3,660 : 56 3rd Qu.:191.00 3rd Qu.: 18250   
## 4,110 : 56 Max. :588.00 Max. :1207359   
## (Other):9232   
## revol.util inq.last.6mths delinq.2yrs pubrec   
## Min. : 0.0 Min. : 0.000 Min. : 0.0000 Min. :0.00000   
## 1st Qu.: 22.6 1st Qu.: 0.000 1st Qu.: 0.0000 1st Qu.:0.00000   
## Median : 46.3 Median : 1.000 Median : 0.0000 Median :0.00000   
## Mean : 46.8 Mean : 1.577 Mean : 0.1637 Mean :0.06212   
## 3rd Qu.: 70.9 3rd Qu.: 2.000 3rd Qu.: 0.0000 3rd Qu.:0.00000   
## Max. :119.0 Max. :33.000 Max. :13.0000 Max. :5.00000   
##   
## notfullypaid   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.1601   
## 3rd Qu.:0.0000   
## Max. :1.0000   
##

# Investigate data types of the columns

str(loans)

## 'data.frame': 9578 obs. of 16 variables:  
## $ credit.policy : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ purpose : Factor w/ 7 levels "all\_other","credit\_card",..: 3 2 3 3 2 2 3 1 5 3 ...  
## $ int.rate : num 0.119 0.107 0.136 0.101 0.143 ...  
## $ installment : num 829 228 367 162 103 ...  
## $ log.annual.inc : num 11.4 11.1 10.4 11.4 11.3 ...  
## $ dti : num 19.5 14.3 11.6 8.1 15 ...  
## $ fico : int 737 707 682 712 667 727 667 722 682 707 ...  
## $ days.with.cr.line : Factor w/ 1888 levels " 1,019 "," 1,020 ",..: 1333 570 1102 554 944 1444 676 1213 911 564 ...  
## $ months.with.cr.line: num 188 92 157 90 136 ...  
## $ long\_term : Factor w/ 2 levels " - "," 1.00 ": 2 2 2 2 2 2 2 2 2 2 ...  
## $ revol.bal : int 28854 33623 3511 33667 4740 50807 3839 24220 69909 5630 ...  
## $ revol.util : num 52.1 76.7 25.6 73.2 39.5 51 76.8 68.6 51.1 23 ...  
## $ inq.last.6mths : int 0 0 1 1 0 0 0 0 1 1 ...  
## $ delinq.2yrs : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ pubrec : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ notfullypaid : int 0 0 0 0 0 0 1 1 0 0 ...

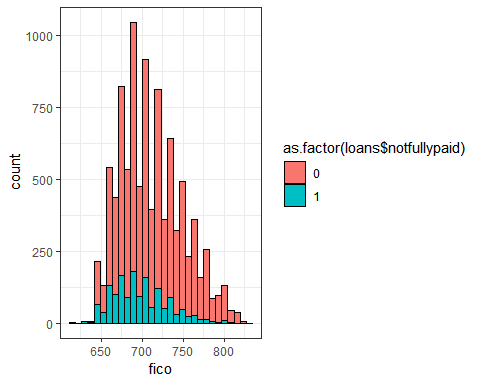
# Transfer the data type to correct format

loans$days.with.cr.line <- as.numeric(loans$days.with.cr.line)  
loans$delinq.2yrs <- as.numeric(loans$delinq.2yrs)  
loans$inq.last.6mths <- as.numeric(loans$inq.last.6mths)  
loans$long\_term <- as.numeric(loans$long\_term)

# Understand data: use ggplot to visualize

ggplot(data = loans,aes(fico)) + geom\_histogram(aes(fill=as.factor(loans$notfullypaid)), color = 'black') + theme\_bw()

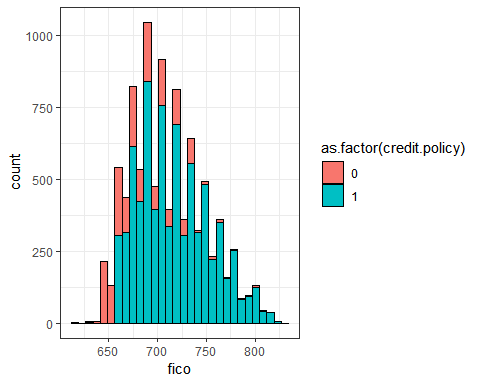
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The higher fico is, the less likely the loan is default

ggplot(data = loans, aes(fico)) + geom\_histogram(aes(fill = as.factor(credit.policy)), color = 'black') + theme\_bw()

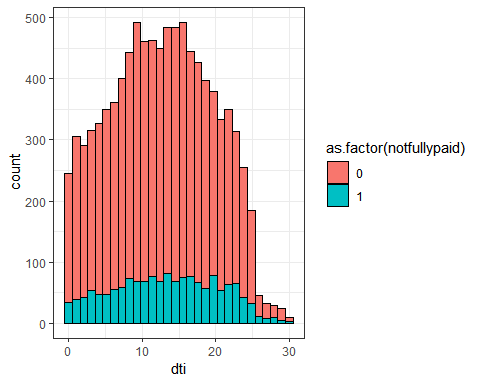
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The higher the fico is, the more likely the loan is accepted by the credit policy.

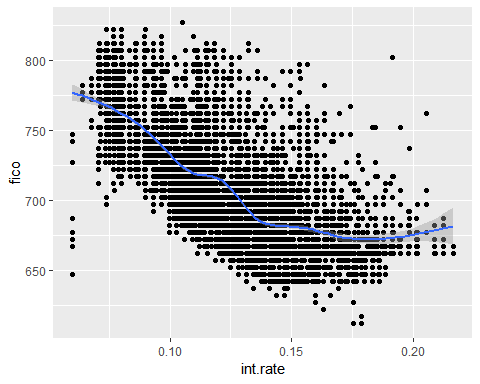
ggplot(data = loans,aes(dti)) + geom\_histogram(aes(fill = as.factor(notfullypaid)),color = "black") +theme\_bw()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



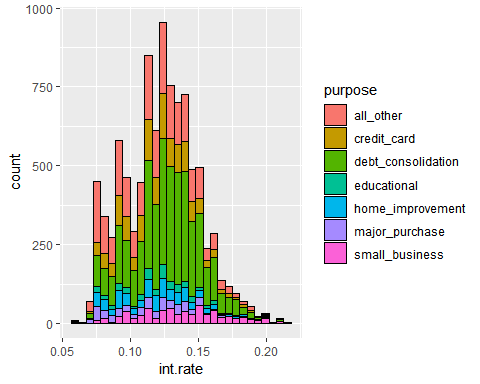
ggplot(data = loans, aes(int.rate,fico)) + geom\_point() + stat\_smooth()

## `geom\_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'



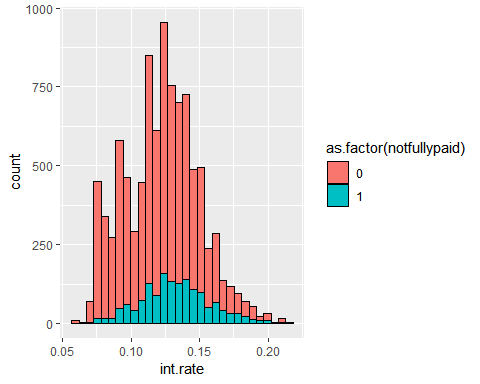
ggplot(data = loans, aes(int.rate)) + geom\_histogram(aes(fill = purpose), color = "black")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(data = loans, aes(int.rate)) + geom\_histogram(aes(fill = as.factor(notfullypaid)), color = "black")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

 The higher fico is the lower the interest is Interest rate seems irrelevant to loan purpose and loan payment ability

set.seed(123)  
sample <- sample.int(n = nrow(loans), size = floor(.9\*nrow(loans)), replace = F)  
loanstrain <- loans[sample,]  
loanstest <- loans[-sample,]

# Model 1: How purpose relates to loan payment

mod1 <- glm(notfullypaid ~ purpose, data = loanstrain)  
summary(mod1)

##   
## Call:  
## glm(formula = notfullypaid ~ purpose, data = loanstrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.2773 -0.1658 -0.1536 -0.1164 0.8836   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.165796 0.007979 20.778 < 2e-16 \*\*\*  
## purposecredit\_card -0.049394 0.013486 -3.663 0.000251 \*\*\*  
## purposedebt\_consolidation -0.012208 0.010062 -1.213 0.225027   
## purposeeducational 0.043207 0.022240 1.943 0.052078 .   
## purposehome\_improvement 0.004494 0.017506 0.257 0.797407   
## purposemajor\_purchase -0.045591 0.020161 -2.261 0.023760 \*   
## purposesmall\_business 0.111485 0.017419 6.400 1.63e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1340283)  
##   
## Null deviance: 1165.9 on 8619 degrees of freedom  
## Residual deviance: 1154.4 on 8613 degrees of freedom  
## AIC: 7147.8  
##   
## Number of Fisher Scoring iterations: 2

mod1rsquared <- (mod1$null.deviance- mod1$deviance)/mod1$null.deviance  
mod1rsquared

## [1] 0.009840628

# Model 2: How installment amount, annual income, and interest rate relate to loan payment ability

mod2 <- glm(notfullypaid ~ installment + log.annual.inc + int.rate, data = loanstrain)  
summary(mod2)

##   
## Call:  
## glm(formula = notfullypaid ~ installment + log.annual.inc + int.rate,   
## data = loanstrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.37615 -0.18926 -0.14518 -0.07403 0.98035   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.402e-01 7.815e-02 3.073 0.00212 \*\*   
## installment 6.450e-05 2.198e-05 2.935 0.00335 \*\*   
## log.annual.inc -3.246e-02 7.104e-03 -4.569 4.96e-06 \*\*\*  
## int.rate 2.082e+00 1.520e-01 13.701 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.131569)  
##   
## Null deviance: 1165.9 on 8619 degrees of freedom  
## Residual deviance: 1133.6 on 8616 degrees of freedom  
## AIC: 6985.2  
##   
## Number of Fisher Scoring iterations: 2

mod2rsquared <- (mod2$null.deviance- mod2$deviance)/mod2$null.deviance  
mod2rsquared

## [1] 0.02767048

str(loanstrain)

## 'data.frame': 8620 obs. of 16 variables:  
## $ credit.policy : int 1 1 1 0 0 1 1 0 1 1 ...  
## $ purpose : Factor w/ 7 levels "all\_other","credit\_card",..: 1 1 2 3 3 3 4 2 3 6 ...  
## $ int.rate : num 0.144 0.139 0.122 0.161 0.15 ...  
## $ installment : num 51.6 40.9 116.6 490.1 503 ...  
## $ log.annual.inc : num 11.2 11 10.6 10.6 10.1 ...  
## $ dti : num 11.4 3.08 8.28 18.94 5.75 ...  
## $ fico : int 677 707 687 652 697 682 792 692 727 692 ...  
## $ days.with.cr.line : num 770 28 1246 962 584 ...  
## $ months.with.cr.line: num 117 39 175 138 94 89 169 115 280 107 ...  
## $ long\_term : num 2 2 2 2 2 2 2 2 2 2 ...  
## $ revol.bal : int 457 2823 3900 17059 3695 15680 0 39486 20017 22002 ...  
## $ revol.util : num 18 65.7 43.7 82.4 46.2 89.6 0 86.4 32 81.2 ...  
## $ inq.last.6mths : num 0 3 0 1 6 1 0 1 0 3 ...  
## $ delinq.2yrs : num 0 0 1 0 0 0 0 0 0 0 ...  
## $ pubrec : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ notfullypaid : int 0 1 0 0 1 0 0 0 1 0 ...

# Model 3: How all variables impact the loan payment ability

mod3 <- glm(notfullypaid ~ ., data = loanstrain)  
summary(mod3)

##   
## Call:  
## glm(formula = notfullypaid ~ ., data = loanstrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.70976 -0.18534 -0.12478 -0.05185 1.02006   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.470e+00 2.047e-01 7.180 7.53e-13 \*\*\*  
## credit.policy -5.286e-02 1.263e-02 -4.186 2.87e-05 \*\*\*  
## purposecredit\_card -6.325e-02 1.349e-02 -4.689 2.79e-06 \*\*\*  
## purposedebt\_consolidation -4.309e-02 1.036e-02 -4.158 3.24e-05 \*\*\*  
## purposeeducational 2.080e-02 2.178e-02 0.955 0.3397   
## purposehome\_improvement 9.274e-03 1.728e-02 0.537 0.5915   
## purposemajor\_purchase -2.829e-02 1.968e-02 -1.437 0.1506   
## purposesmall\_business 8.111e-02 1.786e-02 4.541 5.66e-06 \*\*\*  
## int.rate 3.318e-01 2.478e-01 1.339 0.1807   
## installment 1.463e-04 2.419e-05 6.048 1.53e-09 \*\*\*  
## log.annual.inc -5.129e-02 7.865e-03 -6.522 7.34e-11 \*\*\*  
## dti -1.451e-04 6.198e-04 -0.234 0.8149   
## fico -9.313e-04 1.892e-04 -4.923 8.69e-07 \*\*\*  
## days.with.cr.line 2.123e-05 9.239e-06 2.298 0.0216 \*   
## months.with.cr.line 4.085e-06 5.833e-05 0.070 0.9442   
## long\_term -8.748e-02 6.357e-02 -1.376 0.1688   
## revol.bal 5.668e-07 1.392e-07 4.071 4.72e-05 \*\*\*  
## revol.util 2.760e-04 1.772e-04 1.558 0.1193   
## inq.last.6mths 1.470e-02 2.090e-03 7.032 2.18e-12 \*\*\*  
## delinq.2yrs -1.234e-02 7.569e-03 -1.630 0.1031   
## pubrec 3.380e-02 1.496e-02 2.260 0.0238 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1271239)  
##   
## Null deviance: 1165.9 on 8619 degrees of freedom  
## Residual deviance: 1093.1 on 8599 degrees of freedom  
## AIC: 6705.9  
##   
## Number of Fisher Scoring iterations: 2

mod3rsquared <- (mod3$null.deviance- mod3$deviance)/mod3$null.deviance  
mod3rsquared

## [1] 0.06237456

# Reduced Model 3: Use stepwise to remove insignificant variables from Model 3

redmod3 <- stepwise(mod3)

##   
## Direction: backward/forward  
## Criterion: BIC   
##   
## Start: AIC=6854.22  
## notfullypaid ~ credit.policy + purpose + int.rate + installment +   
## log.annual.inc + dti + fico + days.with.cr.line + months.with.cr.line +   
## long\_term + revol.bal + revol.util + inq.last.6mths + delinq.2yrs +   
## pubrec  
##   
## Df Deviance AIC  
## - months.with.cr.line 1 1093.1 6845.2  
## - dti 1 1093.2 6845.2  
## - int.rate 1 1093.4 6847.0  
## - long\_term 1 1093.4 6847.1  
## - revol.util 1 1093.5 6847.6  
## - delinq.2yrs 1 1093.5 6847.8  
## - pubrec 1 1093.8 6850.3  
## - days.with.cr.line 1 1093.8 6850.5  
## <none> 1093.1 6854.2  
## - revol.bal 1 1095.2 6861.8  
## - credit.policy 1 1095.4 6862.7  
## - fico 1 1096.2 6869.4  
## - purpose 6 1103.3 6879.4  
## - installment 1 1097.8 6881.8  
## - log.annual.inc 1 1098.5 6887.7  
## - inq.last.6mths 1 1099.4 6894.6  
##   
## Step: AIC=6845.17  
## notfullypaid ~ credit.policy + purpose + int.rate + installment +   
## log.annual.inc + dti + fico + days.with.cr.line + long\_term +   
## revol.bal + revol.util + inq.last.6mths + delinq.2yrs + pubrec  
##   
## Df Deviance AIC  
## - dti 1 1093.2 6836.2  
## - int.rate 1 1093.4 6837.9  
## - long\_term 1 1093.4 6838.0  
## - revol.util 1 1093.5 6838.6  
## - delinq.2yrs 1 1093.5 6838.8  
## - pubrec 1 1093.8 6841.3  
## - days.with.cr.line 1 1094.0 6842.8  
## <none> 1093.1 6845.2  
## - revol.bal 1 1095.3 6853.0  
## - credit.policy 1 1095.4 6853.7  
## + months.with.cr.line 1 1093.1 6854.2  
## - fico 1 1096.3 6861.2  
## - purpose 6 1103.3 6870.4  
## - installment 1 1097.8 6872.7  
## - log.annual.inc 1 1098.6 6879.4  
## - inq.last.6mths 1 1099.4 6885.7  
##   
## Step: AIC=6836.16  
## notfullypaid ~ credit.policy + purpose + int.rate + installment +   
## log.annual.inc + fico + days.with.cr.line + long\_term + revol.bal +   
## revol.util + inq.last.6mths + delinq.2yrs + pubrec  
##   
## Df Deviance AIC  
## - int.rate 1 1093.4 6828.9  
## - long\_term 1 1093.4 6829.0  
## - revol.util 1 1093.5 6829.5  
## - delinq.2yrs 1 1093.5 6829.8  
## - pubrec 1 1093.8 6832.3  
## - days.with.cr.line 1 1094.0 6833.7  
## <none> 1093.2 6836.2  
## - revol.bal 1 1095.3 6844.2  
## - credit.policy 1 1095.4 6844.7  
## + dti 1 1093.1 6845.2  
## + months.with.cr.line 1 1093.2 6845.2  
## - fico 1 1096.3 6852.2  
## - installment 1 1097.8 6863.7  
## - purpose 6 1103.6 6863.7  
## - log.annual.inc 1 1098.7 6870.9  
## - inq.last.6mths 1 1099.4 6876.6  
##   
## Step: AIC=6828.88  
## notfullypaid ~ credit.policy + purpose + installment + log.annual.inc +   
## fico + days.with.cr.line + long\_term + revol.bal + revol.util +   
## inq.last.6mths + delinq.2yrs + pubrec  
##   
## Df Deviance AIC  
## - long\_term 1 1093.6 6821.5  
## - delinq.2yrs 1 1093.7 6822.4  
## - revol.util 1 1093.8 6822.8  
## - pubrec 1 1094.0 6825.0  
## - days.with.cr.line 1 1094.2 6826.1  
## <none> 1093.4 6828.9  
## + int.rate 1 1093.2 6836.2  
## - revol.bal 1 1095.5 6836.5  
## + dti 1 1093.4 6837.9  
## + months.with.cr.line 1 1093.4 6837.9  
## - credit.policy 1 1095.7 6838.4  
## - log.annual.inc 1 1099.0 6863.9  
## - purpose 6 1105.2 6867.3  
## - inq.last.6mths 1 1099.8 6870.6  
## - installment 1 1100.3 6874.1  
## - fico 1 1100.8 6877.8  
##   
## Step: AIC=6821.49  
## notfullypaid ~ credit.policy + purpose + installment + log.annual.inc +   
## fico + days.with.cr.line + revol.bal + revol.util + inq.last.6mths +   
## delinq.2yrs + pubrec  
##   
## Df Deviance AIC  
## - delinq.2yrs 1 1093.9 6815.2  
## - revol.util 1 1093.9 6815.2  
## - pubrec 1 1094.2 6817.5  
## - days.with.cr.line 1 1094.4 6818.7  
## <none> 1093.6 6821.5  
## + long\_term 1 1093.4 6828.9  
## - revol.bal 1 1095.7 6828.9  
## + int.rate 1 1093.4 6829.0  
## + dti 1 1093.6 6830.5  
## + months.with.cr.line 1 1093.6 6830.6  
## - credit.policy 1 1096.1 6832.5  
## - log.annual.inc 1 1099.3 6857.7  
## - purpose 6 1105.5 6860.5  
## - inq.last.6mths 1 1100.0 6862.6  
## - installment 1 1100.5 6867.1  
## - fico 1 1101.1 6871.2  
##   
## Step: AIC=6815.17  
## notfullypaid ~ credit.policy + purpose + installment + log.annual.inc +   
## fico + days.with.cr.line + revol.bal + revol.util + inq.last.6mths +   
## pubrec  
##   
## Df Deviance AIC  
## - revol.util 1 1094.5 6810.3  
## - pubrec 1 1094.6 6811.5  
## - days.with.cr.line 1 1094.6 6811.5  
## <none> 1093.9 6815.2  
## + delinq.2yrs 1 1093.6 6821.5  
## + long\_term 1 1093.7 6822.4  
## + int.rate 1 1093.8 6822.8  
## - revol.bal 1 1096.1 6823.0  
## + dti 1 1093.9 6824.2  
## + months.with.cr.line 1 1093.9 6824.2  
## - credit.policy 1 1096.5 6826.2  
## - log.annual.inc 1 1099.9 6852.8  
## - purpose 6 1105.8 6854.1  
## - inq.last.6mths 1 1100.5 6857.3  
## - installment 1 1100.9 6860.5  
## - fico 1 1101.1 6862.6  
##   
## Step: AIC=6810.31  
## notfullypaid ~ credit.policy + purpose + installment + log.annual.inc +   
## fico + days.with.cr.line + revol.bal + inq.last.6mths + pubrec  
##   
## Df Deviance AIC  
## - pubrec 1 1095.1 6806.5  
## - days.with.cr.line 1 1095.2 6807.4  
## <none> 1094.5 6810.3  
## + revol.util 1 1093.9 6815.2  
## + delinq.2yrs 1 1093.9 6815.2  
## + int.rate 1 1094.2 6817.3  
## + long\_term 1 1094.3 6817.7  
## + dti 1 1094.5 6819.3  
## + months.with.cr.line 1 1094.5 6819.4  
## - credit.policy 1 1096.8 6819.8  
## - revol.bal 1 1097.2 6822.5  
## - purpose 6 1105.9 6845.3  
## - log.annual.inc 1 1100.4 6847.6  
## - inq.last.6mths 1 1100.8 6850.7  
## - installment 1 1101.6 6857.3  
## - fico 1 1108.0 6907.2  
##   
## Step: AIC=6806.5  
## notfullypaid ~ credit.policy + purpose + installment + log.annual.inc +   
## fico + days.with.cr.line + revol.bal + inq.last.6mths  
##   
## Df Deviance AIC  
## - days.with.cr.line 1 1096.1 6804.9  
## <none> 1095.1 6806.5  
## + pubrec 1 1094.5 6810.3  
## + delinq.2yrs 1 1094.6 6811.1  
## + revol.util 1 1094.6 6811.5  
## + int.rate 1 1094.9 6813.5  
## + long\_term 1 1094.9 6814.0  
## + dti 1 1095.1 6815.6  
## + months.with.cr.line 1 1095.1 6815.6  
## - credit.policy 1 1097.5 6815.6  
## - revol.bal 1 1097.7 6817.7  
## - purpose 6 1106.5 6841.1  
## - log.annual.inc 1 1100.9 6842.8  
## - inq.last.6mths 1 1101.6 6848.5  
## - installment 1 1102.1 6852.0  
## - fico 1 1109.9 6912.7  
##   
## Step: AIC=6804.91  
## notfullypaid ~ credit.policy + purpose + installment + log.annual.inc +   
## fico + revol.bal + inq.last.6mths  
##   
## Df Deviance AIC  
## <none> 1096.1 6804.9  
## + days.with.cr.line 1 1095.1 6806.5  
## + pubrec 1 1095.2 6807.4  
## + revol.util 1 1095.5 6809.2  
## + delinq.2yrs 1 1095.7 6810.6  
## + int.rate 1 1095.9 6812.3  
## + months.with.cr.line 1 1095.9 6812.3  
## + long\_term 1 1095.9 6812.5  
## + dti 1 1096.1 6813.9  
## - credit.policy 1 1098.7 6816.2  
## - revol.bal 1 1098.9 6817.7  
## - log.annual.inc 1 1101.2 6835.6  
## - purpose 6 1107.3 6838.2  
## - inq.last.6mths 1 1102.4 6845.1  
## - installment 1 1103.1 6850.6  
## - fico 1 1110.0 6904.7

summary(redmod3)

##   
## Call:  
## glm(formula = notfullypaid ~ credit.policy + purpose + installment +   
## log.annual.inc + fico + revol.bal + inq.last.6mths, data = loanstrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.71230 -0.18423 -0.12728 -0.05298 1.04603   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.485e+00 1.065e-01 13.942 < 2e-16 \*\*\*  
## credit.policy -5.587e-02 1.239e-02 -4.508 6.63e-06 \*\*\*  
## purposecredit\_card -5.974e-02 1.331e-02 -4.490 7.23e-06 \*\*\*  
## purposedebt\_consolidation -3.950e-02 1.020e-02 -3.874 0.000108 \*\*\*  
## purposeeducational 2.057e-02 2.179e-02 0.944 0.345104   
## purposehome\_improvement 1.185e-02 1.725e-02 0.687 0.492131   
## purposemajor\_purchase -2.905e-02 1.968e-02 -1.476 0.139942   
## purposesmall\_business 8.722e-02 1.742e-02 5.007 5.63e-07 \*\*\*  
## installment 1.608e-04 2.171e-05 7.406 1.42e-13 \*\*\*  
## log.annual.inc -4.735e-02 7.509e-03 -6.305 3.01e-10 \*\*\*  
## fico -1.165e-03 1.114e-04 -10.460 < 2e-16 \*\*\*  
## revol.bal 6.201e-07 1.328e-07 4.669 3.08e-06 \*\*\*  
## inq.last.6mths 1.456e-02 2.073e-03 7.025 2.31e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1273478)  
##   
## Null deviance: 1165.9 on 8619 degrees of freedom  
## Residual deviance: 1096.1 on 8607 degrees of freedom  
## AIC: 6713.1  
##   
## Number of Fisher Scoring iterations: 2

redmod3rsquared <- (redmod3$null.deviance- redmod3$deviance)/redmod3$null.deviance  
redmod3rsquared

## [1] 0.05984925

# Model 4: Manually reduced model: From Reduced model 3, select statistically significant variables

mod4 <- glm(notfullypaid ~ credit.policy + installment + log.annual.inc + fico + revol.bal + inq.last.6mths, data = loanstrain)  
summary(mod4)

##   
## Call:  
## glm(formula = notfullypaid ~ credit.policy + installment + log.annual.inc +   
## fico + revol.bal + inq.last.6mths, data = loanstrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.71384 -0.18106 -0.13126 -0.06394 1.01163   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.359e+00 1.033e-01 13.165 < 2e-16 \*\*\*  
## credit.policy -5.892e-02 1.243e-02 -4.739 2.19e-06 \*\*\*  
## installment 1.596e-04 2.100e-05 7.600 3.28e-14 \*\*\*  
## log.annual.inc -4.682e-02 7.433e-03 -6.299 3.14e-10 \*\*\*  
## fico -1.021e-03 1.095e-04 -9.326 < 2e-16 \*\*\*  
## revol.bal 6.224e-07 1.330e-07 4.680 2.92e-06 \*\*\*  
## inq.last.6mths 1.589e-02 2.077e-03 7.649 2.24e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1285604)  
##   
## Null deviance: 1165.9 on 8619 degrees of freedom  
## Residual deviance: 1107.3 on 8613 degrees of freedom  
## AIC: 6788.8  
##   
## Number of Fisher Scoring iterations: 2

mod4rsquared <- (mod4$null.deviance- mod4$deviance)/mod4$null.deviance  
mod4rsquared

## [1] 0.05023555

# Final Model: Reduced Model 3 is chosen to predict the probability of not fully payment

loanstrain$Mod3PredTrain <- predict(redmod3, newdata = loanstrain, type = "response")  
loanstrain$Mod3ClassTrain <- ifelse(loanstrain$Mod3PredTrain > 0.5, 1,0)  
confusionMatrixTrain <- table(loanstrain$Mod3ClassTrain, loanstrain$notfullypaid)  
confusionMatrixTrain

##   
## 0 1  
## 0 7214 1380  
## 1 16 10

AccuracyTrain <- (7214+10)/(7214+16+1389+10)  
AccuracyTrain

## [1] 0.837177

loanstest$Mod3PredTest <- predict(redmod3, newdata = loanstest, type = "response")  
loanstest$Mod3ClassTest <- ifelse(loanstest$Mod3PredTest > 0.5, 1,0)  
confusionMatrixTest <- table(loanstest$Mod3ClassTest, loanstest$notfullypaid)  
confusionMatrixTest

##   
## 0 1  
## 0 814 142  
## 1 1 1

AccuracyTest <- (814+1)/(814+1+142+11)  
AccuracyTest

## [1] 0.8419421

# Prediction on original data

loans$Mod3Pred <- predict(redmod3, newdata = loans, type = "response")  
loans$Mod3Class <- ifelse(loans$Mod3Pred > 0.5, 1,0)  
confusionMatrix <- table(loans$Mod3Class, loans$notfullypaid)  
confusionMatrix

##   
## 0 1  
## 0 8028 1522  
## 1 17 11

Accuracy <- (8028+17)/(8028+17+1522+11)  
Accuracy

## [1] 0.8399457

# New loans data with prob and class fields

str(loans)

## 'data.frame': 9578 obs. of 18 variables:  
## $ credit.policy : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ purpose : Factor w/ 7 levels "all\_other","credit\_card",..: 3 2 3 3 2 2 3 1 5 3 ...  
## $ int.rate : num 0.119 0.107 0.136 0.101 0.143 ...  
## $ installment : num 829 228 367 162 103 ...  
## $ log.annual.inc : num 11.4 11.1 10.4 11.4 11.3 ...  
## $ dti : num 19.5 14.3 11.6 8.1 15 ...  
## $ fico : int 737 707 682 712 667 727 667 722 682 707 ...  
## $ days.with.cr.line : num 1333 570 1102 554 944 ...  
## $ months.with.cr.line: num 188 92 157 90 136 ...  
## $ long\_term : num 2 2 2 2 2 2 2 2 2 2 ...  
## $ revol.bal : int 28854 33623 3511 33667 4740 50807 3839 24220 69909 5630 ...  
## $ revol.util : num 52.1 76.7 25.6 73.2 39.5 51 76.8 68.6 51.1 23 ...  
## $ inq.last.6mths : num 0 0 1 1 0 0 0 0 1 1 ...  
## $ delinq.2yrs : num 0 0 0 0 1 0 0 0 0 0 ...  
## $ pubrec : int 0 0 0 0 0 0 1 0 0 0 ...  
## $ notfullypaid : int 0 0 0 0 0 0 1 1 0 0 ...  
## $ Mod3Pred : num 0.1451 0.0788 0.1799 0.0845 0.0771 ...  
## $ Mod3Class : num 0 0 0 0 0 0 0 0 0 0 ...