Final Project

Amol Gote

08/09/2020

# Dataset Info

1. For this project will be using the lending loan club dataset, located at <https://www.kaggle.com/wordsforthewise/lending-club>
2. Since this datasize is huge and due to comouting constraints, will using the data for 2018 only. Only approved loans dataset will be in scope for this project.
3. Effective dataset size if of 495,242 observations/loans

# About lending club

The Lending Club is a peer-to-peer lending service (it lends money to customers by matching lenders to borrowers), based in the United States. This company enables borrowers to create loan listings on its website by supplying details about themselves and the loans that they would like to request.

# Dataset Attributes

1. Dataset contains various attributes based on which the credit lending decision takes place, this includes Debt to income ratio, annual income, home ownership type (Rented,Mortgage, Owned), employment details. Dataset also contains geographical information like state and zip code. It also has issued loan details like funded Loan Amount, issued month and year, loan Term (36 or 60 months) and interest rate. Post loan has been issued, loan performance needs to be tracked, and this dataset also contains those details loan Status (Current, Paid Off, Charged Off, Delinquent), loan Grades and delinquency details.

# Research Objectives

1. Qualitative - Build a model which will indicate if the loan would ever be delinquent, the delinquency could be of 30, 60, 90 or 120 days.
2. Quantitative - Build a model which will provide the interest rate when the person applies for the loan.
3. Unsupervised - Identify correlation between various credit decisioning variables like fico, dti, annual income etc using PCA and K-Mean.

Please note that some of the research objective have changed than what were quoted in the project out line document in Week-5,  
a. Instead of loan amount, now predicting interest rate. b. For qualitative predicting only delinquency and not charge offs.

# Feature engineering

Load the whole dataset

lendingClubLoanData <- fread("data/Lending\_club\_dataset.csv", header = TRUE)

Create a origination year fild on the dataset, so that all loans which have originated in 2018 can be extracted.

nrow(lendingClubLoanData)

## [1] 2260701

lendingClubLoanData$orig\_year<-substr(lendingClubLoanData$issue\_d,5,8)

lendingClubLoanData\_2018 <- lendingClubLoanData %>%  
 filter(orig\_year == 2018)

lc\_ds\_size <- nrow(lendingClubLoanData\_2018)  
lc\_ds\_size

## [1] 495242

Dataset which will used for this analysis would comprise of 495242 observations.

Create file for all loans which originated in 2018

nrow(lendingClubLoanData\_2018)

## [1] 495242

write.csv(lendingClubLoanData\_2018, file = "data/lending\_club\_loan\_data\_2018\_final.csv", row.names=FALSE)

Load the final data set for analysis

final\_dataset <- fread("data/lending\_club\_loan\_data\_2018\_final.csv", header = TRUE)  
drops <- c("id","member\_id", "url", "desc")  
#final\_dataset[ , !(names(final\_dataset) %in% drops)]

List down all varibales in the dataset

names(final\_dataset)

## [1] "id"   
## [2] "member\_id"   
## [3] "loan\_amnt"   
## [4] "funded\_amnt"   
## [5] "funded\_amnt\_inv"   
## [6] "term"   
## [7] "int\_rate"   
## [8] "installment"   
## [9] "grade"   
## [10] "sub\_grade"   
## [11] "emp\_title"   
## [12] "emp\_length"   
## [13] "home\_ownership"   
## [14] "annual\_inc"   
## [15] "verification\_status"   
## [16] "issue\_d"   
## [17] "loan\_status"   
## [18] "pymnt\_plan"   
## [19] "url"   
## [20] "desc"   
## [21] "purpose"   
## [22] "title"   
## [23] "zip\_code"   
## [24] "addr\_state"   
## [25] "dti"   
## [26] "delinq\_2yrs"   
## [27] "earliest\_cr\_line"   
## [28] "fico\_range\_low"   
## [29] "fico\_range\_high"   
## [30] "inq\_last\_6mths"   
## [31] "mths\_since\_last\_delinq"   
## [32] "mths\_since\_last\_record"   
## [33] "open\_acc"   
## [34] "pub\_rec"   
## [35] "revol\_bal"   
## [36] "revol\_util"   
## [37] "total\_acc"   
## [38] "initial\_list\_status"   
## [39] "out\_prncp"   
## [40] "out\_prncp\_inv"   
## [41] "total\_pymnt"   
## [42] "total\_pymnt\_inv"   
## [43] "total\_rec\_prncp"   
## [44] "total\_rec\_int"   
## [45] "total\_rec\_late\_fee"   
## [46] "recoveries"   
## [47] "collection\_recovery\_fee"   
## [48] "last\_pymnt\_d"   
## [49] "last\_pymnt\_amnt"   
## [50] "next\_pymnt\_d"   
## [51] "last\_credit\_pull\_d"   
## [52] "last\_fico\_range\_high"   
## [53] "last\_fico\_range\_low"   
## [54] "collections\_12\_mths\_ex\_med"   
## [55] "mths\_since\_last\_major\_derog"   
## [56] "policy\_code"   
## [57] "application\_type"   
## [58] "annual\_inc\_joint"   
## [59] "dti\_joint"   
## [60] "verification\_status\_joint"   
## [61] "acc\_now\_delinq"   
## [62] "tot\_coll\_amt"   
## [63] "tot\_cur\_bal"   
## [64] "open\_acc\_6m"   
## [65] "open\_act\_il"   
## [66] "open\_il\_12m"   
## [67] "open\_il\_24m"   
## [68] "mths\_since\_rcnt\_il"   
## [69] "total\_bal\_il"   
## [70] "il\_util"   
## [71] "open\_rv\_12m"   
## [72] "open\_rv\_24m"   
## [73] "max\_bal\_bc"   
## [74] "all\_util"   
## [75] "total\_rev\_hi\_lim"   
## [76] "inq\_fi"   
## [77] "total\_cu\_tl"   
## [78] "inq\_last\_12m"   
## [79] "acc\_open\_past\_24mths"   
## [80] "avg\_cur\_bal"   
## [81] "bc\_open\_to\_buy"   
## [82] "bc\_util"   
## [83] "chargeoff\_within\_12\_mths"   
## [84] "delinq\_amnt"   
## [85] "mo\_sin\_old\_il\_acct"   
## [86] "mo\_sin\_old\_rev\_tl\_op"   
## [87] "mo\_sin\_rcnt\_rev\_tl\_op"   
## [88] "mo\_sin\_rcnt\_tl"   
## [89] "mort\_acc"   
## [90] "mths\_since\_recent\_bc"   
## [91] "mths\_since\_recent\_bc\_dlq"   
## [92] "mths\_since\_recent\_inq"   
## [93] "mths\_since\_recent\_revol\_delinq"   
## [94] "num\_accts\_ever\_120\_pd"   
## [95] "num\_actv\_bc\_tl"   
## [96] "num\_actv\_rev\_tl"   
## [97] "num\_bc\_sats"   
## [98] "num\_bc\_tl"   
## [99] "num\_il\_tl"   
## [100] "num\_op\_rev\_tl"   
## [101] "num\_rev\_accts"   
## [102] "num\_rev\_tl\_bal\_gt\_0"   
## [103] "num\_sats"   
## [104] "num\_tl\_120dpd\_2m"   
## [105] "num\_tl\_30dpd"   
## [106] "num\_tl\_90g\_dpd\_24m"   
## [107] "num\_tl\_op\_past\_12m"   
## [108] "pct\_tl\_nvr\_dlq"   
## [109] "percent\_bc\_gt\_75"   
## [110] "pub\_rec\_bankruptcies"   
## [111] "tax\_liens"   
## [112] "tot\_hi\_cred\_lim"   
## [113] "total\_bal\_ex\_mort"   
## [114] "total\_bc\_limit"   
## [115] "total\_il\_high\_credit\_limit"   
## [116] "revol\_bal\_joint"   
## [117] "sec\_app\_fico\_range\_low"   
## [118] "sec\_app\_fico\_range\_high"   
## [119] "sec\_app\_earliest\_cr\_line"   
## [120] "sec\_app\_inq\_last\_6mths"   
## [121] "sec\_app\_mort\_acc"   
## [122] "sec\_app\_open\_acc"   
## [123] "sec\_app\_revol\_util"   
## [124] "sec\_app\_open\_act\_il"   
## [125] "sec\_app\_num\_rev\_accts"   
## [126] "sec\_app\_chargeoff\_within\_12\_mths"   
## [127] "sec\_app\_collections\_12\_mths\_ex\_med"   
## [128] "sec\_app\_mths\_since\_last\_major\_derog"   
## [129] "hardship\_flag"   
## [130] "hardship\_type"   
## [131] "hardship\_reason"   
## [132] "hardship\_status"   
## [133] "deferral\_term"   
## [134] "hardship\_amount"   
## [135] "hardship\_start\_date"   
## [136] "hardship\_end\_date"   
## [137] "payment\_plan\_start\_date"   
## [138] "hardship\_length"   
## [139] "hardship\_dpd"   
## [140] "hardship\_loan\_status"   
## [141] "orig\_projected\_additional\_accrued\_interest"  
## [142] "hardship\_payoff\_balance\_amount"   
## [143] "hardship\_last\_payment\_amount"   
## [144] "disbursement\_method"   
## [145] "debt\_settlement\_flag"   
## [146] "debt\_settlement\_flag\_date"   
## [147] "settlement\_status"   
## [148] "settlement\_date"   
## [149] "settlement\_amount"   
## [150] "settlement\_percentage"   
## [151] "settlement\_term"   
## [152] "orig\_year"

Summary of all the varibales in the dataset

summary(final\_dataset)

## id member\_id loan\_amnt funded\_amnt   
## Min. : 70102325 Mode:logical Min. : 1000 Min. : 1000   
## 1st Qu.:131267288 NA's:495242 1st Qu.: 8000 1st Qu.: 8000   
## Median :136195162 Median :14000 Median :14000   
## Mean :136087208 Mean :16025 Mean :16025   
## 3rd Qu.:141042201 3rd Qu.:22000 3rd Qu.:22000   
## Max. :145647287 Max. :40000 Max. :40000   
##   
## funded\_amnt\_inv term int\_rate installment   
## Min. : 725 Length:495242 Min. : 5.31 Min. : 29.76   
## 1st Qu.: 8000 Class :character 1st Qu.: 8.46 1st Qu.: 254.56   
## Median :14000 Mode :character Median :11.80 Median : 386.82   
## Mean :16022 Mean :12.73 Mean : 466.61   
## 3rd Qu.:22000 3rd Qu.:16.01 3rd Qu.: 629.04   
## Max. :40000 Max. :30.99 Max. :1670.15   
##   
## grade sub\_grade emp\_title emp\_length   
## Length:495242 Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## home\_ownership annual\_inc verification\_status issue\_d   
## Length:495242 Min. : 0 Length:495242 Length:495242   
## Class :character 1st Qu.: 46000 Class :character Class :character   
## Mode :character Median : 66000 Mode :character Mode :character   
## Mean : 80094   
## 3rd Qu.: 96000   
## Max. :9930475   
##   
## loan\_status pymnt\_plan url desc   
## Length:495242 Length:495242 Length:495242 Mode:logical   
## Class :character Class :character Class :character NA's:495242   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## purpose title zip\_code addr\_state   
## Length:495242 Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## dti delinq\_2yrs earliest\_cr\_line fico\_range\_low   
## Min. : 0.00 Min. : 0.0000 Length:495242 Min. :660.0   
## 1st Qu.: 11.43 1st Qu.: 0.0000 Class :character 1st Qu.:680.0   
## Median : 17.71 Median : 0.0000 Mode :character Median :700.0   
## Mean : 19.67 Mean : 0.2293 Mean :706.4   
## 3rd Qu.: 25.03 3rd Qu.: 0.0000 3rd Qu.:725.0   
## Max. :999.00 Max. :58.0000 Max. :845.0   
## NA's :1132   
## fico\_range\_high inq\_last\_6mths mths\_since\_last\_delinq mths\_since\_last\_record  
## Min. :664.0 Min. :0.0000 Min. : 0.00 Min. : 1.0   
## 1st Qu.:684.0 1st Qu.:0.0000 1st Qu.: 19.00 1st Qu.: 69.0   
## Median :704.0 Median :0.0000 Median : 34.00 Median : 87.0   
## Mean :710.4 Mean :0.4422 Mean : 36.89 Mean : 83.3   
## 3rd Qu.:729.0 3rd Qu.:1.0000 3rd Qu.: 53.00 3rd Qu.:102.0   
## Max. :850.0 Max. :5.0000 Max. :226.00 Max. :127.0   
## NA's :276652 NA's :432258   
## open\_acc pub\_rec revol\_bal revol\_util   
## Min. : 0.00 Min. : 0.0000 Min. : 0 Min. : 0.00   
## 1st Qu.: 7.00 1st Qu.: 0.0000 1st Qu.: 5304 1st Qu.: 24.20   
## Median : 10.00 Median : 0.0000 Median : 10832 Median : 42.10   
## Mean : 11.49 Mean : 0.1345 Mean : 16271 Mean : 43.88   
## 3rd Qu.: 14.00 3rd Qu.: 0.0000 3rd Qu.: 19867 3rd Qu.: 62.30   
## Max. :101.00 Max. :52.0000 Max. :2358150 Max. :191.00   
## NA's :592   
## total\_acc initial\_list\_status out\_prncp out\_prncp\_inv   
## Min. : 2.00 Length:495242 Min. : 0 Min. : 0   
## 1st Qu.: 14.00 Class :character 1st Qu.: 4623 1st Qu.: 4622   
## Median : 21.00 Mode :character Median : 9742 Median : 9739   
## Mean : 22.62 Mean :12007 Mean :12005   
## 3rd Qu.: 29.00 3rd Qu.:17801 3rd Qu.:17801   
## Max. :160.00 Max. :40000 Max. :40000   
##   
## total\_pymnt total\_pymnt\_inv total\_rec\_prncp total\_rec\_int   
## Min. : 0 Min. : 0 Min. : 0 Min. : 0.0   
## 1st Qu.: 1758 1st Qu.: 1758 1st Qu.: 1115 1st Qu.: 401.6   
## Median : 3221 Median : 3221 Median : 2075 Median : 814.7   
## Mean : 4925 Mean : 4924 Mean : 3735 Mean : 1179.6   
## 3rd Qu.: 5909 3rd Qu.: 5908 3rd Qu.: 4000 3rd Qu.: 1575.2   
## Max. :51653 Max. :51653 Max. :40000 Max. :12795.0   
##   
## total\_rec\_late\_fee recoveries collection\_recovery\_fee  
## Min. : 0.0000 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.0000 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.0000 Median : 0.00 Median : 0.000   
## Mean : 0.5128 Mean : 9.94 Mean : 1.748   
## 3rd Qu.: 0.0000 3rd Qu.: 0.00 3rd Qu.: 0.000   
## Max. :458.5700 Max. :33122.07 Max. :5961.973   
##   
## last\_pymnt\_d last\_pymnt\_amnt next\_pymnt\_d last\_credit\_pull\_d  
## Length:495242 Min. : 0.0 Length:495242 Length:495242   
## Class :character 1st Qu.: 270.7 Class :character Class :character   
## Mode :character Median : 436.9 Mode :character Mode :character   
## Mean : 1599.8   
## 3rd Qu.: 747.9   
## Max. :41353.7   
##   
## last\_fico\_range\_high last\_fico\_range\_low collections\_12\_mths\_ex\_med  
## Min. : 0.0 Min. : 0.0 Min. :0.00000   
## 1st Qu.:679.0 1st Qu.:675.0 1st Qu.:0.00000   
## Median :709.0 Median :705.0 Median :0.00000   
## Mean :709.4 Mean :703.7 Mean :0.01768   
## 3rd Qu.:744.0 3rd Qu.:740.0 3rd Qu.:0.00000   
## Max. :850.0 Max. :845.0 Max. :9.00000   
##   
## mths\_since\_last\_major\_derog policy\_code application\_type annual\_inc\_joint   
## Min. : 0.0 Min. :1 Length:495242 Min. : 5694   
## 1st Qu.: 29.0 1st Qu.:1 Class :character 1st Qu.: 85000   
## Median : 46.0 Median :1 Mode :character Median : 114000   
## Mean : 46.3 Mean :1 Mean : 128342   
## 3rd Qu.: 64.0 3rd Qu.:1 3rd Qu.: 152000   
## Max. :226.0 Max. :1 Max. :7874821   
## NA's :380409 NA's :426257   
## dti\_joint verification\_status\_joint acc\_now\_delinq   
## Min. : 0.0 Length:495242 Min. :0.00e+00   
## 1st Qu.:13.4 Class :character 1st Qu.:0.00e+00   
## Median :18.9 Mode :character Median :0.00e+00   
## Mean :19.4 Mean :5.25e-05   
## 3rd Qu.:25.0 3rd Qu.:0.00e+00   
## Max. :40.0 Max. :1.00e+00   
## NA's :426257   
## tot\_coll\_amt tot\_cur\_bal open\_acc\_6m open\_act\_il   
## Min. : 0 Min. : 0 Min. : 0.0000 Min. : 0.000   
## 1st Qu.: 0 1st Qu.: 26885 1st Qu.: 0.0000 1st Qu.: 1.000   
## Median : 0 Median : 74550 Median : 1.0000 Median : 2.000   
## Mean : 214 Mean : 143960 Mean : 0.8981 Mean : 2.701   
## 3rd Qu.: 0 3rd Qu.: 218044 3rd Qu.: 1.0000 3rd Qu.: 3.000   
## Max. :6214661 Max. :9971659 Max. :15.0000 Max. :56.000   
##   
## open\_il\_12m open\_il\_24m mths\_since\_rcnt\_il total\_bal\_il   
## Min. :0.0000 Min. : 0.000 Min. : 0.0 Min. : 0   
## 1st Qu.:0.0000 1st Qu.: 0.000 1st Qu.: 7.0 1st Qu.: 7851   
## Median :0.0000 Median : 1.000 Median : 13.0 Median : 22452   
## Mean :0.6729 Mean : 1.527 Mean : 21.1 Mean : 35324   
## 3rd Qu.:1.0000 3rd Qu.: 2.000 3rd Qu.: 25.0 3rd Qu.: 45845   
## Max. :8.0000 Max. :20.000 Max. :507.0 Max. :1837038   
## NA's :18410   
## il\_util open\_rv\_12m open\_rv\_24m max\_bal\_bc   
## Min. : 0 Min. : 0.000 Min. : 0.000 Min. : 0   
## 1st Qu.: 53 1st Qu.: 0.000 1st Qu.: 1.000 1st Qu.: 2183   
## Median : 71 Median : 1.000 Median : 2.000 Median : 4441   
## Mean : 68 Mean : 1.228 Mean : 2.624 Mean : 5863   
## 3rd Qu.: 85 3rd Qu.: 2.000 3rd Qu.: 4.000 3rd Qu.: 7772   
## Max. :1000 Max. :26.000 Max. :50.000 Max. :1170668   
## NA's :80824   
## all\_util total\_rev\_hi\_lim inq\_fi total\_cu\_tl   
## Min. : 0.00 Min. : 0 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 40.00 1st Qu.: 16200 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 55.00 Median : 28700 Median : 1.000 Median : 0.000   
## Mean : 54.09 Mean : 38345 Mean : 1.087 Mean : 1.485   
## 3rd Qu.: 69.00 3rd Qu.: 48700 3rd Qu.: 2.000 3rd Qu.: 2.000   
## Max. :239.00 Max. :2087500 Max. :38.000 Max. :68.000   
## NA's :129   
## inq\_last\_12m acc\_open\_past\_24mths avg\_cur\_bal bc\_open\_to\_buy   
## Min. : 0.000 Min. : 0.000 Min. : 0 Min. : 0   
## 1st Qu.: 0.000 1st Qu.: 2.000 1st Qu.: 2904 1st Qu.: 2867   
## Median : 1.000 Median : 4.000 Median : 7033 Median : 8301   
## Mean : 1.937 Mean : 4.427 Mean : 13709 Mean : 15057   
## 3rd Qu.: 3.000 3rd Qu.: 6.000 3rd Qu.: 19031 3rd Qu.: 19694   
## Max. :67.000 Max. :54.000 Max. :623229 Max. :605996   
## NA's :40 NA's :6588   
## bc\_util chargeoff\_within\_12\_mths delinq\_amnt   
## Min. : 0.00 Min. :0.000000 Min. : 0.00   
## 1st Qu.: 25.95 1st Qu.:0.000000 1st Qu.: 0.00   
## Median : 48.80 Median :0.000000 Median : 0.00   
## Mean : 49.87 Mean :0.006823 Mean : 1.79   
## 3rd Qu.: 74.40 3rd Qu.:0.000000 3rd Qu.: 0.00   
## Max. :201.60 Max. :9.000000 Max. :65000.00   
## NA's :6803   
## mo\_sin\_old\_il\_acct mo\_sin\_old\_rev\_tl\_op mo\_sin\_rcnt\_rev\_tl\_op  
## Min. : 0 Min. : 1.0 Min. : 0.00   
## 1st Qu.: 85 1st Qu.:102.0 1st Qu.: 4.00   
## Median :129 Median :156.0 Median : 9.00   
## Mean :123 Mean :174.6 Mean : 15.13   
## 3rd Qu.:155 3rd Qu.:226.0 3rd Qu.: 19.00   
## Max. :848 Max. :826.0 Max. :502.00   
## NA's :18410   
## mo\_sin\_rcnt\_tl mort\_acc mths\_since\_recent\_bc  
## Min. : 0.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 3.000 1st Qu.: 0.000 1st Qu.: 6.00   
## Median : 6.000 Median : 1.000 Median : 15.00   
## Mean : 8.719 Mean : 1.336 Mean : 25.55   
## 3rd Qu.: 11.000 3rd Qu.: 2.000 3rd Qu.: 30.00   
## Max. :382.000 Max. :87.000 Max. :661.00   
## NA's :6198   
## mths\_since\_recent\_bc\_dlq mths\_since\_recent\_inq mths\_since\_recent\_revol\_delinq  
## Min. : 0.0 Min. : 0.00 Min. : 0.0   
## 1st Qu.: 23.0 1st Qu.: 2.00 1st Qu.: 20.0   
## Median : 38.0 Median : 6.00 Median : 35.0   
## Mean : 40.5 Mean : 7.49 Mean : 37.8   
## 3rd Qu.: 57.0 3rd Qu.:11.00 3rd Qu.: 53.0   
## Max. :194.0 Max. :25.00 Max. :190.0   
## NA's :397132 NA's :61305 NA's :352552   
## num\_accts\_ever\_120\_pd num\_actv\_bc\_tl num\_actv\_rev\_tl num\_bc\_sats   
## Min. : 0.0000 Min. : 0.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 0.0000 1st Qu.: 2.000 1st Qu.: 3.000 1st Qu.: 3.00   
## Median : 0.0000 Median : 3.000 Median : 5.000 Median : 4.00   
## Mean : 0.4679 Mean : 3.614 Mean : 5.362 Mean : 4.84   
## 3rd Qu.: 0.0000 3rd Qu.: 5.000 3rd Qu.: 7.000 3rd Qu.: 6.00   
## Max. :58.0000 Max. :50.000 Max. :72.000 Max. :69.00   
##   
## num\_bc\_tl num\_il\_tl num\_op\_rev\_tl num\_rev\_accts   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 2.00   
## 1st Qu.: 4.000 1st Qu.: 3.000 1st Qu.: 5.000 1st Qu.: 7.00   
## Median : 6.000 Median : 6.000 Median : 7.000 Median : 11.00   
## Mean : 7.093 Mean : 8.169 Mean : 8.163 Mean : 12.92   
## 3rd Qu.: 9.000 3rd Qu.: 11.000 3rd Qu.:10.000 3rd Qu.: 17.00   
## Max. :86.000 Max. :130.000 Max. :91.000 Max. :151.00   
##   
## num\_rev\_tl\_bal\_gt\_0 num\_sats num\_tl\_120dpd\_2m num\_tl\_30dpd   
## Min. : 0.000 Min. : 0.00 Min. :0 Min. :0.00e+00   
## 1st Qu.: 3.000 1st Qu.: 7.00 1st Qu.:0 1st Qu.:0.00e+00   
## Median : 5.000 Median : 10.00 Median :0 Median :0.00e+00   
## Mean : 5.325 Mean : 11.47 Mean :0 Mean :4.64e-05   
## 3rd Qu.: 7.000 3rd Qu.: 14.00 3rd Qu.:0 3rd Qu.:0.00e+00   
## Max. :65.000 Max. :101.00 Max. :0 Max. :1.00e+00   
## NA's :12404   
## num\_tl\_90g\_dpd\_24m num\_tl\_op\_past\_12m pct\_tl\_nvr\_dlq percent\_bc\_gt\_75  
## Min. : 0.00000 Min. : 0.00 Min. : 0.00 Min. : 0.00   
## 1st Qu.: 0.00000 1st Qu.: 1.00 1st Qu.: 92.30 1st Qu.: 0.00   
## Median : 0.00000 Median : 2.00 Median :100.00 Median : 25.00   
## Mean : 0.05993 Mean : 2.03 Mean : 94.58 Mean : 32.89   
## 3rd Qu.: 0.00000 3rd Qu.: 3.00 3rd Qu.:100.00 3rd Qu.: 57.10   
## Max. :58.00000 Max. :26.00 Max. :100.00 Max. :100.00   
## NA's :2 NA's :6596   
## pub\_rec\_bankruptcies tax\_liens tot\_hi\_cred\_lim total\_bal\_ex\_mort  
## Min. :0.0000 Min. : 0.00000 Min. : 0 Min. : 0   
## 1st Qu.:0.0000 1st Qu.: 0.00000 1st Qu.: 52059 1st Qu.: 19530   
## Median :0.0000 Median : 0.00000 Median : 116292 Median : 37556   
## Mean :0.1235 Mean : 0.01092 Mean : 184942 Mean : 51922   
## 3rd Qu.:0.0000 3rd Qu.: 0.00000 3rd Qu.: 269292 3rd Qu.: 66119   
## Max. :7.0000 Max. :52.00000 Max. :9999999 Max. :2622906   
##   
## total\_bc\_limit total\_il\_high\_credit\_limit revol\_bal\_joint   
## Min. : 0 Min. : 0 Min. : 0   
## 1st Qu.: 9800 1st Qu.: 14926 1st Qu.: 15477   
## Median : 19200 Median : 33754 Median : 27350   
## Mean : 26680 Mean : 45688 Mean : 34757   
## 3rd Qu.: 35200 3rd Qu.: 61914 3rd Qu.: 45215   
## Max. :1569000 Max. :2118996 Max. :1110019   
## NA's :426257   
## sec\_app\_fico\_range\_low sec\_app\_fico\_range\_high sec\_app\_earliest\_cr\_line  
## Min. :540.0 Min. :544.0 Length:495242   
## 1st Qu.:645.0 1st Qu.:649.0 Class :character   
## Median :670.0 Median :674.0 Mode :character   
## Mean :671.8 Mean :675.8   
## 3rd Qu.:700.0 3rd Qu.:704.0   
## Max. :845.0 Max. :850.0   
## NA's :426257 NA's :426257   
## sec\_app\_inq\_last\_6mths sec\_app\_mort\_acc sec\_app\_open\_acc sec\_app\_revol\_util  
## Min. :0.0 Min. : 0.0 Min. : 0.0 Min. : 0.0   
## 1st Qu.:0.0 1st Qu.: 0.0 1st Qu.: 7.0 1st Qu.: 38.3   
## Median :0.0 Median : 1.0 Median :10.0 Median : 59.1   
## Mean :0.6 Mean : 1.5 Mean :11.5 Mean : 57.1   
## 3rd Qu.:1.0 3rd Qu.: 2.0 3rd Qu.:15.0 3rd Qu.: 77.8   
## Max. :6.0 Max. :27.0 Max. :67.0 Max. :434.3   
## NA's :426257 NA's :426257 NA's :426257 NA's :427454   
## sec\_app\_open\_act\_il sec\_app\_num\_rev\_accts sec\_app\_chargeoff\_within\_12\_mths  
## Min. : 0 Min. : 0.0 Min. : 0   
## 1st Qu.: 1 1st Qu.: 7.0 1st Qu.: 0   
## Median : 2 Median : 11.0 Median : 0   
## Mean : 3 Mean : 12.5 Mean : 0   
## 3rd Qu.: 4 3rd Qu.: 17.0 3rd Qu.: 0   
## Max. :43 Max. :106.0 Max. :21   
## NA's :426257 NA's :426257 NA's :426257   
## sec\_app\_collections\_12\_mths\_ex\_med sec\_app\_mths\_since\_last\_major\_derog  
## Min. : 0.0 Min. : 0.0   
## 1st Qu.: 0.0 1st Qu.: 17.0   
## Median : 0.0 Median : 36.0   
## Mean : 0.1 Mean : 37.3   
## 3rd Qu.: 0.0 3rd Qu.: 57.0   
## Max. :23.0 Max. :185.0   
## NA's :426257 NA's :472865   
## hardship\_flag hardship\_type hardship\_reason hardship\_status   
## Length:495242 Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## deferral\_term hardship\_amount hardship\_start\_date hardship\_end\_date   
## Min. :3 Min. : 6.0 Length:495242 Length:495242   
## 1st Qu.:3 1st Qu.: 87.4 Class :character Class :character   
## Median :3 Median :160.1 Mode :character Mode :character   
## Mean :3 Mean :196.1   
## 3rd Qu.:3 3rd Qu.:271.9   
## Max. :3 Max. :845.2   
## NA's :494874 NA's :494874   
## payment\_plan\_start\_date hardship\_length hardship\_dpd hardship\_loan\_status  
## Length:495242 Min. :3 Min. : 0.0 Length:495242   
## Class :character 1st Qu.:3 1st Qu.: 5.0 Class :character   
## Mode :character Median :3 Median :13.0 Mode :character   
## Mean :3 Mean :12.8   
## 3rd Qu.:3 3rd Qu.:20.0   
## Max. :3 Max. :29.0   
## NA's :494874 NA's :494874   
## orig\_projected\_additional\_accrued\_interest hardship\_payoff\_balance\_amount  
## Min. : 31.5 Min. : 424.1   
## 1st Qu.: 258.7 1st Qu.: 7966.5   
## Median : 479.8 Median :13323.6   
## Mean : 595.4 Mean :15633.3   
## 3rd Qu.: 819.9 3rd Qu.:22590.2   
## Max. :2535.7 Max. :40149.3   
## NA's :494921 NA's :494874   
## hardship\_last\_payment\_amount disbursement\_method debt\_settlement\_flag  
## Min. : 0.1 Length:495242 Length:495242   
## 1st Qu.: 50.9 Class :character Class :character   
## Median : 153.2 Mode :character Mode :character   
## Mean : 215.7   
## 3rd Qu.: 327.4   
## Max. :1159.6   
## NA's :494874   
## debt\_settlement\_flag\_date settlement\_status settlement\_date   
## Length:495242 Length:495242 Length:495242   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## settlement\_amount settlement\_percentage settlement\_term orig\_year   
## Min. : 413.9 Min. :29.9 Min. : 1 Min. :2018   
## 1st Qu.: 3367.1 1st Qu.:45.0 1st Qu.:17 1st Qu.:2018   
## Median : 5626.1 Median :55.0 Median :18 Median :2018   
## Mean : 7145.7 Mean :54.6 Mean :18 Mean :2018   
## 3rd Qu.: 9761.5 3rd Qu.:65.0 3rd Qu.:24 3rd Qu.:2018   
## Max. :28503.0 Max. :80.0 Max. :24 Max. :2018   
## NA's :494762 NA's :494762 NA's :494762

## Variable Selection

1. Dataset contains lot of variables, these variables are attributes associated with existing loans and what we are trying to predict is when loan gets originated, at that point in time identify in its life time whether loan would get delinquent at any point in time. So selecting attributes which are available only when loan gets onboarded. Most of these attributes come from customers credit bureau report.

ds\_lc <- final\_dataset[,c("loan\_amnt",  
 "funded\_amnt",  
 "funded\_amnt\_inv",  
 "term",  
 "int\_rate",  
 "installment",  
 "grade",  
 "sub\_grade",  
 "home\_ownership",  
 "annual\_inc",  
 "loan\_status",  
 "dti",  
 "delinq\_2yrs",  
 "fico\_range\_low",  
 "fico\_range\_high",  
 "inq\_last\_6mths",  
 "mths\_since\_last\_delinq",  
 "mths\_since\_last\_record",  
 "open\_acc",  
 "pub\_rec",  
 "revol\_bal",  
 "revol\_util",  
 "total\_acc",  
 "out\_prncp",  
 "out\_prncp\_inv",  
 "total\_pymnt",  
 "total\_pymnt\_inv",  
 "total\_rec\_prncp",  
 "total\_rec\_int",  
 "total\_rec\_late\_fee",  
 "recoveries",  
 "collection\_recovery\_fee",  
 "last\_fico\_range\_high",  
 "last\_fico\_range\_low",  
 "mths\_since\_last\_major\_derog",  
 "acc\_now\_delinq",  
 "avg\_cur\_bal",  
 "inq\_last\_12m", "num\_tl\_30dpd", "num\_tl\_90g\_dpd\_24m", "tot\_hi\_cred\_lim", "num\_rev\_accts"  
 )]

## Clean up the datset

1. convert all categorical variables to factors
2. Remove the NA values
3. Convert variables to numeric or integer
4. Create a delinquent account flag, loan account becomes deliquent when loan status turns to either of the following (is\_acct\_delinquent = 1) a. Late (16-30 days) b. Late (31-120 days) c. Charged Off d. Default e. In Grace Period

Loan status with following is one which is not delinquent (is\_acct\_delinquent = 0)  
a. Current  
b. Fully Paid

nrow(ds\_lc)

## [1] 495242

ds\_lc <- na.omit(ds\_lc, cols = c("dti", "int\_rate"))  
nrow(ds\_lc)

## [1] 494110

levels(factor(ds\_lc$loan\_status))

## [1] "Charged Off" "Current" "Default"   
## [4] "Fully Paid" "In Grace Period" "Late (16-30 days)"   
## [7] "Late (31-120 days)"

levels(factor(ds\_lc$grade))

## [1] "A" "B" "C" "D" "E" "F" "G"

levels(factor(ds\_lc$sub\_grade))

## [1] "A1" "A2" "A3" "A4" "A5" "B1" "B2" "B3" "B4" "B5" "C1" "C2" "C3" "C4" "C5"  
## [16] "D1" "D2" "D3" "D4" "D5" "E1" "E2" "E3" "E4" "E5" "F1" "F2" "F3" "F4" "F5"  
## [31] "G1" "G2" "G3" "G4" "G5"

levels(factor(ds\_lc$home\_ownership))

## [1] "ANY" "MORTGAGE" "OWN" "RENT"

ds\_lc$term <- as.integer(as.factor(ds\_lc$term))   
ds\_lc$grade <- as.integer(as.factor(ds\_lc$grade))  
ds\_lc$sub\_grade <- as.integer(as.factor(ds\_lc$sub\_grade))  
ds\_lc$home\_ownership <- as.integer(as.factor(ds\_lc$home\_ownership))  
ds\_lc$int\_rate <- as.numeric(ds\_lc$int\_rate)  
ds\_lc$loan\_status <- as.integer(as.factor(ds\_lc$loan\_status))  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 1] <- 1  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 2] <- 0  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 3] <- 1  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 4] <- 0  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 5] <- 1  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 6] <- 0  
ds\_lc$is\_acct\_delinquent[ds\_lc$loan\_status == 7] <- 1  
ds\_lc$is\_acct\_delinquent <- factor(ds\_lc$is\_acct\_delinquent)  
levels(factor(ds\_lc$is\_acct\_delinquent))

## [1] "0" "1"

ds\_lc = subset(ds\_lc, select = -c(loan\_status))

Perform regular subset varibale selection test

regular\_subset\_full <- regsubsets(is\_acct\_delinquent∼.,ds\_lc, nvmax = 25)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 3 linear dependencies found

## Reordering variables and trying again:

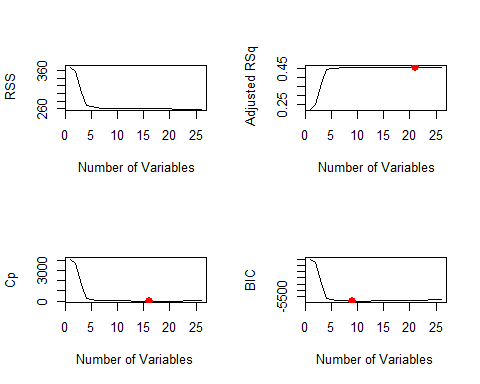
regular\_subset\_full\_summary <- summary(regular\_subset\_full)  
regular\_subset\_full\_summary

## Subset selection object  
## Call: regsubsets.formula(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 25)  
## 41 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## funded\_amnt\_inv FALSE FALSE  
## term FALSE FALSE  
## int\_rate FALSE FALSE  
## installment FALSE FALSE  
## grade FALSE FALSE  
## sub\_grade FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_low FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## mths\_since\_last\_delinq FALSE FALSE  
## mths\_since\_last\_record FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## revol\_util FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_pymnt\_inv FALSE FALSE  
## total\_rec\_prncp FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## last\_fico\_range\_low FALSE FALSE  
## mths\_since\_last\_major\_derog FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## avg\_cur\_bal FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## funded\_amnt FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## 1 subsets of each size up to 26  
## Selection Algorithm: exhaustive  
## loan\_amnt funded\_amnt funded\_amnt\_inv term int\_rate installment grade  
## 1 ( 1 ) " " " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " " " "   
## 3 ( 1 ) " " " " "\*" " " " " " " " "   
## 4 ( 1 ) " " " " "\*" " " " " " " " "   
## 5 ( 1 ) " " " " "\*" " " " " " " " "   
## 6 ( 1 ) " " " " "\*" " " " " " " " "   
## 7 ( 1 ) " " " " "\*" " " " " " " " "   
## 8 ( 1 ) " " " " "\*" " " " " " " " "   
## 9 ( 1 ) " " " " "\*" " " " " " " " "   
## 10 ( 1 ) "\*" " " " " " " "\*" " " " "   
## 11 ( 1 ) " " "\*" " " " " " " " " " "   
## 12 ( 1 ) "\*" " " " " " " "\*" " " " "   
## 13 ( 1 ) " " "\*" " " " " "\*" " " " "   
## 14 ( 1 ) " " "\*" " " " " "\*" " " " "   
## 15 ( 1 ) " " " " "\*" "\*" "\*" "\*" " "   
## 16 ( 1 ) " " " " "\*" "\*" "\*" "\*" " "   
## 17 ( 1 ) " " " " "\*" "\*" "\*" "\*" " "   
## 18 ( 1 ) " " " " "\*" "\*" "\*" "\*" " "   
## 19 ( 1 ) " " " " "\*" "\*" "\*" "\*" " "   
## 20 ( 1 ) "\*" " " " " " " "\*" " " " "   
## 21 ( 1 ) "\*" " " " " " " "\*" " " " "   
## 22 ( 1 ) "\*" " " " " " " "\*" " " " "   
## 23 ( 1 ) "\*" " " " " "\*" "\*" " " " "   
## 24 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 25 ( 1 ) " " " " "\*" "\*" "\*" "\*" " "   
## 26 ( 1 ) " " " " "\*" "\*" "\*" "\*" "\*"   
## sub\_grade home\_ownership annual\_inc dti delinq\_2yrs fico\_range\_low  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " " " "   
## 7 ( 1 ) " " " " " " " " " " " "   
## 8 ( 1 ) " " " " " " " " " " " "   
## 9 ( 1 ) "\*" " " " " " " " " " "   
## 10 ( 1 ) " " " " " " " " " " "\*"   
## 11 ( 1 ) "\*" " " " " " " " " " "   
## 12 ( 1 ) " " " " " " " " " " "\*"   
## 13 ( 1 ) "\*" " " " " " " " " "\*"   
## 14 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 15 ( 1 ) "\*" " " " " " " " " "\*"   
## 16 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 17 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 18 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 19 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 20 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 21 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 22 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 23 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 24 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 25 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 26 ( 1 ) "\*" " " "\*" " " " " "\*"   
## fico\_range\_high inq\_last\_6mths mths\_since\_last\_delinq  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) "\*" " " " "   
## 7 ( 1 ) "\*" " " " "   
## 8 ( 1 ) "\*" " " " "   
## 9 ( 1 ) "\*" " " " "   
## 10 ( 1 ) "\*" " " " "   
## 11 ( 1 ) "\*" " " " "   
## 12 ( 1 ) "\*" " " " "   
## 13 ( 1 ) "\*" " " " "   
## 14 ( 1 ) "\*" " " " "   
## 15 ( 1 ) "\*" " " " "   
## 16 ( 1 ) "\*" " " " "   
## 17 ( 1 ) "\*" " " " "   
## 18 ( 1 ) "\*" " " " "   
## 19 ( 1 ) "\*" " " " "   
## 20 ( 1 ) "\*" " " " "   
## 21 ( 1 ) "\*" " " " "   
## 22 ( 1 ) "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" " "   
## 24 ( 1 ) " " " " " "   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## mths\_since\_last\_record open\_acc pub\_rec revol\_bal revol\_util  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " "   
## 7 ( 1 ) " " " " " " " " " "   
## 8 ( 1 ) "\*" " " " " " " " "   
## 9 ( 1 ) " " " " " " " " " "   
## 10 ( 1 ) " " " " " " " " " "   
## 11 ( 1 ) "\*" " " " " " " " "   
## 12 ( 1 ) "\*" " " " " " " " "   
## 13 ( 1 ) "\*" " " " " " " " "   
## 14 ( 1 ) "\*" " " " " " " " "   
## 15 ( 1 ) "\*" " " " " " " " "   
## 16 ( 1 ) "\*" " " " " " " " "   
## 17 ( 1 ) "\*" " " " " "\*" " "   
## 18 ( 1 ) "\*" " " " " " " "\*"   
## 19 ( 1 ) "\*" " " "\*" " " "\*"   
## 20 ( 1 ) "\*" " " " " " " "\*"   
## 21 ( 1 ) "\*" " " "\*" " " "\*"   
## 22 ( 1 ) "\*" " " "\*" " " "\*"   
## 23 ( 1 ) "\*" " " "\*" " " "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" " " "\*" "\*" "\*"   
## 26 ( 1 ) "\*" " " "\*" "\*" "\*"   
## total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_pymnt\_inv  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " "\*" " " " " " "   
## 4 ( 1 ) " " "\*" " " " " " "   
## 5 ( 1 ) " " "\*" " " " " " "   
## 6 ( 1 ) " " "\*" " " " " " "   
## 7 ( 1 ) " " "\*" " " " " " "   
## 8 ( 1 ) " " "\*" " " " " " "   
## 9 ( 1 ) " " "\*" " " " " "\*"   
## 10 ( 1 ) " " "\*" " " "\*" " "   
## 11 ( 1 ) " " "\*" " " "\*" " "   
## 12 ( 1 ) " " "\*" " " "\*" " "   
## 13 ( 1 ) " " "\*" " " "\*" " "   
## 14 ( 1 ) " " "\*" " " "\*" " "   
## 15 ( 1 ) " " "\*" " " " " "\*"   
## 16 ( 1 ) " " "\*" " " " " "\*"   
## 17 ( 1 ) " " "\*" " " " " "\*"   
## 18 ( 1 ) "\*" "\*" " " " " "\*"   
## 19 ( 1 ) "\*" "\*" " " " " "\*"   
## 20 ( 1 ) "\*" "\*" " " "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" " "   
## 22 ( 1 ) "\*" "\*" " " "\*" " "   
## 23 ( 1 ) "\*" "\*" " " "\*" " "   
## 24 ( 1 ) "\*" " " " " "\*" " "   
## 25 ( 1 ) "\*" "\*" " " " " "\*"   
## 26 ( 1 ) "\*" "\*" " " " " "\*"   
## total\_rec\_prncp total\_rec\_int total\_rec\_late\_fee recoveries  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " "\*"   
## 3 ( 1 ) "\*" " " " " " "   
## 4 ( 1 ) "\*" " " " " " "   
## 5 ( 1 ) "\*" " " "\*" " "   
## 6 ( 1 ) "\*" " " "\*" " "   
## 7 ( 1 ) "\*" " " "\*" " "   
## 8 ( 1 ) "\*" " " "\*" " "   
## 9 ( 1 ) "\*" " " "\*" " "   
## 10 ( 1 ) "\*" " " "\*" " "   
## 11 ( 1 ) "\*" " " "\*" " "   
## 12 ( 1 ) "\*" " " "\*" " "   
## 13 ( 1 ) "\*" " " "\*" " "   
## 14 ( 1 ) "\*" " " "\*" " "   
## 15 ( 1 ) "\*" " " "\*" " "   
## 16 ( 1 ) "\*" " " "\*" " "   
## 17 ( 1 ) "\*" " " "\*" " "   
## 18 ( 1 ) "\*" " " "\*" " "   
## 19 ( 1 ) "\*" " " "\*" " "   
## 20 ( 1 ) "\*" "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*"   
## 25 ( 1 ) " " "\*" "\*" "\*"   
## 26 ( 1 ) " " "\*" "\*" "\*"   
## collection\_recovery\_fee last\_fico\_range\_high last\_fico\_range\_low  
## 1 ( 1 ) " " "\*" " "   
## 2 ( 1 ) " " "\*" " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " "\*" " "   
## 5 ( 1 ) " " "\*" " "   
## 6 ( 1 ) " " "\*" " "   
## 7 ( 1 ) " " "\*" "\*"   
## 8 ( 1 ) " " "\*" "\*"   
## 9 ( 1 ) " " "\*" "\*"   
## 10 ( 1 ) " " "\*" "\*"   
## 11 ( 1 ) " " "\*" "\*"   
## 12 ( 1 ) " " "\*" "\*"   
## 13 ( 1 ) " " "\*" "\*"   
## 14 ( 1 ) " " "\*" "\*"   
## 15 ( 1 ) " " "\*" "\*"   
## 16 ( 1 ) " " "\*" "\*"   
## 17 ( 1 ) " " "\*" "\*"   
## 18 ( 1 ) " " "\*" "\*"   
## 19 ( 1 ) " " "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## mths\_since\_last\_major\_derog acc\_now\_delinq avg\_cur\_bal inq\_last\_12m  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) " " " " " " " "   
## 11 ( 1 ) "\*" " " " " " "   
## 12 ( 1 ) "\*" " " " " " "   
## 13 ( 1 ) "\*" " " " " " "   
## 14 ( 1 ) "\*" " " " " " "   
## 15 ( 1 ) "\*" " " " " " "   
## 16 ( 1 ) "\*" " " " " " "   
## 17 ( 1 ) "\*" " " " " " "   
## 18 ( 1 ) "\*" " " " " " "   
## 19 ( 1 ) "\*" " " " " " "   
## 20 ( 1 ) "\*" " " " " "\*"   
## 21 ( 1 ) "\*" " " " " "\*"   
## 22 ( 1 ) "\*" " " " " "\*"   
## 23 ( 1 ) "\*" " " " " "\*"   
## 24 ( 1 ) "\*" " " "\*" "\*"   
## 25 ( 1 ) "\*" " " " " "\*"   
## 26 ( 1 ) "\*" " " " " "\*"   
## num\_tl\_30dpd num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) " " " " " " " "   
## 11 ( 1 ) " " " " " " " "   
## 12 ( 1 ) " " " " " " " "   
## 13 ( 1 ) " " " " " " " "   
## 14 ( 1 ) " " " " " " " "   
## 15 ( 1 ) " " " " " " " "   
## 16 ( 1 ) " " " " " " " "   
## 17 ( 1 ) " " " " " " " "   
## 18 ( 1 ) " " " " " " " "   
## 19 ( 1 ) " " " " " " " "   
## 20 ( 1 ) " " " " " " " "   
## 21 ( 1 ) " " " " " " " "   
## 22 ( 1 ) " " " " " " " "   
## 23 ( 1 ) " " " " " " " "   
## 24 ( 1 ) " " " " " " " "   
## 25 ( 1 ) " " " " " " " "   
## 26 ( 1 ) " " " " " " " "

regular\_subset\_full\_summary$rsq

## [1] 0.2260611 0.2490137 0.3661593 0.4413045 0.4470236 0.4518187 0.4540051  
## [8] 0.4546793 0.4555957 0.4560975 0.4566145 0.4571093 0.4574210 0.4575402  
## [15] 0.4576965 0.4578167 0.4579268 0.4580230 0.4580943 0.4581741 0.4582452  
## [22] 0.4582946 0.4583257 0.4583702 0.4584169 0.4584542

par(mfrow =c(2,2))  
plot(regular\_subset\_full\_summary$rss ,xlab=" Number of Variables", ylab=" RSS", type="l")  
plot(regular\_subset\_full\_summary$adjr2 ,xlab =" Number of Variables", ylab=" Adjusted RSq",type="l")  
points (which.max(regular\_subset\_full\_summary$adjr2), regular\_subset\_full\_summary$adjr2[which.max(regular\_subset\_full\_summary$adjr2)], col ="red",cex =2, pch =20)  
plot(regular\_subset\_full\_summary$cp ,xlab =" Number of Variables ",ylab="Cp", type="l")  
points (which.min (regular\_subset\_full\_summary$cp ), regular\_subset\_full\_summary$cp [which.min (regular\_subset\_full\_summary$cp )], col ="red",cex =2, pch =20)  
plot(regular\_subset\_full\_summary$bic ,xlab=" Number of Variables ",ylab="BIC", type="l")  
points(which.min (regular\_subset\_full\_summary$bic), regular\_subset\_full\_summary$bic[which.min(regular\_subset\_full\_summary$bic )], col ="red",cex =2, pch =20)



Forward and backward subset selection test.

forward\_subset\_full <- regsubsets(is\_acct\_delinquent∼., ds\_lc, nvmax = 25, method="forward")

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 3 linear dependencies found

## Reordering variables and trying again:

summary(forward\_subset\_full)

## Subset selection object  
## Call: regsubsets.formula(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 25,   
## method = "forward")  
## 41 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## funded\_amnt\_inv FALSE FALSE  
## term FALSE FALSE  
## int\_rate FALSE FALSE  
## installment FALSE FALSE  
## grade FALSE FALSE  
## sub\_grade FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_low FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## mths\_since\_last\_delinq FALSE FALSE  
## mths\_since\_last\_record FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## revol\_util FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_pymnt\_inv FALSE FALSE  
## total\_rec\_prncp FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## last\_fico\_range\_low FALSE FALSE  
## mths\_since\_last\_major\_derog FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## avg\_cur\_bal FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## funded\_amnt FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## 1 subsets of each size up to 26  
## Selection Algorithm: forward  
## loan\_amnt funded\_amnt funded\_amnt\_inv term int\_rate installment grade  
## 1 ( 1 ) " " " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " " " " " "   
## 6 ( 1 ) " " " " "\*" " " " " " " " "   
## 7 ( 1 ) " " " " "\*" " " " " " " " "   
## 8 ( 1 ) " " " " "\*" " " " " " " " "   
## 9 ( 1 ) " " " " "\*" " " " " " " " "   
## 10 ( 1 ) " " " " "\*" " " " " " " " "   
## 11 ( 1 ) " " " " "\*" " " " " " " " "   
## 12 ( 1 ) " " " " "\*" " " " " " " " "   
## 13 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 14 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 15 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 16 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 17 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 18 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 19 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 20 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 21 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 22 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 23 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 24 ( 1 ) " " " " "\*" " " "\*" " " " "   
## 25 ( 1 ) " " " " "\*" " " "\*" " " "\*"   
## 26 ( 1 ) " " " " "\*" " " "\*" " " "\*"   
## sub\_grade home\_ownership annual\_inc dti delinq\_2yrs fico\_range\_low  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " " " "   
## 7 ( 1 ) " " " " " " " " " " " "   
## 8 ( 1 ) " " " " " " " " " " " "   
## 9 ( 1 ) " " " " " " " " " " " "   
## 10 ( 1 ) " " " " " " " " " " " "   
## 11 ( 1 ) "\*" " " " " " " " " " "   
## 12 ( 1 ) "\*" " " " " " " " " " "   
## 13 ( 1 ) "\*" " " " " " " " " " "   
## 14 ( 1 ) "\*" " " "\*" " " " " " "   
## 15 ( 1 ) "\*" " " "\*" " " " " " "   
## 16 ( 1 ) "\*" " " "\*" " " " " " "   
## 17 ( 1 ) "\*" " " "\*" " " " " " "   
## 18 ( 1 ) "\*" " " "\*" " " " " " "   
## 19 ( 1 ) "\*" " " "\*" " " " " " "   
## 20 ( 1 ) "\*" " " "\*" " " " " " "   
## 21 ( 1 ) "\*" " " "\*" " " " " " "   
## 22 ( 1 ) "\*" " " "\*" " " " " " "   
## 23 ( 1 ) "\*" " " "\*" "\*" " " " "   
## 24 ( 1 ) "\*" "\*" "\*" "\*" " " " "   
## 25 ( 1 ) "\*" "\*" "\*" "\*" " " " "   
## 26 ( 1 ) "\*" "\*" "\*" "\*" " " " "   
## fico\_range\_high inq\_last\_6mths mths\_since\_last\_delinq  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) "\*" " " " "   
## 4 ( 1 ) "\*" " " " "   
## 5 ( 1 ) "\*" " " " "   
## 6 ( 1 ) "\*" " " " "   
## 7 ( 1 ) "\*" " " " "   
## 8 ( 1 ) "\*" " " " "   
## 9 ( 1 ) "\*" " " " "   
## 10 ( 1 ) "\*" " " " "   
## 11 ( 1 ) "\*" " " " "   
## 12 ( 1 ) "\*" " " " "   
## 13 ( 1 ) "\*" " " " "   
## 14 ( 1 ) "\*" " " " "   
## 15 ( 1 ) "\*" " " " "   
## 16 ( 1 ) "\*" " " " "   
## 17 ( 1 ) "\*" " " " "   
## 18 ( 1 ) "\*" " " " "   
## 19 ( 1 ) "\*" " " " "   
## 20 ( 1 ) "\*" " " " "   
## 21 ( 1 ) "\*" " " " "   
## 22 ( 1 ) "\*" " " "\*"   
## 23 ( 1 ) "\*" " " "\*"   
## 24 ( 1 ) "\*" " " "\*"   
## 25 ( 1 ) "\*" " " "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## mths\_since\_last\_record open\_acc pub\_rec revol\_bal revol\_util  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " "   
## 7 ( 1 ) " " " " " " " " " "   
## 8 ( 1 ) " " " " " " " " " "   
## 9 ( 1 ) "\*" " " " " " " " "   
## 10 ( 1 ) "\*" " " " " " " " "   
## 11 ( 1 ) "\*" " " " " " " " "   
## 12 ( 1 ) "\*" " " " " " " " "   
## 13 ( 1 ) "\*" " " " " " " " "   
## 14 ( 1 ) "\*" " " " " " " " "   
## 15 ( 1 ) "\*" " " " " "\*" " "   
## 16 ( 1 ) "\*" " " " " "\*" " "   
## 17 ( 1 ) "\*" " " " " "\*" " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " "   
## 19 ( 1 ) "\*" "\*" " " "\*" " "   
## 20 ( 1 ) "\*" "\*" "\*" "\*" " "   
## 21 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_pymnt\_inv  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " "   
## 7 ( 1 ) " " "\*" " " " " " "   
## 8 ( 1 ) " " "\*" " " " " " "   
## 9 ( 1 ) " " "\*" " " " " " "   
## 10 ( 1 ) " " "\*" " " " " " "   
## 11 ( 1 ) " " "\*" " " " " " "   
## 12 ( 1 ) " " "\*" " " " " "\*"   
## 13 ( 1 ) " " "\*" " " " " "\*"   
## 14 ( 1 ) " " "\*" " " " " "\*"   
## 15 ( 1 ) " " "\*" " " " " "\*"   
## 16 ( 1 ) " " "\*" " " " " "\*"   
## 17 ( 1 ) "\*" "\*" " " " " "\*"   
## 18 ( 1 ) "\*" "\*" " " " " "\*"   
## 19 ( 1 ) "\*" "\*" " " " " "\*"   
## 20 ( 1 ) "\*" "\*" " " " " "\*"   
## 21 ( 1 ) "\*" "\*" " " " " "\*"   
## 22 ( 1 ) "\*" "\*" " " " " "\*"   
## 23 ( 1 ) "\*" "\*" " " " " "\*"   
## 24 ( 1 ) "\*" "\*" " " " " "\*"   
## 25 ( 1 ) "\*" "\*" " " " " "\*"   
## 26 ( 1 ) "\*" "\*" " " " " "\*"   
## total\_rec\_prncp total\_rec\_int total\_rec\_late\_fee recoveries  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " "\*"   
## 3 ( 1 ) " " " " " " "\*"   
## 4 ( 1 ) " " " " "\*" "\*"   
## 5 ( 1 ) " " " " "\*" "\*"   
## 6 ( 1 ) " " " " "\*" "\*"   
## 7 ( 1 ) " " " " "\*" "\*"   
## 8 ( 1 ) "\*" " " "\*" "\*"   
## 9 ( 1 ) "\*" " " "\*" "\*"   
## 10 ( 1 ) "\*" " " "\*" "\*"   
## 11 ( 1 ) "\*" " " "\*" "\*"   
## 12 ( 1 ) "\*" " " "\*" "\*"   
## 13 ( 1 ) "\*" " " "\*" "\*"   
## 14 ( 1 ) "\*" " " "\*" "\*"   
## 15 ( 1 ) "\*" " " "\*" "\*"   
## 16 ( 1 ) "\*" " " "\*" "\*"   
## 17 ( 1 ) "\*" " " "\*" "\*"   
## 18 ( 1 ) "\*" " " "\*" "\*"   
## 19 ( 1 ) "\*" " " "\*" "\*"   
## 20 ( 1 ) "\*" " " "\*" "\*"   
## 21 ( 1 ) "\*" " " "\*" "\*"   
## 22 ( 1 ) "\*" " " "\*" "\*"   
## 23 ( 1 ) "\*" " " "\*" "\*"   
## 24 ( 1 ) "\*" " " "\*" "\*"   
## 25 ( 1 ) "\*" " " "\*" "\*"   
## 26 ( 1 ) "\*" " " "\*" "\*"   
## collection\_recovery\_fee last\_fico\_range\_high last\_fico\_range\_low  
## 1 ( 1 ) " " "\*" " "   
## 2 ( 1 ) " " "\*" " "   
## 3 ( 1 ) " " "\*" " "   
## 4 ( 1 ) " " "\*" " "   
## 5 ( 1 ) " " "\*" "\*"   
## 6 ( 1 ) " " "\*" "\*"   
## 7 ( 1 ) " " "\*" "\*"   
## 8 ( 1 ) " " "\*" "\*"   
## 9 ( 1 ) " " "\*" "\*"   
## 10 ( 1 ) " " "\*" "\*"   
## 11 ( 1 ) " " "\*" "\*"   
## 12 ( 1 ) " " "\*" "\*"   
## 13 ( 1 ) " " "\*" "\*"   
## 14 ( 1 ) " " "\*" "\*"   
## 15 ( 1 ) " " "\*" "\*"   
## 16 ( 1 ) "\*" "\*" "\*"   
## 17 ( 1 ) "\*" "\*" "\*"   
## 18 ( 1 ) "\*" "\*" "\*"   
## 19 ( 1 ) "\*" "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## mths\_since\_last\_major\_derog acc\_now\_delinq avg\_cur\_bal inq\_last\_12m  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) "\*" " " " " " "   
## 11 ( 1 ) "\*" " " " " " "   
## 12 ( 1 ) "\*" " " " " " "   
## 13 ( 1 ) "\*" " " " " " "   
## 14 ( 1 ) "\*" " " " " " "   
## 15 ( 1 ) "\*" " " " " " "   
## 16 ( 1 ) "\*" " " " " " "   
## 17 ( 1 ) "\*" " " " " " "   
## 18 ( 1 ) "\*" " " " " " "   
## 19 ( 1 ) "\*" " " "\*" " "   
## 20 ( 1 ) "\*" " " "\*" " "   
## 21 ( 1 ) "\*" " " "\*" " "   
## 22 ( 1 ) "\*" " " "\*" " "   
## 23 ( 1 ) "\*" " " "\*" " "   
## 24 ( 1 ) "\*" " " "\*" " "   
## 25 ( 1 ) "\*" " " "\*" " "   
## 26 ( 1 ) "\*" " " "\*" " "   
## num\_tl\_30dpd num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) " " " " " " " "   
## 11 ( 1 ) " " " " " " " "   
## 12 ( 1 ) " " " " " " " "   
## 13 ( 1 ) " " " " " " " "   
## 14 ( 1 ) " " " " " " " "   
## 15 ( 1 ) " " " " " " " "   
## 16 ( 1 ) " " " " " " " "   
## 17 ( 1 ) " " " " " " " "   
## 18 ( 1 ) " " " " " " " "   
## 19 ( 1 ) " " " " " " " "   
## 20 ( 1 ) " " " " " " " "   
## 21 ( 1 ) " " " " " " " "   
## 22 ( 1 ) " " " " " " " "   
## 23 ( 1 ) " " " " " " " "   
## 24 ( 1 ) " " " " " " " "   
## 25 ( 1 ) " " " " " " " "   
## 26 ( 1 ) " " " " " " " "

backward\_subset\_full <- regsubsets(is\_acct\_delinquent∼., ds\_lc, nvmax = 25, method="backward")

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =  
## force.in, : 3 linear dependencies found

## Reordering variables and trying again:

summary(backward\_subset\_full)

## Subset selection object  
## Call: regsubsets.formula(is\_acct\_delinquent ~ ., ds\_lc, nvmax = 25,   
## method = "backward")  
## 41 Variables (and intercept)  
## Forced in Forced out  
## loan\_amnt FALSE FALSE  
## funded\_amnt\_inv FALSE FALSE  
## term FALSE FALSE  
## int\_rate FALSE FALSE  
## installment FALSE FALSE  
## grade FALSE FALSE  
## sub\_grade FALSE FALSE  
## home\_ownership FALSE FALSE  
## annual\_inc FALSE FALSE  
## dti FALSE FALSE  
## delinq\_2yrs FALSE FALSE  
## fico\_range\_low FALSE FALSE  
## inq\_last\_6mths FALSE FALSE  
## mths\_since\_last\_delinq FALSE FALSE  
## mths\_since\_last\_record FALSE FALSE  
## open\_acc FALSE FALSE  
## pub\_rec FALSE FALSE  
## revol\_bal FALSE FALSE  
## revol\_util FALSE FALSE  
## total\_acc FALSE FALSE  
## out\_prncp FALSE FALSE  
## out\_prncp\_inv FALSE FALSE  
## total\_pymnt FALSE FALSE  
## total\_pymnt\_inv FALSE FALSE  
## total\_rec\_prncp FALSE FALSE  
## total\_rec\_int FALSE FALSE  
## total\_rec\_late\_fee FALSE FALSE  
## recoveries FALSE FALSE  
## collection\_recovery\_fee FALSE FALSE  
## last\_fico\_range\_high FALSE FALSE  
## last\_fico\_range\_low FALSE FALSE  
## mths\_since\_last\_major\_derog FALSE FALSE  
## acc\_now\_delinq FALSE FALSE  
## avg\_cur\_bal FALSE FALSE  
## inq\_last\_12m FALSE FALSE  
## num\_tl\_90g\_dpd\_24m FALSE FALSE  
## tot\_hi\_cred\_lim FALSE FALSE  
## num\_rev\_accts FALSE FALSE  
## funded\_amnt FALSE FALSE  
## fico\_range\_high FALSE FALSE  
## num\_tl\_30dpd FALSE FALSE  
## 1 subsets of each size up to 26  
## Selection Algorithm: backward  
## loan\_amnt funded\_amnt funded\_amnt\_inv term int\_rate installment grade  
## 1 ( 1 ) " " " " " " " " " " " " " "   
## 2 ( 1 ) "\*" " " " " " " " " " " " "   
## 3 ( 1 ) "\*" " " " " " " " " " " " "   
## 4 ( 1 ) "\*" " " " " " " " " " " " "   
## 5 ( 1 ) "\*" " " " " " " " " " " " "   
## 6 ( 1 ) "\*" " " " " " " " " " " " "   
## 7 ( 1 ) "\*" " " " " " " " " " " " "   
## 8 ( 1 ) "\*" " " " " " " " " " " " "   
## 9 ( 1 ) "\*" " " " " " " " " " " " "   
## 10 ( 1 ) "\*" " " " " " " " " " " " "   
## 11 ( 1 ) "\*" " " " " " " " " " " " "   
## 12 ( 1 ) "\*" " " " " " " " " " " " "   
## 13 ( 1 ) "\*" " " " " " " "\*" " " " "   
## 14 ( 1 ) "\*" " " " " "\*" "\*" " " " "   
## 15 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 16 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 17 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 18 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 19 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 20 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 21 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 22 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 23 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 24 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 25 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## 26 ( 1 ) "\*" " " " " "\*" "\*" "\*" " "   
## sub\_grade home\_ownership annual\_inc dti delinq\_2yrs fico\_range\_low  
## 1 ( 1 ) " " " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " " " "   
## 7 ( 1 ) " " " " " " " " " " " "   
## 8 ( 1 ) " " " " " " " " " " "\*"   
## 9 ( 1 ) " " " " " " " " " " "\*"   
## 10 ( 1 ) "\*" " " " " " " " " "\*"   
## 11 ( 1 ) "\*" " " " " " " " " "\*"   
## 12 ( 1 ) "\*" " " " " " " " " "\*"   
## 13 ( 1 ) "\*" " " " " " " " " "\*"   
## 14 ( 1 ) "\*" " " " " " " " " "\*"   
## 15 ( 1 ) "\*" " " " " " " " " "\*"   
## 16 ( 1 ) "\*" " " " " " " " " "\*"   
## 17 ( 1 ) "\*" " " " " " " " " "\*"   
## 18 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 19 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 20 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 21 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 22 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 23 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 24 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 25 ( 1 ) "\*" " " "\*" " " " " "\*"   
## 26 ( 1 ) "\*" " " "\*" " " " " "\*"   
## fico\_range\_high inq\_last\_6mths mths\_since\_last\_delinq  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " " " " "   
## 5 ( 1 ) " " " " " "   
## 6 ( 1 ) " " " " " "   
## 7 ( 1 ) " " " " " "   
## 8 ( 1 ) " " " " " "   
## 9 ( 1 ) " " " " " "   
## 10 ( 1 ) " " " " " "   
## 11 ( 1 ) " " " " " "   
## 12 ( 1 ) " " " " " "   
## 13 ( 1 ) " " " " " "   
## 14 ( 1 ) " " " " " "   
## 15 ( 1 ) " " " " " "   
## 16 ( 1 ) " " " " " "   
## 17 ( 1 ) " " " " " "   
## 18 ( 1 ) " " " " " "   
## 19 ( 1 ) " " " " " "   
## 20 ( 1 ) " " " " " "   
## 21 ( 1 ) " " " " " "   
## 22 ( 1 ) " " " " " "   
## 23 ( 1 ) " " " " "\*"   
## 24 ( 1 ) " " " " "\*"   
## 25 ( 1 ) " " "\*" "\*"   
## 26 ( 1 ) " " "\*" "\*"   
## mths\_since\_last\_record open\_acc pub\_rec revol\_bal revol\_util  
## 1 ( 1 ) " " " " " " " " " "   
## 2 ( 1 ) " " " " " " " " " "   
## 3 ( 1 ) " " " " " " " " " "   
## 4 ( 1 ) " " " " " " " " " "   
## 5 ( 1 ) " " " " " " " " " "   
## 6 ( 1 ) " " " " " " " " " "   
## 7 ( 1 ) " " " " " " " " " "   
## 8 ( 1 ) " " " " " " " " " "   
## 9 ( 1 ) " " " " " " " " " "   
## 10 ( 1 ) " " " " " " " " " "   
## 11 ( 1 ) " " " " " " " " " "   
## 12 ( 1 ) "\*" " " " " " " " "   
## 13 ( 1 ) "\*" " " " " " " " "   
## 14 ( 1 ) "\*" " " " " " " " "   
## 15 ( 1 ) "\*" " " " " " " " "   
## 16 ( 1 ) "\*" " " " " " " " "   
## 17 ( 1 ) "\*" "\*" " " " " " "   
## 18 ( 1 ) "\*" "\*" " " " " " "   
## 19 ( 1 ) "\*" "\*" " " "\*" " "   
## 20 ( 1 ) "\*" "\*" " " "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" " "   
## 22 ( 1 ) "\*" "\*" "\*" "\*" " "   
## 23 ( 1 ) "\*" "\*" "\*" "\*" " "   
## 24 ( 1 ) "\*" "\*" "\*" "\*" " "   
## 25 ( 1 ) "\*" "\*" "\*" "\*" " "   
## 26 ( 1 ) "\*" "\*" "\*" "\*" "\*"   
## total\_acc out\_prncp out\_prncp\_inv total\_pymnt total\_pymnt\_inv  
## 1 ( 1 ) " " "\*" " " " " " "   
## 2 ( 1 ) " " "\*" " " " " " "   
## 3 ( 1 ) " " "\*" " " " " " "   
## 4 ( 1 ) " " "\*" " " " " " "   
## 5 ( 1 ) " " "\*" " " " " " "   
## 6 ( 1 ) " " "\*" " " " " " "   
## 7 ( 1 ) " " "\*" " " "\*" " "   
## 8 ( 1 ) " " "\*" " " "\*" " "   
## 9 ( 1 ) " " "\*" " " "\*" " "   
## 10 ( 1 ) " " "\*" " " "\*" " "   
## 11 ( 1 ) " " "\*" " " "\*" " "   
## 12 ( 1 ) " " "\*" " " "\*" " "   
## 13 ( 1 ) " " "\*" " " "\*" " "   
## 14 ( 1 ) " " "\*" " " "\*" " "   
## 15 ( 1 ) " " "\*" " " "\*" " "   
## 16 ( 1 ) "\*" "\*" " " "\*" " "   
## 17 ( 1 ) "\*" "\*" " " "\*" " "   
## 18 ( 1 ) "\*" "\*" " " "\*" " "   
## 19 ( 1 ) "\*" "\*" " " "\*" " "   
## 20 ( 1 ) "\*" "\*" " " "\*" " "   
## 21 ( 1 ) "\*" "\*" " " "\*" " "   
## 22 ( 1 ) "\*" "\*" " " "\*" " "   
## 23 ( 1 ) "\*" "\*" " " "\*" " "   
## 24 ( 1 ) "\*" "\*" " " "\*" " "   
## 25 ( 1 ) "\*" "\*" " " "\*" " "   
## 26 ( 1 ) "\*" "\*" " " "\*" " "   
## total\_rec\_prncp total\_rec\_int total\_rec\_late\_fee recoveries  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) "\*" " " " " " "   
## 4 ( 1 ) "\*" " " " " " "   
## 5 ( 1 ) "\*" " " " " "\*"   
## 6 ( 1 ) "\*" "\*" " " "\*"   
## 7 ( 1 ) "\*" "\*" " " "\*"   
## 8 ( 1 ) "\*" "\*" " " "\*"   
## 9 ( 1 ) "\*" "\*" " " "\*"   
## 10 ( 1 ) "\*" "\*" " " "\*"   
## 11 ( 1 ) "\*" "\*" " " "\*"   
## 12 ( 1 ) "\*" "\*" " " "\*"   
## 13 ( 1 ) "\*" "\*" " " "\*"   
## 14 ( 1 ) "\*" "\*" " " "\*"   
## 15 ( 1 ) "\*" "\*" " " "\*"   
## 16 ( 1 ) "\*" "\*" " " "\*"   
## 17 ( 1 ) "\*" "\*" " " "\*"   
## 18 ( 1 ) "\*" "\*" " " "\*"   
## 19 ( 1 ) "\*" "\*" " " "\*"   
## 20 ( 1 ) "\*" "\*" " " "\*"   
## 21 ( 1 ) "\*" "\*" " " "\*"   
## 22 ( 1 ) "\*" "\*" " " "\*"   
## 23 ( 1 ) "\*" "\*" " " "\*"   
## 24 ( 1 ) "\*" "\*" " " "\*"   
## 25 ( 1 ) "\*" "\*" " " "\*"   
## 26 ( 1 ) "\*" "\*" " " "\*"   
## collection\_recovery\_fee last\_fico\_range\_high last\_fico\_range\_low  
## 1 ( 1 ) " " " " " "   
## 2 ( 1 ) " " " " " "   
## 3 ( 1 ) " " " " " "   
## 4 ( 1 ) " " "\*" " "   
## 5 ( 1 ) " " "\*" " "   
## 6 ( 1 ) " " "\*" " "   
## 7 ( 1 ) " " "\*" " "   
## 8 ( 1 ) " " "\*" " "   
## 9 ( 1 ) " " "\*" "\*"   
## 10 ( 1 ) " " "\*" "\*"   
## 11 ( 1 ) " " "\*" "\*"   
## 12 ( 1 ) " " "\*" "\*"   
## 13 ( 1 ) " " "\*" "\*"   
## 14 ( 1 ) " " "\*" "\*"   
## 15 ( 1 ) " " "\*" "\*"   
## 16 ( 1 ) " " "\*" "\*"   
## 17 ( 1 ) " " "\*" "\*"   
## 18 ( 1 ) " " "\*" "\*"   
## 19 ( 1 ) " " "\*" "\*"   
## 20 ( 1 ) "\*" "\*" "\*"   
## 21 ( 1 ) "\*" "\*" "\*"   
## 22 ( 1 ) "\*" "\*" "\*"   
## 23 ( 1 ) "\*" "\*" "\*"   
## 24 ( 1 ) "\*" "\*" "\*"   
## 25 ( 1 ) "\*" "\*" "\*"   
## 26 ( 1 ) "\*" "\*" "\*"   
## mths\_since\_last\_major\_derog acc\_now\_delinq avg\_cur\_bal inq\_last\_12m  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) " " " " " " " "   
## 11 ( 1 ) "\*" " " " " " "   
## 12 ( 1 ) "\*" " " " " " "   
## 13 ( 1 ) "\*" " " " " " "   
## 14 ( 1 ) "\*" " " " " " "   
## 15 ( 1 ) "\*" " " " " " "   
## 16 ( 1 ) "\*" " " " " " "   
## 17 ( 1 ) "\*" " " " " " "   
## 18 ( 1 ) "\*" " " " " " "   
## 19 ( 1 ) "\*" " " " " " "   
## 20 ( 1 ) "\*" " " " " " "   
## 21 ( 1 ) "\*" " " "\*" " "   
## 22 ( 1 ) "\*" " " "\*" " "   
## 23 ( 1 ) "\*" " " "\*" " "   
## 24 ( 1 ) "\*" " " "\*" "\*"   
## 25 ( 1 ) "\*" " " "\*" "\*"   
## 26 ( 1 ) "\*" " " "\*" "\*"   
## num\_tl\_30dpd num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1 ( 1 ) " " " " " " " "   
## 2 ( 1 ) " " " " " " " "   
## 3 ( 1 ) " " " " " " " "   
## 4 ( 1 ) " " " " " " " "   
## 5 ( 1 ) " " " " " " " "   
## 6 ( 1 ) " " " " " " " "   
## 7 ( 1 ) " " " " " " " "   
## 8 ( 1 ) " " " " " " " "   
## 9 ( 1 ) " " " " " " " "   
## 10 ( 1 ) " " " " " " " "   
## 11 ( 1 ) " " " " " " " "   
## 12 ( 1 ) " " " " " " " "   
## 13 ( 1 ) " " " " " " " "   
## 14 ( 1 ) " " " " " " " "   
## 15 ( 1 ) " " " " " " " "   
## 16 ( 1 ) " " " " " " " "   
## 17 ( 1 ) " " " " " " " "   
## 18 ( 1 ) " " " " " " " "   
## 19 ( 1 ) " " " " " " " "   
## 20 ( 1 ) " " " " " " " "   
## 21 ( 1 ) " " " " " " " "   
## 22 ( 1 ) " " " " " " " "   
## 23 ( 1 ) " " " " " " " "   
## 24 ( 1 ) " " " " " " " "   
## 25 ( 1 ) " " " " " " " "   
## 26 ( 1 ) " " " " " " " "

Based on the above subset selection test, identified following set of final variables.

ds\_lc <- ds\_lc[,c(  
 "loan\_amnt",  
 "term",  
 "int\_rate",  
 "installment",  
 "grade",  
 "sub\_grade",  
 "home\_ownership",  
 "annual\_inc",  
 "is\_acct\_delinquent",  
 "dti",  
 "delinq\_2yrs",  
 "fico\_range\_low",  
 "fico\_range\_high",  
 "inq\_last\_6mths",  
 "open\_acc",  
 "pub\_rec",  
 "total\_acc",  
 "last\_fico\_range\_high",   
 "last\_fico\_range\_low",   
 "inq\_last\_12m",   
 "num\_tl\_30dpd",   
 "num\_tl\_90g\_dpd\_24m",   
 "tot\_hi\_cred\_lim",   
 "num\_rev\_accts"   
 )]

Check data types for all variables is int or numeric

str(ds\_lc)

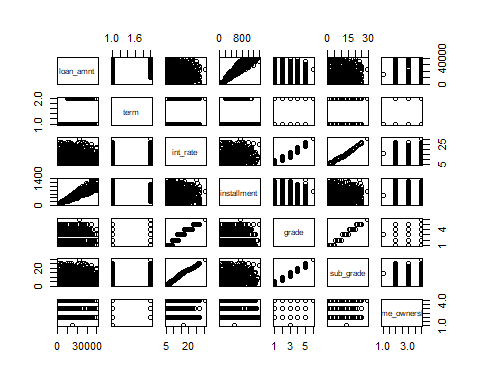
## Classes 'data.table' and 'data.frame': 494110 obs. of 24 variables:  
## $ loan\_amnt : int 5000 15000 11200 25000 3000 17000 20000 19200 6500 10000 ...  
## $ term : int 1 1 2 2 1 2 1 2 1 2 ...  
## $ int\_rate : num 20.39 9.92 30.79 21.85 7.34 ...  
## $ installment : num 186.8 483.4 367.8 688.4 93.1 ...  
## $ grade : int 4 2 7 4 1 4 2 4 1 3 ...  
## $ sub\_grade : int 19 7 31 20 4 19 8 16 2 11 ...  
## $ home\_ownership : int 4 3 4 2 4 4 2 4 2 2 ...  
## $ annual\_inc : num 50000 196000 44000 65000 52000 52000 19000 36500 50000 80000 ...  
## $ is\_acct\_delinquent : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ dti : num 21.8 18.29 43.97 12.89 0.58 ...  
## $ delinq\_2yrs : int 1 0 1 1 0 0 0 3 0 0 ...  
## $ fico\_range\_low : int 665 700 665 665 760 670 795 675 705 660 ...  
## $ fico\_range\_high : int 669 704 669 669 764 674 799 679 709 664 ...  
## $ inq\_last\_6mths : int 0 0 2 1 0 0 1 0 0 1 ...  
## $ open\_acc : int 5 19 8 7 7 9 6 14 7 17 ...  
## $ pub\_rec : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ total\_acc : int 18 53 14 16 30 14 9 18 16 23 ...  
## $ last\_fico\_range\_high: int 609 694 629 669 764 634 799 559 714 634 ...  
## $ last\_fico\_range\_low : int 605 690 625 665 760 630 795 555 710 630 ...  
## $ inq\_last\_12m : int 5 7 10 3 2 2 1 0 0 4 ...  
## $ num\_tl\_30dpd : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ num\_tl\_90g\_dpd\_24m : int 1 0 0 1 0 0 0 3 0 0 ...  
## $ tot\_hi\_cred\_lim : num 33430 605228 80367 101234 191216 ...  
## $ num\_rev\_accts : int 2 37 6 9 19 12 8 15 12 18 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

Summary for all variables in the dataset along with NA check.

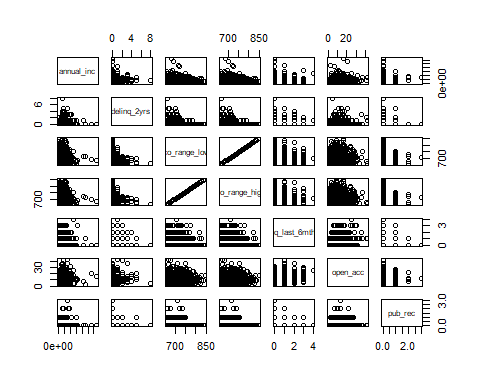
summary(ds\_lc)

## loan\_amnt term int\_rate installment   
## Min. : 1000 Min. :1.000 Min. : 5.31 Min. : 29.76   
## 1st Qu.: 8000 1st Qu.:1.000 1st Qu.: 8.46 1st Qu.: 254.50   
## Median :14000 Median :1.000 Median :11.80 Median : 386.76   
## Mean :16017 Mean :1.304 Mean :12.73 Mean : 466.38   
## 3rd Qu.:22000 3rd Qu.:2.000 3rd Qu.:16.01 3rd Qu.: 628.67   
## Max. :40000 Max. :2.000 Max. :30.99 Max. :1670.15   
## grade sub\_grade home\_ownership annual\_inc   
## Min. :1.000 Min. : 1 Min. :1.00 Min. : 0   
## 1st Qu.:1.000 1st Qu.: 5 1st Qu.:2.00 1st Qu.: 46000   
## Median :2.000 Median : 9 Median :3.00 Median : 66000   
## Mean :2.409 Mean :10 Mean :2.91 Mean : 80277   
## 3rd Qu.:3.000 3rd Qu.:14 3rd Qu.:4.00 3rd Qu.: 96000   
## Max. :7.000 Max. :35 Max. :4.00 Max. :9930475   
## is\_acct\_delinquent dti delinq\_2yrs fico\_range\_low   
## 0:475138 Min. : 0.00 Min. : 0.0000 Min. :660.0   
## 1: 18972 1st Qu.: 11.43 1st Qu.: 0.0000 1st Qu.:680.0   
## Median : 17.71 Median : 0.0000 Median :700.0   
## Mean : 19.67 Mean : 0.2295 Mean :706.4   
## 3rd Qu.: 25.03 3rd Qu.: 0.0000 3rd Qu.:725.0   
## Max. :999.00 Max. :58.0000 Max. :845.0   
## fico\_range\_high inq\_last\_6mths open\_acc pub\_rec   
## Min. :664.0 Min. :0.0000 Min. : 0.0 Min. : 0.0000   
## 1st Qu.:684.0 1st Qu.:0.0000 1st Qu.: 7.0 1st Qu.: 0.0000   
## Median :704.0 Median :0.0000 Median : 10.0 Median : 0.0000   
## Mean :710.4 Mean :0.4426 Mean : 11.5 Mean : 0.1346   
## 3rd Qu.:729.0 3rd Qu.:1.0000 3rd Qu.: 14.0 3rd Qu.: 0.0000   
## Max. :850.0 Max. :5.0000 Max. :101.0 Max. :52.0000   
## total\_acc last\_fico\_range\_high last\_fico\_range\_low inq\_last\_12m   
## Min. : 2.00 Min. : 0.0 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 14.00 1st Qu.:679.0 1st Qu.:675.0 1st Qu.: 0.000   
## Median : 21.00 Median :709.0 Median :705.0 Median : 1.000   
## Mean : 22.64 Mean :709.4 Mean :703.7 Mean : 1.938   
## 3rd Qu.: 29.00 3rd Qu.:744.0 3rd Qu.:740.0 3rd Qu.: 3.000   
## Max. :160.00 Max. :850.0 Max. :845.0 Max. :67.000   
## num\_tl\_30dpd num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts   
## Min. :0.00e+00 Min. : 0.00000 Min. : 0 Min. : 2.00   
## 1st Qu.:0.00e+00 1st Qu.: 0.00000 1st Qu.: 52075 1st Qu.: 7.00   
## Median :0.00e+00 Median : 0.00000 Median : 116346 Median : 11.00   
## Mean :4.65e-05 Mean : 0.06003 Mean : 184961 Mean : 12.92   
## 3rd Qu.:0.00e+00 3rd Qu.: 0.00000 3rd Qu.: 269331 3rd Qu.: 17.00   
## Max. :1.00e+00 Max. :58.00000 Max. :9999999 Max. :151.00

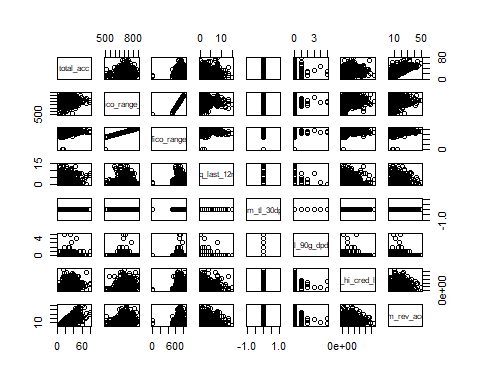
corr\_ds\_subset <- sample(nrow(ds\_lc), 1000)  
corr\_ds <- ds\_lc[corr\_ds\_subset, ]  
pairs(~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership  
 , corr\_ds)



pairs(~   
 annual\_inc +   
 delinq\_2yrs +  
 fico\_range\_low +  
 fico\_range\_high +  
 inq\_last\_6mths +  
 open\_acc +  
 pub\_rec  
 ,corr\_ds)



pairs(~   
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +  
 num\_rev\_accts   
 ,corr\_ds)

 Peak at initial set of records.

head(ds\_lc)

## loan\_amnt term int\_rate installment grade sub\_grade home\_ownership  
## 1: 5000 1 20.39 186.82 4 19 4  
## 2: 15000 1 9.92 483.45 2 7 3  
## 3: 11200 2 30.79 367.82 7 31 4  
## 4: 25000 2 21.85 688.35 4 20 2  
## 5: 3000 1 7.34 93.10 1 4 4  
## 6: 17000 2 20.39 454.10 4 19 4  
## annual\_inc is\_acct\_delinquent dti delinq\_2yrs fico\_range\_low  
## 1: 50000 0 21.80 1 665  
## 2: 196000 0 18.29 0 700  
## 3: 44000 0 43.97 1 665  
## 4: 65000 0 12.89 1 665  
## 5: 52000 0 0.58 0 760  
## 6: 52000 0 15.65 0 670  
## fico\_range\_high inq\_last\_6mths open\_acc pub\_rec total\_acc  
## 1: 669 0 5 0 18  
## 2: 704 0 19 0 53  
## 3: 669 2 8 0 14  
## 4: 669 1 7 0 16  
## 5: 764 0 7 0 30  
## 6: 674 0 9 0 14  
## last\_fico\_range\_high last\_fico\_range\_low inq\_last\_12m num\_tl\_30dpd  
## 1: 609 605 5 0  
## 2: 694 690 7 0  
## 3: 629 625 10 0  
## 4: 669 665 3 0  
## 5: 764 760 2 0  
## 6: 634 630 2 0  
## num\_tl\_90g\_dpd\_24m tot\_hi\_cred\_lim num\_rev\_accts  
## 1: 1 33430 2  
## 2: 0 605228 37  
## 3: 0 80367 6  
## 4: 1 101234 9  
## 5: 0 191216 19  
## 6: 0 32500 12

## Validation set

Identify whats right the subset distribution for training and test dataset

sample\_set<-sample(nrow(ds\_lc),nrow(ds\_lc)\*0.7)  
training\_dataset <- ds\_lc[sample\_set,]  
test\_dataset <- ds\_lc[-sample\_set,]  
regression\_model <- glm(is\_acct\_delinquent ~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership +   
 annual\_inc +   
 dti +   
 delinq\_2yrs +   
 fico\_range\_low +   
 fico\_range\_high +   
 inq\_last\_6mths +   
 open\_acc +   
 pub\_rec,  
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +   
 num\_rev\_accts,   
 data = training\_dataset, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

probablity<-predict(regression\_model,test\_dataset,type="response")  
prediction<-ifelse(probablity > 0.5, 1, 0)  
validation\_error\_rate2 <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
validation\_error\_rate2

## [1] 0.03852044

sample\_set<-sample(nrow(ds\_lc),nrow(ds\_lc)\*0.75)  
training\_dataset <- ds\_lc[sample\_set,]  
test\_dataset <- ds\_lc[-sample\_set,]  
regression\_model <- glm(is\_acct\_delinquent ~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership +   
 annual\_inc +   
 dti +   
 delinq\_2yrs +   
 fico\_range\_low +   
 fico\_range\_high +   
 inq\_last\_6mths +   
 open\_acc +   
 pub\_rec,  
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +   
 num\_rev\_accts,   
 data = training\_dataset, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

probablity<-predict(regression\_model,test\_dataset,type="response")  
prediction<-ifelse(probablity > 0.5, 1, 0)  
validation\_error\_rate3 <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
validation\_error\_rate3

## [1] 0.03838806

sample\_set<-sample(nrow(ds\_lc),nrow(ds\_lc)\*0.8)  
training\_dataset <- ds\_lc[sample\_set,]  
test\_dataset <- ds\_lc[-sample\_set,]  
regression\_model <- glm(is\_acct\_delinquent ~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership +   
 annual\_inc +   
 dti +   
 delinq\_2yrs +   
 fico\_range\_low +   
 fico\_range\_high +   
 inq\_last\_6mths +   
 open\_acc +   
 pub\_rec,  
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +   
 num\_rev\_accts,   
 data = training\_dataset, family = "binomial")

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

probablity<-predict(regression\_model,test\_dataset,type="response")  
prediction<-ifelse(probablity > 0.5, 1, 0)  
validation\_error\_rate4 <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
validation\_error\_rate4

## [1] 0.03825059

Following is the data split in training and test data ration along with validation error rate/misclassification

1. Tranining : Test -> 70 : 30, validation error rate 0.0385204. or 3.8520437%   
2. Tranining : Test -> 75 : 25, validation error rate 0.0383881. or 3.8388058%   
3. Tranining : Test -> 80 : 20, validation error rate 0.0383881. or 3.8250592%

Validation error rate is different for different sample splits of taining and test dataset. This indicates that validation error rate varies by which observations are in the training/validation sets.

Partition the dataset based on results of above validation set approach.

data\_partition <- createDataPartition(y = ds\_lc$is\_acct\_delinquent, p = 0.75, list = FALSE)  
train\_dataset <- ds\_lc[data\_partition,]  
test\_dataset <- ds\_lc[-data\_partition,]

nrow(train\_dataset)

## [1] 370583

nrow(test\_dataset)

## [1] 123527

## Logistic Regression

logistic\_regression <- glm(is\_acct\_delinquent ~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership +   
 annual\_inc +   
 dti +   
 delinq\_2yrs +   
 fico\_range\_low +   
 fico\_range\_high +   
 inq\_last\_6mths +   
 open\_acc +   
 pub\_rec +  
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +   
 num\_rev\_accts   
 ,data = train\_dataset, family = binomial)  
  
predictions <- predict(logistic\_regression, test\_dataset, type="response")  
predicted\_direction <- ifelse(predictions > 0.5, 1, 0)  
error\_rate\_lr <- mean(predicted\_direction != test\_dataset$is\_acct\_delinquent)  
error\_rate\_lr

## [1] 0.02681195

## LDA - Linear Discriminant analysis

lda\_model <- lda(is\_acct\_delinquent ~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership +   
 annual\_inc +   
 dti +   
 delinq\_2yrs +   
 fico\_range\_low +   
 fico\_range\_high +   
 inq\_last\_6mths +   
 open\_acc +   
 pub\_rec +  
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +   
 num\_rev\_accts  
 ,data = train\_dataset)  
predictions <- predict(lda\_model, test\_dataset, type="response")  
confusion\_matrix <- table(predictions$class,   
 test\_dataset$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 117810 2454  
## 1 974 2289

error\_rate\_lda <- mean(predictions$class != test\_dataset$is\_acct\_delinquent)  
error\_rate\_lda

## [1] 0.02775102

## QDA - Quadratic Discriminant Analysis

qda\_model <- qda(is\_acct\_delinquent ~   
 loan\_amnt +   
 term +   
 int\_rate +   
 installment +   
 grade +   
 sub\_grade +   
 home\_ownership +   
 annual\_inc +   
 dti +   
 delinq\_2yrs +   
 fico\_range\_low +   
 fico\_range\_high +   
 inq\_last\_6mths +   
 open\_acc +   
 pub\_rec +  
 total\_acc +  
 last\_fico\_range\_high +  
 last\_fico\_range\_low +  
 inq\_last\_12m +  
 num\_tl\_30dpd +   
 num\_tl\_90g\_dpd\_24m +  
 tot\_hi\_cred\_lim +   
 num\_rev\_accts,   
 data = train\_dataset)  
predictions <- predict(qda\_model, test\_dataset, type="response")  
confusion\_matrix <- table(predictions$class,   
 test\_dataset$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 117721 2983  
## 1 1063 1760

error\_rate\_qda <- mean(predictions$class != test\_dataset$is\_acct\_delinquent)  
error\_rate\_qda

## [1] 0.03275397

## KNN

train\_dataset\_matrix <- as.matrix(head(train\_dataset, 80000))  
test\_dataset\_matrix <- as.matrix(head(test\_dataset, 80000))  
train\_dataset\_mini <- head(train\_dataset, 80000)  
predictions <- knn(train\_dataset\_matrix, test\_dataset\_matrix, train\_dataset\_mini$is\_acct\_delinquent, 10)  
confusion\_matrix <- table(predictions,   
 train\_dataset\_mini$is\_acct\_delinquent,   
 dnn = c("Predicted Status", "Observed Status"))  
confusion\_matrix

## Observed Status  
## Predicted Status 0 1  
## 0 74661 5317  
## 1 20 2

error\_rate\_knn <- mean(predictions != train\_dataset\_mini$is\_acct\_delinquent)  
error\_rate\_knn

## [1] 0.0667125

## Tree

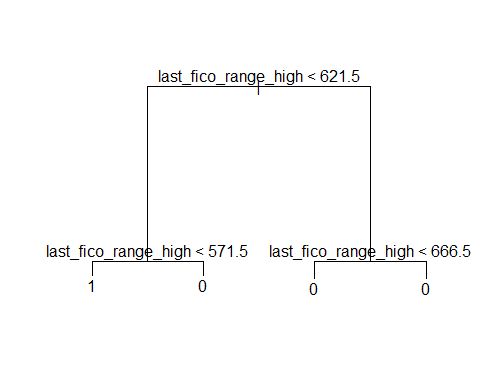
tree\_lc <- tree(is\_acct\_delinquent ~ .,train\_dataset)  
summary(tree\_lc)

##   
## Classification tree:  
## tree(formula = is\_acct\_delinquent ~ ., data = train\_dataset)  
## Variables actually used in tree construction:  
## [1] "last\_fico\_range\_high"  
## Number of terminal nodes: 4   
## Residual mean deviance: 0.1976 = 73220 / 370600   
## Misclassification error rate: 0.02797 = 10364 / 370583

tree\_lc

## node), split, n, deviance, yval, (yprob)  
## \* denotes terminal node  
##   
## 1) root 370583 120700 0 ( 0.961604 0.038396 )   
## 2) last\_fico\_range\_high < 621.5 16315 22570 1 ( 0.472939 0.527061 )   
## 4) last\_fico\_range\_high < 571.5 6941 7342 1 ( 0.221582 0.778418 ) \*  
## 5) last\_fico\_range\_high > 571.5 9374 12030 0 ( 0.659057 0.340943 ) \*  
## 3) last\_fico\_range\_high > 621.5 354268 57810 0 ( 0.984108 0.015892 )   
## 6) last\_fico\_range\_high < 666.5 40837 18850 0 ( 0.938585 0.061415 ) \*  
## 7) last\_fico\_range\_high > 666.5 313431 34990 0 ( 0.990039 0.009961 ) \*

plot(tree\_lc)  
text(tree\_lc, pretty=0)



prediction <- predict(tree\_lc, test\_dataset, type="class")  
error\_rate\_tree <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
error\_rate\_tree

## [1] 0.0283096

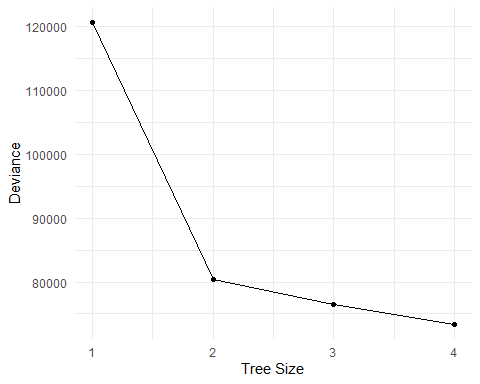
The prime indicator/factor for is loan account delinquent is last\_fico\_range\_high When high fico reaches below 581, then probablity of loan account becoming delinquent is high.

optimal\_tree <- cv.tree(tree\_lc, FUN = prune.tree)  
optimal\_tree

## $size  
## [1] 4 3 2 1  
##   
## $dev  
## [1] 73371.35 76515.52 80381.44 120672.15  
##   
## $k  
## [1] -Inf 3197.527 3962.416 40293.862  
##   
## $method  
## [1] "deviance"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

OPtimal tree size is 4.

tree\_plot <- data.frame(x=optimal\_tree$size, y=optimal\_tree$dev)  
ggplot(tree\_plot, aes(x=x,y=y)) +   
 geom\_point() +   
 geom\_line() +   
 xlab("Tree Size") +   
 ylab("Deviance") +  
 theme\_minimal()



pruned\_tree <- prune.tree(tree\_lc, best = 4)  
optimal\_tree

## $size  
## [1] 4 3 2 1  
##   
## $dev  
## [1] 73371.35 76515.52 80381.44 120672.15  
##   
## $k  
## [1] -Inf 3197.527 3962.416 40293.862  
##   
## $method  
## [1] "deviance"  
##   
## attr(,"class")  
## [1] "prune" "tree.sequence"

#optimal\_tree <- cv.tree(tree\_lc, FUN = prune.tree)  
#optimal\_tree

summary(pruned\_tree)

##   
## Classification tree:  
## tree(formula = is\_acct\_delinquent ~ ., data = train\_dataset)  
## Variables actually used in tree construction:  
## [1] "last\_fico\_range\_high"  
## Number of terminal nodes: 4   
## Residual mean deviance: 0.1976 = 73220 / 370600   
## Misclassification error rate: 0.02797 = 10364 / 370583

summary(tree\_lc)

##   
## Classification tree:  
## tree(formula = is\_acct\_delinquent ~ ., data = train\_dataset)  
## Variables actually used in tree construction:  
## [1] "last\_fico\_range\_high"  
## Number of terminal nodes: 4   
## Residual mean deviance: 0.1976 = 73220 / 370600   
## Misclassification error rate: 0.02797 = 10364 / 370583

test\_error\_rate\_pruned <- mean(predict(pruned\_tree, test\_dataset, type = "class") != test\_dataset$is\_acct\_delinquent)  
test\_error\_rate\_pruned

## [1] 0.0283096

test\_error\_rate\_unpruned <- mean(prediction != test\_dataset$is\_acct\_delinquent)  
test\_error\_rate\_unpruned

## [1] 0.0283096

Test error rate of pruned tree is 0.0283096 and unpruned tree is 0.0283096.

Tree error rate is 0.0283096. Logistic regression error rate is 0.026812. LDA error rate is 0.027751. QDA error rate is 0.032754. KNN error rate is 0.0667125.

All the error rates are almost identical except QDA and KNN, but KNN did not run agains full dataset, so ignoring KNN based models, QDA error rate is bit on higher side, so models to predict loan account delinquency would be logistic regression, LDA or Tree

# Predicting the Interest Rate

## Data Preparation, reset the training and test dataset, identify the interest mean for test dataset

train\_dataset\_subset <- sample(nrow(ds\_lc) \* 0.75)  
train\_dataset <- ds\_lc[train\_dataset\_subset, ]  
test\_dataset <- ds\_lc[-train\_dataset\_subset, ]  
nrow(train\_dataset)

## [1] 370582

nrow(test\_dataset)

## [1] 123528

test\_interest\_mean <- mean(test\_dataset$int\_rate)  
test\_interest\_mse <- mean((test\_dataset$int\_rate - test\_interest\_mean)^2)  
test\_interest\_mean

## [1] 12.55028

test\_interest\_mse

## [1] 25.85992

## Fit a linear model using least squares on the training set, and identify the test error obtained.

lm\_fit <- lm(int\_rate ~ . , data = train\_dataset)  
lm\_predictions <- predict(lm\_fit, test\_dataset)  
lm\_mse <- mean((lm\_predictions - test\_dataset$int\_rate)^2)  
lm\_mse

## [1] 0.4408567

## Fit a ridge regression model and identify the test error obtained.

train\_ds\_matrix <- model.matrix(int\_rate ~ ., data = train\_dataset)  
test\_ds\_matrix <- model.matrix(int\_rate ~ ., data = test\_dataset)  
  
grid <- 10 ^ seq(4, -2, length = 100)  
  
ridge\_reg\_model <- cv.glmnet(train\_ds\_matrix, train\_dataset$int\_rate, alpha = 0, lambda = grid, thresh = 1e-12)  
ridge\_reg\_predictions <- predict(ridge\_reg\_model, test\_ds\_matrix, s = ridge\_reg\_model$lambda.min)  
ridge\_mse <- mean((test\_dataset$int\_rate - ridge\_reg\_predictions)^2)  
ridge\_mse

## [1] 0.4445531

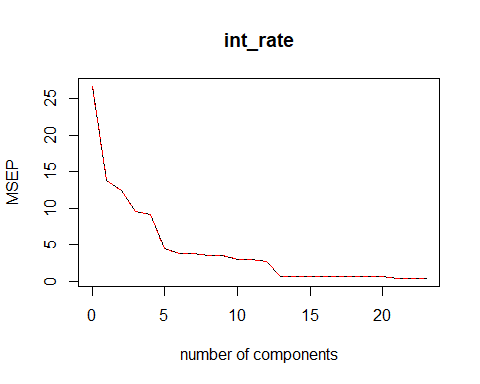
## Fit a Lasso model and identify the test error obtained.

lasso\_model <- cv.glmnet(train\_ds\_matrix, train\_dataset$int\_rate, alpha = 1, lambda = grid, thresh = 1e-12)  
lasso\_predictions <- predict(lasso\_model, test\_ds\_matrix, s = lasso\_model$lambda.min)  
lasso\_mse <- mean((test\_dataset$int\_rate - lasso\_predictions)^2)  
lasso\_rmse <- sqrt(mean((test\_dataset$int\_rate - lasso\_predictions)^2))  
lasso\_mse

## [1] 0.4593279

## Fit a PCR model and identify the test error obtained.

pcr\_model <- pcr(int\_rate ~ . , data = train\_dataset, scale=T, validation="CV")  
validationplot(pcr\_model, val.type = "MSEP")

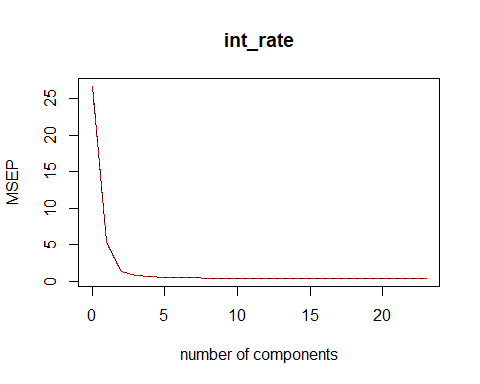


pcr\_predictions <- predict(pcr\_model, test\_dataset, ncomp = 15)  
pcr\_mse <- mean((test\_dataset$int\_rate - pcr\_predictions)^2)  
pcr\_rmse <- sqrt(mean((test\_dataset$int\_rate - pcr\_predictions)^2))  
pcr\_mse

## [1] 0.6963504

## Fit a PLS model and identify the test error obtained.

pls\_model <- plsr(int\_rate ~ . , data = train\_dataset, scale=T, validation="CV")  
validationplot(pls\_model, val.type = "MSEP")



pls\_predictions <- predict(pls\_model, test\_dataset, ncomp = 15)  
pls\_mse <- mean((test\_dataset$int\_rate - pls\_predictions)^2)  
pls\_rmse <- sqrt(mean((test\_dataset$int\_rate - pls\_predictions)^2))  
pls\_mse

## [1] 0.4419671

Compare R2 values for various models

lm\_test\_r2 <- (1 - (lm\_mse/test\_interest\_mse))  
ridge\_test\_r2 <- (1 - (ridge\_mse/test\_interest\_mse))  
lasso\_test\_r2 <- (1 - (lasso\_mse/test\_interest\_mse))  
pcr\_test\_r2 <- (1 - (pcr\_mse/test\_interest\_mse))  
pls\_test\_r2 <- (1 - (pls\_mse/test\_interest\_mse))  
  
cat("R square with linear model : ", lm\_test\_r2, "\n")

## R square with linear model : 0.9829521

cat("R square with ridge model : ", ridge\_test\_r2, "\n")

## R square with ridge model : 0.9828092

cat("R square with lasso model : ", lasso\_test\_r2, "\n")

## R square with lasso model : 0.9822378

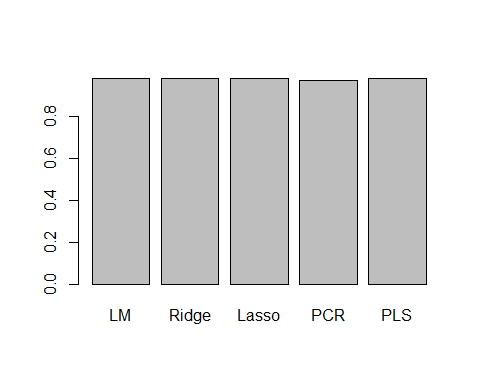
cat("R square with pcr : ", pcr\_test\_r2, "\n")

## R square with pcr : 0.9730722

cat("R square with pls : ", pls\_test\_r2, "\n")

## R square with pls : 0.9829092

barplot(c(lm\_test\_r2, ridge\_test\_r2, lasso\_test\_r2, pcr\_test\_r2, pls\_test\_r2),  
 names.arg = c("LM", "Ridge", "Lasso", "PCR", "PLS"))



All models have R Square values near to 0.9. All models predict interest rate with high accuracy, pcr has lesser accuracy than others.

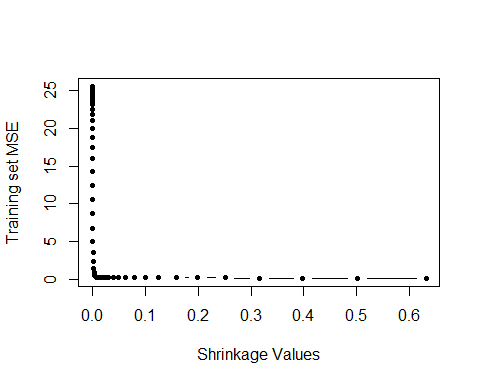
## Boosting

Changed the dataset size due to systme constraints

train\_dataset\_mini <- train\_dataset[1:1000,]  
test\_dataset\_mini <- test\_dataset[(1:1000),]  
  
pows <- seq(-10, -0.2, by = 0.1)  
lambdas <- 10^pows  
training\_errors <- rep(NA, length(lambdas))  
  
  
for (i in 1:length(lambdas)) {  
 boosting\_model <- gbm(int\_rate ~ . , data = train\_dataset\_mini, distribution = "gaussian",   
 n.trees = 1000, shrinkage = lambdas[i])  
   
 training\_predictions <- predict(boosting\_model, train\_dataset\_mini, n.trees = 1000)  
 training\_errors[i] <- mean((training\_predictions - train\_dataset\_mini$int\_rate)^2)  
}

## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.

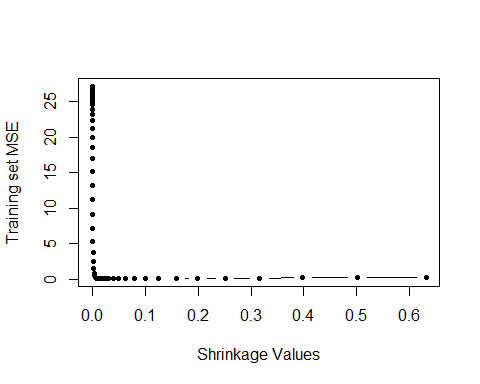
plot(lambdas, training\_errors, xlab = "Shrinkage Values", ylab = "Training set MSE", type = "b", pch = 20)



test\_errors <- rep(NA, length(lambdas))  
  
  
for (i in 1:length(lambdas)) {  
 boosting\_model <- gbm(int\_rate ~ . , data = train\_dataset\_mini, distribution = "gaussian",   
 n.trees = 1000, shrinkage = lambdas[i])  
   
 test\_predictions <- predict(boosting\_model, test\_dataset\_mini, n.trees = 1000)  
 test\_errors[i] <- mean((test\_predictions - test\_dataset\_mini$int\_rate)^2)  
}

## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.  
  
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.

plot(lambdas, test\_errors, xlab = "Shrinkage Values", ylab = "Training set MSE", type = "b", pch = 20)



min(test\_errors)

## [1] 0.1504024

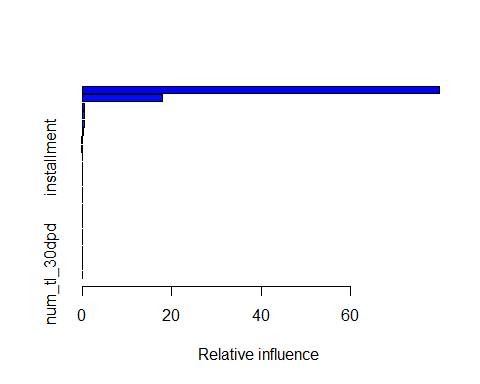
min\_test\_err <- min(test\_errors)  
min\_test\_err\_at <- lambdas[which.min(test\_errors)]

Both regression approaches lm and lasso have higher MSE compared to that of boosting.

boosting\_model <- gbm(int\_rate ~ . , data = train\_dataset\_mini, distribution = "gaussian",   
 n.trees = 1000, shrinkage = min\_test\_err)

## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :  
## variable 20: num\_tl\_30dpd has no variation.

summary(boosting\_model)



## var rel.inf  
## sub\_grade sub\_grade 79.891036299  
## grade grade 17.942497876  
## annual\_inc annual\_inc 0.582023852  
## total\_acc total\_acc 0.410452983  
## dti dti 0.384645610  
## tot\_hi\_cred\_lim tot\_hi\_cred\_lim 0.180489477  
## last\_fico\_range\_high last\_fico\_range\_high 0.140427731  
## num\_rev\_accts num\_rev\_accts 0.122036227  
## installment installment 0.097504361  
## fico\_range\_low fico\_range\_low 0.053584541  
## is\_acct\_delinquent is\_acct\_delinquent 0.047294498  
## delinq\_2yrs delinq\_2yrs 0.034917329  
## open\_acc open\_acc 0.028521733  
## loan\_amnt loan\_amnt 0.027860456  
## inq\_last\_12m inq\_last\_12m 0.022732902  
## term term 0.013150538  
## num\_tl\_90g\_dpd\_24m num\_tl\_90g\_dpd\_24m 0.011429789  
## home\_ownership home\_ownership 0.007024980  
## inq\_last\_6mths inq\_last\_6mths 0.002368817  
## fico\_range\_high fico\_range\_high 0.000000000  
## pub\_rec pub\_rec 0.000000000  
## last\_fico\_range\_low last\_fico\_range\_low 0.000000000  
## num\_tl\_30dpd num\_tl\_30dpd 0.000000000

Variables that appear to be most important predictors are 1. sub\_grade 2. grade 3. open\_acc 4. num\_rev\_Accts 5. annual\_inc

## Bagging

random\_forest\_model <- randomForest(int\_rate ~ . , data = train\_dataset\_mini, ntree = 500, mtry = ncol(train\_dataset\_mini)-1)  
random\_forest\_predictions <- predict(random\_forest\_model, test\_dataset\_mini)  
  
random\_forest\_test\_mse <- mean((random\_forest\_predictions - test\_dataset\_mini$int\_rate)^2)  
random\_forest\_test\_mse

## [1] 0.1481762

Test MSE for bagging is 0.1481762 which is better than 0.1504024 which is best MSE from boosting

# Unspervised Learning - PCA

ds\_lc\_pca <- ds\_lc[,c(  
 "loan\_amnt",  
 "term",  
 "int\_rate",  
 "installment",  
 "grade",  
 "sub\_grade",  
 "home\_ownership",  
 "annual\_inc",  
 "dti",  
 "delinq\_2yrs",  
 "fico\_range\_low",  
 "fico\_range\_high",  
 "inq\_last\_6mths",  
 "open\_acc",  
 "pub\_rec",  
 "total\_acc",  
 "last\_fico\_range\_high",   
 "last\_fico\_range\_low",   
 "inq\_last\_12m",   
 "tot\_hi\_cred\_lim",   
 "num\_rev\_accts"   
 )]  
ds\_lc\_pca\_subset <- sample(nrow(ds\_lc), 1000)  
ds\_lc\_pca = ds\_lc\_pca[ds\_lc\_pca\_subset, ]  
  
str(ds\_lc\_pca)

## Classes 'data.table' and 'data.frame': 1000 obs. of 21 variables:  
## $ loan\_amnt : int 16000 3000 12000 35000 40000 10800 20000 4000 16000 8400 ...  
## $ term : int 1 1 2 1 1 1 2 1 1 1 ...  
## $ int\_rate : num 6.67 7.46 16.01 8.81 6.07 ...  
## $ installment : num 491.6 93.3 291.9 1109.9 1218.2 ...  
## $ grade : int 1 1 3 1 1 3 3 3 1 1 ...  
## $ sub\_grade : int 2 4 15 5 2 13 13 13 4 4 ...  
## $ home\_ownership : int 2 4 2 4 2 4 4 4 2 4 ...  
## $ annual\_inc : num 75000 50000 42000 122000 89000 38000 60000 50000 247000 58000 ...  
## $ dti : num 13.22 19.37 11.69 9.57 21.1 ...  
## $ delinq\_2yrs : int 1 0 0 0 0 0 0 1 0 0 ...  
## $ fico\_range\_low : int 760 780 715 790 725 670 680 705 690 730 ...  
## $ fico\_range\_high : int 764 784 719 794 729 674 684 709 694 734 ...  
## $ inq\_last\_6mths : int 0 0 0 0 0 1 0 2 3 0 ...  
## $ open\_acc : int 15 12 4 8 13 9 6 4 17 13 ...  
## $ pub\_rec : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ total\_acc : int 26 24 9 14 30 14 15 7 23 17 ...  
## $ last\_fico\_range\_high: int 769 634 709 819 749 684 574 679 659 754 ...  
## $ last\_fico\_range\_low : int 765 630 705 815 745 680 570 675 655 750 ...  
## $ inq\_last\_12m : int 1 0 2 0 1 2 0 3 12 2 ...  
## $ tot\_hi\_cred\_lim : num 310778 114300 22234 57864 181912 ...  
## $ num\_rev\_accts : int 18 14 5 10 17 12 4 3 13 11 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

pr.out <- prcomp (ds\_lc\_pca, scale =TRUE)  
names(pr.out)

## [1] "sdev" "rotation" "center" "scale" "x"

pr.out$center

## loan\_amnt term int\_rate   
## 16348.77500 1.32400 12.80433   
## installment grade sub\_grade   
## 472.80447 2.42600 10.06600   
## home\_ownership annual\_inc dti   
## 2.88800 83692.20576 18.51897   
## delinq\_2yrs fico\_range\_low fico\_range\_high   
## 0.22300 707.30500 711.30600   
## inq\_last\_6mths open\_acc pub\_rec   
## 0.41900 11.42600 0.12300   
## total\_acc last\_fico\_range\_high last\_fico\_range\_low   
## 22.81400 711.36500 704.39500   
## inq\_last\_12m tot\_hi\_cred\_lim num\_rev\_accts   
## 1.88400 192555.72100 13.09400

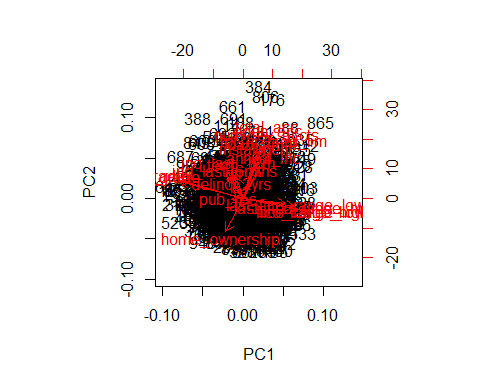
pr.out$scale

## loan\_amnt term int\_rate   
## 1.032236e+04 4.682342e-01 5.230282e+00   
## installment grade sub\_grade   
## 2.901695e+02 1.202485e+00 6.119566e+00   
## home\_ownership annual\_inc dti   
## 9.350563e-01 6.827437e+04 1.266146e+01   
## delinq\_2yrs fico\_range\_low fico\_range\_high   
## 8.049341e-01 3.623154e+01 3.623536e+01   
## inq\_last\_6mths open\_acc pub\_rec   
## 6.942052e-01 5.973961e+00 3.316338e-01   
## total\_acc last\_fico\_range\_high last\_fico\_range\_low   
## 1.220411e+01 5.177898e+01 7.352935e+01   
## inq\_last\_12m tot\_hi\_cred\_lim num\_rev\_accts   
## 2.273700e+00 1.811434e+05 8.384597e+00

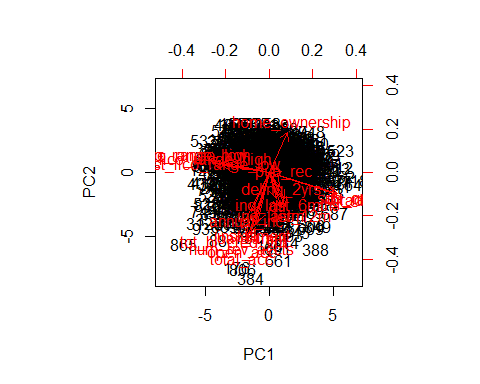
pr.out$rotation

## PC1 PC2 PC3 PC4  
## loan\_amnt 0.08654170 0.304232318 0.41908129 -0.26183120  
## term -0.09500792 0.194799961 0.32128510 0.17254514  
## int\_rate -0.38898465 0.124588135 0.20073268 0.22020725  
## installment 0.06691735 0.285474196 0.37823723 -0.30819008  
## grade -0.38543479 0.127231504 0.19916355 0.21742190  
## sub\_grade -0.39604454 0.118111973 0.19277485 0.21591621  
## home\_ownership -0.10860873 -0.231173221 -0.02343444 0.04352809  
## annual\_inc 0.10313155 0.224598599 0.07214233 -0.38303005  
## dti -0.04240715 0.148783914 0.01781923 0.33996350  
## delinq\_2yrs -0.05839575 0.079620271 -0.06854094 -0.09390503  
## fico\_range\_low 0.34211443 -0.076750177 0.14573708 0.21357637  
## fico\_range\_high 0.34209213 -0.076771434 0.14573968 0.21358804  
## inq\_last\_6mths -0.08934230 0.151679308 -0.25000853 0.03059925  
## open\_acc 0.12418383 0.373608850 -0.20171986 0.19350162  
## pub\_rec -0.07074454 -0.006389139 -0.15428396 -0.07263336  
## total\_acc 0.12210084 0.400577207 -0.22542386 0.18595553  
## last\_fico\_range\_high 0.32115357 -0.058816208 0.19476789 0.28461887  
## last\_fico\_range\_low 0.26492464 -0.037521583 0.19607995 0.30221809  
## inq\_last\_12m -0.06981059 0.201039318 -0.29681462 -0.01674673  
## tot\_hi\_cred\_lim 0.15699636 0.322635373 0.01253402 -0.13517524  
## num\_rev\_accts 0.12005389 0.354574217 -0.25546143 0.17590524  
## PC5 PC6 PC7 PC8  
## loan\_amnt 0.03060060 -0.207147922 -0.18322364 0.02375023  
## term -0.06670474 0.068131855 0.07108629 -0.11911839  
## int\_rate -0.09207974 0.021061227 0.07460643 0.05904363  
## installment 0.04093975 -0.232538007 -0.21121368 0.07504605  
## grade -0.09417080 0.028757822 0.06348067 0.05358915  
## sub\_grade -0.08929097 0.018006485 0.06549040 0.04479305  
## home\_ownership 0.05782884 -0.601808055 -0.11740753 0.28680134  
## annual\_inc -0.18046931 0.041267143 0.11541220 0.12684777  
## dti 0.21907431 0.019915674 -0.17861896 -0.37132693  
## delinq\_2yrs 0.27061952 0.222838710 0.46866992 0.46282534  
## fico\_range\_low -0.26546828 -0.174436580 0.24082196 -0.18810129  
## fico\_range\_high -0.26549618 -0.174437759 0.24084908 -0.18808121  
## inq\_last\_6mths -0.54683877 -0.092461867 -0.08860975 0.21886952  
## open\_acc 0.16964162 -0.216457312 -0.04368282 -0.01135464  
## pub\_rec -0.01858345 0.258779906 -0.59160239 -0.14791401  
## total\_acc 0.16815315 -0.084465839 0.02587632 0.02960310  
## last\_fico\_range\_high -0.05742696 0.194542058 -0.21030030 0.34681429  
## last\_fico\_range\_low -0.01424163 0.276826725 -0.25912070 0.44383892  
## inq\_last\_12m -0.51079055 -0.008659762 -0.09014305 0.07866327  
## tot\_hi\_cred\_lim -0.10316450 0.387390634 0.16664568 -0.19789392  
## num\_rev\_accts 0.19501932 -0.166919531 0.02294196 0.12902217  
## PC9 PC10 PC11 PC12  
## loan\_amnt 0.06900154 0.22316910 8.805669e-02 0.021797385  
## term -0.38747054 0.24616540 4.608218e-01 -0.557247948  
## int\_rate -0.05317931 -0.10163332 -1.672778e-01 0.147419293  
## installment 0.20058568 0.14639497 -9.289577e-02 0.253792590  
## grade -0.06586521 -0.09917381 -1.641214e-01 0.152329499  
## sub\_grade -0.05046429 -0.09901335 -1.488603e-01 0.141333091  
## home\_ownership -0.06555412 -0.11982677 -2.106652e-01 -0.427326634  
## annual\_inc -0.15916285 -0.46758884 -3.407629e-01 -0.367454450  
## dti 0.59750052 0.05620299 -2.319284e-01 -0.384008627  
## delinq\_2yrs 0.11400889 0.53251486 -2.943304e-01 -0.157509233  
## fico\_range\_low -0.10229379 0.16452623 -2.641645e-01 0.073829044  
## fico\_range\_high -0.10232596 0.16450250 -2.642211e-01 0.073843012  
## inq\_last\_6mths 0.19629617 0.21879461 2.028299e-01 0.030216141  
## open\_acc -0.05947387 -0.12715313 -1.078306e-02 0.038946861  
## pub\_rec -0.44414201 0.33789835 -4.479718e-01 -0.029437496  
## total\_acc -0.11700891 -0.05407801 -2.607889e-02 0.043855702  
## last\_fico\_range\_high 0.05483871 -0.07644553 3.302489e-02 -0.030766612  
## last\_fico\_range\_low 0.05364578 -0.14811543 6.472992e-05 -0.009370946  
## inq\_last\_12m 0.19661061 0.09351243 -3.812541e-02 -0.156162292  
## tot\_hi\_cred\_lim 0.06983809 -0.21434244 -6.085676e-02 -0.068858357  
## num\_rev\_accts -0.26860267 0.02704786 6.930954e-02 0.153260331  
## PC13 PC14 PC15 PC16  
## loan\_amnt 0.0452915932 0.037010721 -0.016552613 -0.010655211  
## term 0.0212101513 -0.012009770 0.047674351 -0.025182086  
## int\_rate 0.0048855440 0.022744971 -0.002888111 0.006541779  
## installment 0.0397590105 0.030124197 -0.039637980 0.005488311  
## grade 0.0037612912 0.018724242 0.001293463 0.020719253  
## sub\_grade 0.0014732185 0.004366827 -0.007898434 0.012855503  
## home\_ownership -0.0833928533 0.411583707 -0.212063367 0.057824363  
## annual\_inc -0.0923617347 -0.457573620 0.095761048 0.002027730  
## dti -0.0879802675 -0.231997234 -0.067189703 0.118922311  
## delinq\_2yrs -0.0204009784 0.041093619 0.095706088 -0.013266251  
## fico\_range\_low -0.0127184878 -0.014545437 0.002630315 -0.004565707  
## fico\_range\_high -0.0127106144 -0.014535668 0.002650002 -0.004524185  
## inq\_last\_6mths -0.6290940335 -0.116800081 -0.060337282 -0.020812804  
## open\_acc -0.1302435372 0.233597008 0.769121519 -0.061107062  
## pub\_rec -0.1006536334 0.049891948 0.016414637 -0.006787907  
## total\_acc 0.0643308385 -0.047440735 -0.432564323 -0.695926689  
## last\_fico\_range\_high 0.0614163152 0.018938247 -0.007526681 0.084035569  
## last\_fico\_range\_low -0.0002958659 0.004444301 0.022955514 -0.066089866  
## inq\_last\_12m 0.7052220989 0.109985251 0.074406118 0.031049459  
## tot\_hi\_cred\_lim -0.1956888242 0.653978710 -0.249823155 0.192705946  
## num\_rev\_accts 0.0742701979 -0.221064499 -0.277270946 0.665447589  
## PC17 PC18 PC19 PC20  
## loan\_amnt 0.029784072 -0.0821039210 0.6980935388 -0.1273606906  
## term 0.025064382 0.0259618218 -0.2280681139 0.0355726572  
## int\_rate -0.026664681 0.4200611866 -0.0397838466 -0.6958449372  
## installment -0.002109479 0.0777438582 -0.6406932977 0.1210640191  
## grade -0.056990010 -0.8123446805 -0.0586760067 -0.0035911953  
## sub\_grade -0.003651320 0.3856632907 0.2107836265 0.6949179471  
## home\_ownership 0.014786331 -0.0045560352 -0.0129048306 0.0065526163  
## annual\_inc -0.033337929 0.0024789676 0.0027221708 0.0015904257  
## dti 0.014733112 -0.0085508708 -0.0013394687 -0.0079016303  
## delinq\_2yrs -0.020563092 -0.0006559535 0.0042901620 0.0095853877  
## fico\_range\_low 0.052100950 0.0007736752 0.0019306723 0.0048858724  
## fico\_range\_high 0.052108440 0.0005652350 0.0021452092 0.0050703947  
## inq\_last\_6mths -0.016384249 -0.0024139182 0.0012347189 -0.0050981139  
## open\_acc -0.030125511 0.0070479472 -0.0035998382 0.0084004158  
## pub\_rec -0.024536796 0.0123182674 0.0043005966 0.0001886952  
## total\_acc -0.072157729 -0.0038423421 0.0004396965 0.0047385206  
## last\_fico\_range\_high -0.738455974 0.0280363173 0.0183872317 0.0151240881  
## last\_fico\_range\_low 0.653794607 -0.0168674737 -0.0098163767 -0.0021834834  
## inq\_last\_12m 0.042846992 -0.0005481756 -0.0039756507 0.0026181664  
## tot\_hi\_cred\_lim 0.016371132 0.0059742731 -0.0001877358 0.0091459937  
## num\_rev\_accts 0.072491427 0.0044109851 0.0045979349 -0.0063951984  
## PC21  
## loan\_amnt -8.194941e-05  
## term 5.141501e-05  
## int\_rate 1.369712e-04  
## installment 1.044245e-04  
## grade -1.322016e-04  
## sub\_grade -8.649090e-05  
## home\_ownership -1.515891e-05  
## annual\_inc -1.876081e-05  
## dti 2.025323e-05  
## delinq\_2yrs -1.054338e-05  
## fico\_range\_low -7.071236e-01  
## fico\_range\_high 7.070899e-01  
## inq\_last\_6mths 1.071879e-05  
## open\_acc -2.943608e-06  
## pub\_rec -7.632812e-06  
## total\_acc 2.966452e-05  
## last\_fico\_range\_high 1.646008e-06  
## last\_fico\_range\_low -4.375271e-06  
## inq\_last\_12m -7.488382e-06  
## tot\_hi\_cred\_lim -3.419210e-06  
## num\_rev\_accts -1.088670e-05

biplot(pr.out)



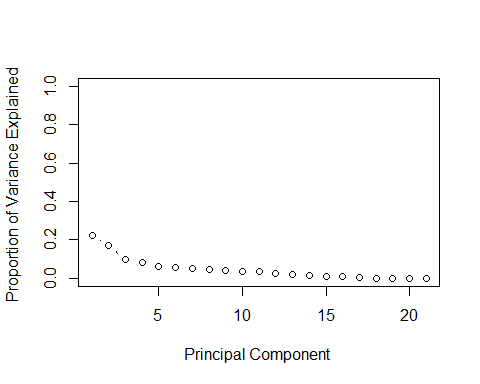
pr.out$rotation=-pr.out$rotation  
pr.out$x=-pr.out$x  
biplot (pr.out , scale =0)



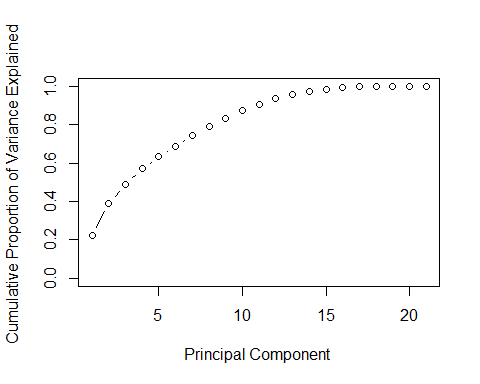
#pr.out$sdev  
pr.var =pr.out$sdev ^2  
pve=pr.var/sum(pr.var )  
pve

## [1] 2.211029e-01 1.694966e-01 1.006923e-01 8.079903e-02 6.126954e-02  
## [6] 5.587240e-02 5.302054e-02 4.845791e-02 4.356585e-02 3.840623e-02  
## [11] 3.502943e-02 2.743101e-02 2.226131e-02 1.522615e-02 1.162247e-02  
## [16] 8.035864e-03 5.143899e-03 1.663147e-03 6.010416e-04 3.024257e-04  
## [21] 1.756110e-08

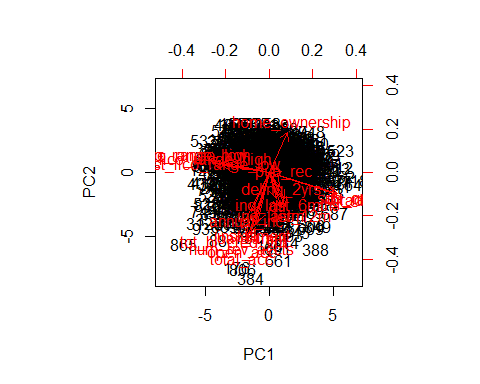
plot(pve , xlab="Principal Component ", ylab="Proportion of Variance Explained ", ylim=c(0,1) ,type="b")



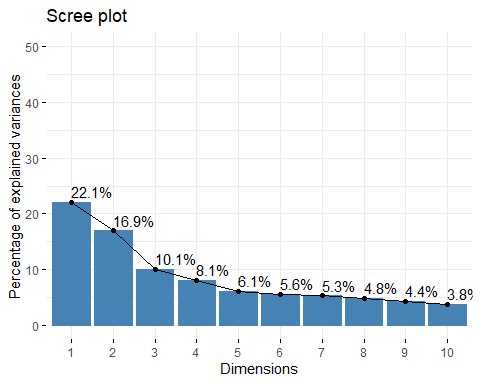
plot(cumsum (pve ), xlab="Principal Component ", ylab ="Cumulative Proportion of Variance Explained ", ylim=c(0,1) , type="b")



#autoplot(pr.out)  
biplot (pr.out , scale =0)



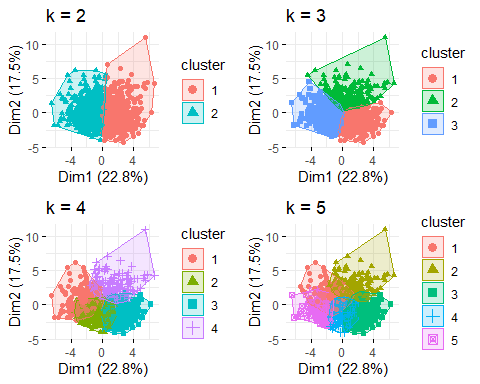
fviz\_eig(pr.out, addlabels = TRUE, ylim = c(0, 50))



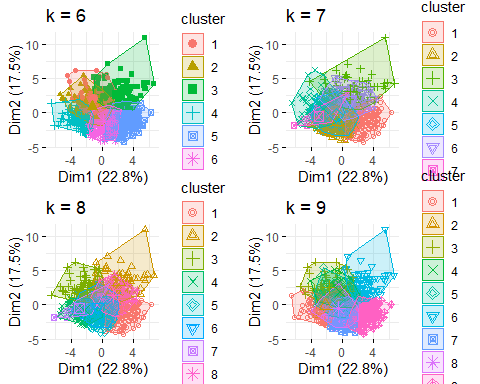
PCA does not bring in much value. Very dimension were identified with PCA.

ds\_lc\_kmean <- ds\_lc[,c(  
 "loan\_amnt",  
 "term",  
 "int\_rate",  
 "installment",  
 "grade",  
 "sub\_grade",  
 "home\_ownership",  
 "annual\_inc",  
 "dti",  
 "delinq\_2yrs",  
 "fico\_range\_low",  
 "fico\_range\_high",  
 "inq\_last\_6mths",  
 "open\_acc",  
 "pub\_rec",  
 "total\_acc",  
 "last\_fico\_range\_high",   
 "last\_fico\_range\_low",   
 "inq\_last\_12m",   
 "tot\_hi\_cred\_lim",   
 "num\_rev\_accts"   
 )]  
ds\_lc\_kmean\_subset <- sample(nrow(ds\_lc\_kmean), 1000)  
ds\_lc\_kmean = ds\_lc\_kmean[ds\_lc\_kmean\_subset, ]  
ds\_lc\_kmean\_scaled <- as.data.frame(scale(ds\_lc\_kmean))

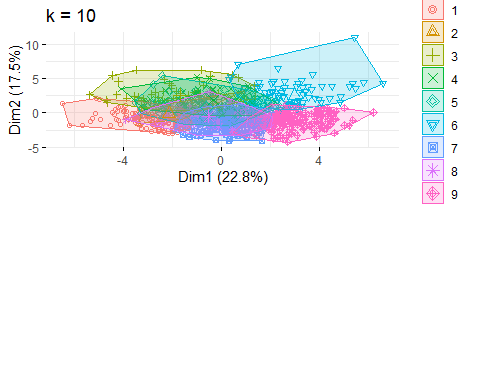
k2 <- kmeans(ds\_lc\_kmean\_scaled, centers = 2, nstart = 25)  
k3 <- kmeans(ds\_lc\_kmean\_scaled, centers = 3, nstart = 25)  
k4 <- kmeans(ds\_lc\_kmean\_scaled, centers = 4, nstart = 25)  
k5 <- kmeans(ds\_lc\_kmean\_scaled, centers = 5, nstart = 25)  
k6 <- kmeans(ds\_lc\_kmean\_scaled, centers = 6, nstart = 25)  
k7 <- kmeans(ds\_lc\_kmean\_scaled, centers = 7, nstart = 25)  
k8 <- kmeans(ds\_lc\_kmean\_scaled, centers = 8, nstart = 25)  
k9 <- kmeans(ds\_lc\_kmean\_scaled, centers = 9, nstart = 25)  
  
  
# plots to compare  
p2 <- fviz\_cluster(k2, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 2")  
p3 <- fviz\_cluster(k3, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 3")  
p4 <- fviz\_cluster(k4, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 4")  
p5 <- fviz\_cluster(k5, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 5")  
p6 <- fviz\_cluster(k6, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 6")  
p7 <- fviz\_cluster(k7, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 7")  
p8 <- fviz\_cluster(k8, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 8")  
p9 <- fviz\_cluster(k9, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 9")  
p10 <- fviz\_cluster(k9, geom = "point", data = ds\_lc\_kmean\_scaled, ggtheme = theme\_minimal()) + ggtitle("k = 10")  
  
  
grid.arrange(p2, p3, p4, p5, nrow = 2)



grid.arrange(p6, p7, p8, p9, nrow = 2)



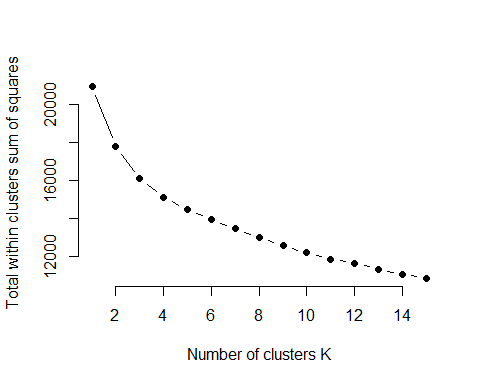
grid.arrange(p10, nrow = 2)



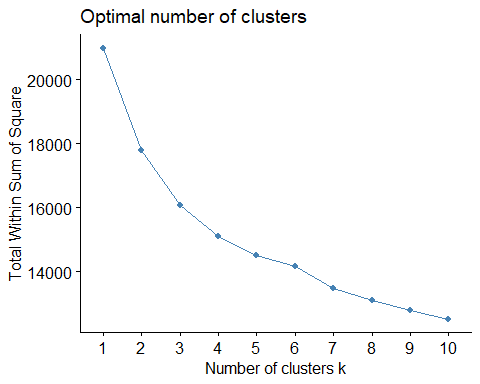
set.seed(1234)  
#Function to compute total within cluster sum of squares   
wss <- function(k) {  
 kmeans(ds\_lc\_kmean\_scaled, k, nstart = 25, iter.max = 10)$tot.withinss  
}  
  
#Compute and plot the within sum of squares (wss) for k = 1 to k = 10  
k.values <- 1:15  
  
#Extract wss for 2 - 10 clusters  
wss\_values <- map\_dbl(k.values, wss)

## Warning: did not converge in 10 iterations  
  
## Warning: did not converge in 10 iterations

plot(k.values, wss\_values,  
 type = "b", pch = 19, frame = FALSE,  
 xlab = "Number of clusters K",  
 ylab = "Total within clusters sum of squares")



set.seed(1234)  
fviz\_nbclust(ds\_lc\_kmean\_scaled, kmeans, method = "wss")



Based on the visualization 3 or 4 clusters seems to be better classification. 3 clusters have very less overlap compared to that of the 4 clusters.

ds\_lc\_kmean$cluster <- k4$cluster

ds\_lc\_kmean$category <- ""  
for(i in 1:nrow(ds\_lc\_kmean)){  
 if (ds\_lc\_kmean$cluster[i] == 1){  
 ds\_lc\_kmean$category[i] <- "Low"  
 }  
 else if (ds\_lc\_kmean$cluster[i] == 2){  
 ds\_lc\_kmean$category[i] <- "Low-Medium"  
 }  
 else if (ds\_lc\_kmean$cluster[i] == 3){  
 ds\_lc\_kmean$category[i] <- "Medium"  
 }  
 else if (ds\_lc\_kmean$cluster[i] == 4){  
 ds\_lc\_kmean$category[i] <- "High"  
 }  
}  
  
low\_risk\_loans <- ds\_lc\_kmean %>%  
 filter(category == "Low")  
  
low\_medium\_risk\_loans <- ds\_lc\_kmean %>%  
 filter(category == "Low-Medium")  
  
medium\_risk\_loans <- ds\_lc\_kmean %>%  
 filter(category == "Medium")  
  
high\_risk\_loans <- ds\_lc\_kmean %>%  
 filter(category == "High")

mean(low\_risk\_loans$int\_rate)

## [1] 18.6983

mean(low\_medium\_risk\_loans$int\_rate)

## [1] 11.06391

mean(medium\_risk\_loans$int\_rate)

## [1] 8.648333

mean(high\_risk\_loans$int\_rate)

## [1] 11.21703

mean(low\_risk\_loans$annual\_inc)

## [1] 65542.35

mean(low\_medium\_risk\_loans$annual\_inc)

## [1] 60948.37

mean(medium\_risk\_loans$annual\_inc)

## [1] 69494.57

mean(high\_risk\_loans$annual\_inc)

## [1] 138060.6

mean(low\_risk\_loans$dti)

## [1] 21.48065

mean(low\_medium\_risk\_loans$dti)

## [1] 16.71315

mean(medium\_risk\_loans$dti)

## [1] 19.07592

mean(high\_risk\_loans$dti)

## [1] 19.55137

mean(low\_risk\_loans$fico\_range\_low)

## [1] 686.2551

mean(low\_medium\_risk\_loans$fico\_range\_low)

## [1] 690.1458

mean(medium\_risk\_loans$fico\_range\_low)

## [1] 755.5263

mean(high\_risk\_loans$fico\_range\_low)

## [1] 706.4286

mean(low\_risk\_loans$fico\_range\_high)

## [1] 690.2551

mean(low\_medium\_risk\_loans$fico\_range\_high)

## [1] 694.1458

mean(medium\_risk\_loans$fico\_range\_high)

## [1] 759.5263

mean(high\_risk\_loans$fico\_range\_high)

## [1] 710.4286

mean(low\_risk\_loans$inq\_last\_6mths)

## [1] 0.6477733

mean(low\_medium\_risk\_loans$inq\_last\_6mths)

## [1] 0.361516

mean(medium\_risk\_loans$inq\_last\_6mths)

## [1] 0.2061404

mean(high\_risk\_loans$inq\_last\_6mths)

## [1] 0.6593407

mean(low\_risk\_loans$delinq\_2yrs)

## [1] 0.2186235

mean(low\_medium\_risk\_loans$delinq\_2yrs)

## [1] 0.2798834

mean(medium\_risk\_loans$delinq\_2yrs)

## [1] 0.03070175

mean(high\_risk\_loans$delinq\_2yrs)

## [1] 0.2527473

mean(low\_risk\_loans$grade)

## [1] 3.740891

mean(low\_medium\_risk\_loans$grade)

## [1] 2.043732

mean(medium\_risk\_loans$grade)

## [1] 1.464912

mean(high\_risk\_loans$grade)

## [1] 2.082418

mean(low\_risk\_loans$sub\_grade)

## [1] 16.90283

mean(low\_medium\_risk\_loans$sub\_grade)

## [1] 8.22449

mean(medium\_risk\_loans$sub\_grade)

## [1] 4.995614

mean(high\_risk\_loans$sub\_grade)

## [1] 8.291209