Homework 4

Alain Kuiete

11/5/2020

## DATA EXPLORATION

auto.insurance <- read.csv('insurance\_training\_data.csv')  
head(auto.insurance)

## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1  
## 1 1 0 0 0 60 0 11 $67,349 No  
## 2 2 0 0 0 43 0 11 $91,449 No  
## 3 4 0 0 0 35 1 10 $16,039 No  
## 4 5 0 0 0 51 0 14 No  
## 5 6 0 0 0 50 0 NA $114,986 No  
## 6 7 1 2946 0 34 1 12 $125,301 Yes  
## HOME\_VAL MSTATUS SEX EDUCATION JOB TRAVTIME CAR\_USE BLUEBOOK  
## 1 $0 z\_No M PhD Professional 14 Private $14,230  
## 2 $257,252 z\_No M z\_High School z\_Blue Collar 22 Commercial $14,940  
## 3 $124,191 Yes z\_F z\_High School Clerical 5 Private $4,010  
## 4 $306,251 Yes M <High School z\_Blue Collar 32 Private $15,440  
## 5 $243,925 Yes z\_F PhD Doctor 36 Private $18,000  
## 6 $0 z\_No z\_F Bachelors z\_Blue Collar 46 Commercial $17,430  
## TIF CAR\_TYPE RED\_CAR OLDCLAIM CLM\_FREQ REVOKED MVR\_PTS CAR\_AGE  
## 1 11 Minivan yes $4,461 2 No 3 18  
## 2 1 Minivan yes $0 0 No 0 1  
## 3 4 z\_SUV no $38,690 2 No 3 10  
## 4 7 Minivan yes $0 0 No 0 6  
## 5 1 z\_SUV no $19,217 2 Yes 3 17  
## 6 1 Sports Car no $0 0 No 0 7  
## URBANICITY  
## 1 Highly Urban/ Urban  
## 2 Highly Urban/ Urban  
## 3 Highly Urban/ Urban  
## 4 Highly Urban/ Urban  
## 5 Highly Urban/ Urban  
## 6 Highly Urban/ Urban

write.csv(head(auto.insurance), "head\_insurance\_data.csv")

cost.ins <- auto.insurance  
#attach(cost.ins)

#### Remove $ sign on INCOME, HOME\_VAL, BLUEBOOK, and OLDCLAIM

dollar.sign <- function(a){  
 a <- as.numeric(gsub("[\\$,]", "", a))  
 return(a)  
}  
cost.ins$INCOME <- dollar.sign(cost.ins$INCOME)  
cost.ins$HOME\_VAL <- dollar.sign(cost.ins$HOME\_VAL)  
cost.ins$BLUEBOOK <- dollar.sign(cost.ins$BLUEBOOK)  
cost.ins$OLDCLAIM <- dollar.sign(cost.ins$OLDCLAIM)  
summary(cost.ins)

## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV   
## Min. : 1 Min. :0.0000 Min. : 0 Min. :0.0000   
## 1st Qu.: 2559 1st Qu.:0.0000 1st Qu.: 0 1st Qu.:0.0000   
## Median : 5133 Median :0.0000 Median : 0 Median :0.0000   
## Mean : 5152 Mean :0.2638 Mean : 1504 Mean :0.1711   
## 3rd Qu.: 7745 3rd Qu.:1.0000 3rd Qu.: 1036 3rd Qu.:0.0000   
## Max. :10302 Max. :1.0000 Max. :107586 Max. :4.0000   
##   
## AGE HOMEKIDS YOJ INCOME PARENT1   
## Min. :16.00 Min. :0.0000 Min. : 0.0 Min. : 0 No :7084   
## 1st Qu.:39.00 1st Qu.:0.0000 1st Qu.: 9.0 1st Qu.: 28097 Yes:1077   
## Median :45.00 Median :0.0000 Median :11.0 Median : 54028   
## Mean :44.79 Mean :0.7212 Mean :10.5 Mean : 61898   
## 3rd Qu.:51.00 3rd Qu.:1.0000 3rd Qu.:13.0 3rd Qu.: 85986   
## Max. :81.00 Max. :5.0000 Max. :23.0 Max. :367030   
## NA's :6 NA's :454 NA's :445   
## HOME\_VAL MSTATUS SEX EDUCATION   
## Min. : 0 Yes :4894 M :3786 <High School :1203   
## 1st Qu.: 0 z\_No:3267 z\_F:4375 Bachelors :2242   
## Median :161160 Masters :1658   
## Mean :154867 PhD : 728   
## 3rd Qu.:238724 z\_High School:2330   
## Max. :885282   
## NA's :464   
## JOB TRAVTIME CAR\_USE BLUEBOOK   
## z\_Blue Collar:1825 Min. : 5.00 Commercial:3029 Min. : 1500   
## Clerical :1271 1st Qu.: 22.00 Private :5132 1st Qu.: 9280   
## Professional :1117 Median : 33.00 Median :14440   
## Manager : 988 Mean : 33.49 Mean :15710   
## Lawyer : 835 3rd Qu.: 44.00 3rd Qu.:20850   
## Student : 712 Max. :142.00 Max. :69740   
## (Other) :1413   
## TIF CAR\_TYPE RED\_CAR OLDCLAIM   
## Min. : 1.000 Minivan :2145 no :5783 Min. : 0   
## 1st Qu.: 1.000 Panel Truck: 676 yes:2378 1st Qu.: 0   
## Median : 4.000 Pickup :1389 Median : 0   
## Mean : 5.351 Sports Car : 907 Mean : 4037   
## 3rd Qu.: 7.000 Van : 750 3rd Qu.: 4636   
## Max. :25.000 z\_SUV :2294 Max. :57037   
##   
## CLM\_FREQ REVOKED MVR\_PTS CAR\_AGE   
## Min. :0.0000 No :7161 Min. : 0.000 Min. :-3.000   
## 1st Qu.:0.0000 Yes:1000 1st Qu.: 0.000 1st Qu.: 1.000   
## Median :0.0000 Median : 1.000 Median : 8.000   
## Mean :0.7986 Mean : 1.696 Mean : 8.328   
## 3rd Qu.:2.0000 3rd Qu.: 3.000 3rd Qu.:12.000   
## Max. :5.0000 Max. :13.000 Max. :28.000   
## NA's :510   
## URBANICITY   
## Highly Urban/ Urban :6492   
## z\_Highly Rural/ Rural:1669   
##   
##   
##   
##   
##

write.csv(summary(cost.ins), "summary\_data.csv")

#### Transforming to categorical variables

cost.ins$TARGET\_FLAG <- factor(cost.ins$TARGET\_FLAG, labels = c("NO","YES"))  
cost.ins$HOMEKIDS <- factor(cost.ins$HOMEKIDS, labels = c("HK1", "HK2", "HK3", "HK4", "HK5", "HK6"))  
cost.ins$CLM\_FREQ <- factor(cost.ins$CLM\_FREQ, labels = c("CF1", "CF2", "CF3", "CF4", "CF5", "CF6"))  
#cost.ins$KIDSDRIV <- factor(cost.ins$KIDSDRIV)  
cost.ins$URBANICITY <- ifelse(cost.ins$URBANICITY=="Highly Urban/ Urban", "U","R")  
cost.ins$MSTATUS <- ifelse(cost.ins$MSTATUS=="Yes", "1","0")  
cost.ins$SEX <- ifelse(cost.ins$SEX=="M", "1","0")  
cost.ins$CAR\_USE <- ifelse(cost.ins$CAR\_USE=="Private", "1","0")  
cost.ins$EDUCATION <- ifelse(cost.ins$EDUCATION=="<High School","1HS",  
 ifelse(cost.ins$EDUCATION=="z\_High School","2HS",  
 ifelse(cost.ins$EDUCATION=="Bachelors","3BA",  
 ifelse(cost.ins$EDUCATION=="Masters","4MA","5PH"))))  
#cost.ins$JOB[cost.ins$JOB==""] <- "UNEMPLOY"  
summary(cost.ins)

## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV AGE   
## Min. : 1 NO :6008 Min. : 0 Min. :0.0000 Min. :16.00   
## 1st Qu.: 2559 YES:2153 1st Qu.: 0 1st Qu.:0.0000 1st Qu.:39.00   
## Median : 5133 Median : 0 Median :0.0000 Median :45.00   
## Mean : 5152 Mean : 1504 Mean :0.1711 Mean :44.79   
## 3rd Qu.: 7745 3rd Qu.: 1036 3rd Qu.:0.0000 3rd Qu.:51.00   
## Max. :10302 Max. :107586 Max. :4.0000 Max. :81.00   
## NA's :6   
## HOMEKIDS YOJ INCOME PARENT1 HOME\_VAL   
## HK1:5289 Min. : 0.0 Min. : 0 No :7084 Min. : 0   
## HK2: 902 1st Qu.: 9.0 1st Qu.: 28097 Yes:1077 1st Qu.: 0   
## HK3:1118 Median :11.0 Median : 54028 Median :161160   
## HK4: 674 Mean :10.5 Mean : 61898 Mean :154867   
## HK5: 164 3rd Qu.:13.0 3rd Qu.: 85986 3rd Qu.:238724   
## HK6: 14 Max. :23.0 Max. :367030 Max. :885282   
## NA's :454 NA's :445 NA's :464   
## MSTATUS SEX EDUCATION JOB   
## Length:8161 Length:8161 Length:8161 z\_Blue Collar:1825   
## Class :character Class :character Class :character Clerical :1271   
## Mode :character Mode :character Mode :character Professional :1117   
## Manager : 988   
## Lawyer : 835   
## Student : 712   
## (Other) :1413   
## TRAVTIME CAR\_USE BLUEBOOK TIF   
## Min. : 5.00 Length:8161 Min. : 1500 Min. : 1.000   
## 1st Qu.: 22.00 Class :character 1st Qu.: 9280 1st Qu.: 1.000   
## Median : 33.00 Mode :character Median :14440 Median : 4.000   
## Mean : 33.49 Mean :15710 Mean : 5.351   
## 3rd Qu.: 44.00 3rd Qu.:20850 3rd Qu.: 7.000   
## Max. :142.00 Max. :69740 Max. :25.000   
##   
## CAR\_TYPE RED\_CAR OLDCLAIM CLM\_FREQ REVOKED   
## Minivan :2145 no :5783 Min. : 0 CF1:5009 No :7161   
## Panel Truck: 676 yes:2378 1st Qu.: 0 CF2: 997 Yes:1000   
## Pickup :1389 Median : 0 CF3:1171   
## Sports Car : 907 Mean : 4037 CF4: 776   
## Van : 750 3rd Qu.: 4636 CF5: 190   
## z\_SUV :2294 Max. :57037 CF6: 18   
##   
## MVR\_PTS CAR\_AGE URBANICITY   
## Min. : 0.000 Min. :-3.000 Length:8161   
## 1st Qu.: 0.000 1st Qu.: 1.000 Class :character   
## Median : 1.000 Median : 8.000 Mode :character   
## Mean : 1.696 Mean : 8.328   
## 3rd Qu.: 3.000 3rd Qu.:12.000   
## Max. :13.000 Max. :28.000   
## NA's :510

head(cost.ins)

## INDEX TARGET\_FLAG TARGET\_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME PARENT1  
## 1 1 NO 0 0 60 HK1 11 67349 No  
## 2 2 NO 0 0 43 HK1 11 91449 No  
## 3 4 NO 0 0 35 HK2 10 16039 No  
## 4 5 NO 0 0 51 HK1 14 NA No  
## 5 6 NO 0 0 50 HK1 NA 114986 No  
## 6 7 YES 2946 0 34 HK2 12 125301 Yes  
## HOME\_VAL MSTATUS SEX EDUCATION JOB TRAVTIME CAR\_USE BLUEBOOK TIF  
## 1 0 0 1 5PH Professional 14 1 14230 11  
## 2 257252 0 1 2HS z\_Blue Collar 22 0 14940 1  
## 3 124191 1 0 2HS Clerical 5 1 4010 4  
## 4 306251 1 1 1HS z\_Blue Collar 32 1 15440 7  
## 5 243925 1 0 5PH Doctor 36 1 18000 1  
## 6 0 0 0 3BA z\_Blue Collar 46 0 17430 1  
## CAR\_TYPE RED\_CAR OLDCLAIM CLM\_FREQ REVOKED MVR\_PTS CAR\_AGE URBANICITY  
## 1 Minivan yes 4461 CF3 No 3 18 U  
## 2 Minivan yes 0 CF1 No 0 1 U  
## 3 z\_SUV no 38690 CF3 No 3 10 U  
## 4 Minivan yes 0 CF1 No 0 6 U  
## 5 z\_SUV no 19217 CF3 Yes 3 17 U  
## 6 Sports Car no 0 CF1 No 0 7 U

attach(cost.ins)  
str(cost.ins)

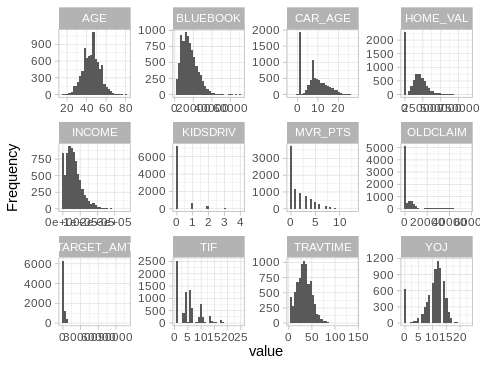
## 'data.frame': 8161 obs. of 26 variables:  
## $ INDEX : int 1 2 4 5 6 7 8 11 12 13 ...  
## $ TARGET\_FLAG: Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 2 2 1 ...  
## $ TARGET\_AMT : num 0 0 0 0 0 ...  
## $ KIDSDRIV : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ AGE : int 60 43 35 51 50 34 54 37 34 50 ...  
## $ HOMEKIDS : Factor w/ 6 levels "HK1","HK2","HK3",..: 1 1 2 1 1 2 1 3 1 1 ...  
## $ YOJ : int 11 11 10 14 NA 12 NA NA 10 7 ...  
## $ INCOME : num 67349 91449 16039 NA 114986 ...  
## $ PARENT1 : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...  
## $ HOME\_VAL : num 0 257252 124191 306251 243925 ...  
## $ MSTATUS : chr "0" "0" "1" "1" ...  
## $ SEX : chr "1" "1" "0" "1" ...  
## $ EDUCATION : chr "5PH" "2HS" "2HS" "1HS" ...  
## $ JOB : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...  
## $ TRAVTIME : int 14 22 5 32 36 46 33 44 34 48 ...  
## $ CAR\_USE : chr "1" "0" "1" "1" ...  
## $ BLUEBOOK : num 14230 14940 4010 15440 18000 ...  
## $ TIF : int 11 1 4 7 1 1 1 1 1 7 ...  
## $ CAR\_TYPE : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...  
## $ RED\_CAR : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...  
## $ OLDCLAIM : num 4461 0 38690 0 19217 ...  
## $ CLM\_FREQ : Factor w/ 6 levels "CF1","CF2","CF3",..: 3 1 3 1 3 1 1 2 1 1 ...  
## $ REVOKED : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...  
## $ MVR\_PTS : int 3 0 3 0 3 0 0 10 0 1 ...  
## $ CAR\_AGE : int 18 1 10 6 17 7 1 7 1 17 ...  
## $ URBANICITY : chr "U" "U" "U" "U" ...

### Univariate Study

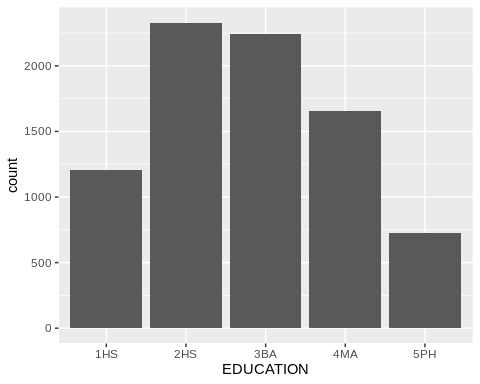
#### Histogram of some variables

We divide the variable in two groups to study

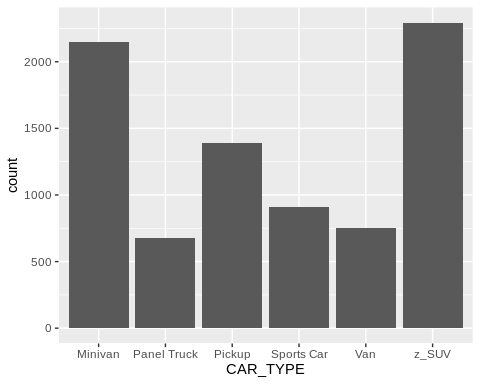
library(DataExplorer)  
var.group1 <-data.frame(TARGET\_AMT, KIDSDRIV, AGE, YOJ, INCOME, HOME\_VAL, TRAVTIME, TIF,   
 OLDCLAIM, BLUEBOOK, MVR\_PTS, CAR\_AGE)  
plot\_histogram(var.group1, ggtheme = theme\_light())



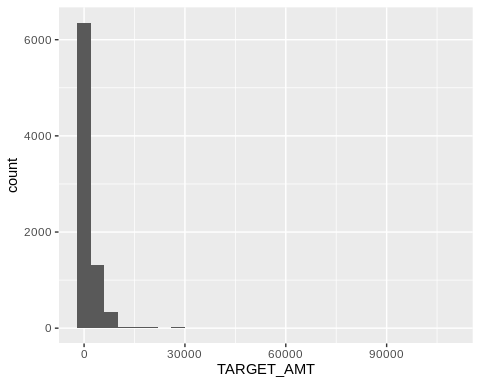
ggplot(cost.ins, aes(x=EDUCATION)) + geom\_bar()



ggplot(cost.ins, aes(CAR\_TYPE)) + geom\_bar()



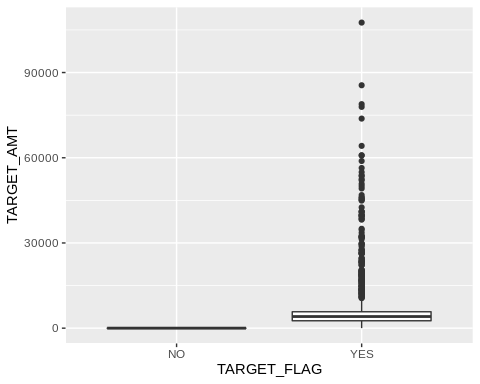
ggplot(cost.ins, aes(x=TARGET\_AMT)) + geom\_histogram(binwidth = 4000)



str(cost.ins)

## 'data.frame': 8161 obs. of 26 variables:  
## $ INDEX : int 1 2 4 5 6 7 8 11 12 13 ...  
## $ TARGET\_FLAG: Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 2 2 1 ...  
## $ TARGET\_AMT : num 0 0 0 0 0 ...  
## $ KIDSDRIV : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ AGE : int 60 43 35 51 50 34 54 37 34 50 ...  
## $ HOMEKIDS : Factor w/ 6 levels "HK1","HK2","HK3",..: 1 1 2 1 1 2 1 3 1 1 ...  
## $ YOJ : int 11 11 10 14 NA 12 NA NA 10 7 ...  
## $ INCOME : num 67349 91449 16039 NA 114986 ...  
## $ PARENT1 : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...  
## $ HOME\_VAL : num 0 257252 124191 306251 243925 ...  
## $ MSTATUS : chr "0" "0" "1" "1" ...  
## $ SEX : chr "1" "1" "0" "1" ...  
## $ EDUCATION : chr "5PH" "2HS" "2HS" "1HS" ...  
## $ JOB : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...  
## $ TRAVTIME : int 14 22 5 32 36 46 33 44 34 48 ...  
## $ CAR\_USE : chr "1" "0" "1" "1" ...  
## $ BLUEBOOK : num 14230 14940 4010 15440 18000 ...  
## $ TIF : int 11 1 4 7 1 1 1 1 1 7 ...  
## $ CAR\_TYPE : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...  
## $ RED\_CAR : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...  
## $ OLDCLAIM : num 4461 0 38690 0 19217 ...  
## $ CLM\_FREQ : Factor w/ 6 levels "CF1","CF2","CF3",..: 3 1 3 1 3 1 1 2 1 1 ...  
## $ REVOKED : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...  
## $ MVR\_PTS : int 3 0 3 0 3 0 0 10 0 1 ...  
## $ CAR\_AGE : int 18 1 10 6 17 7 1 7 1 17 ...  
## $ URBANICITY : chr "U" "U" "U" "U" ...

ggplot(cost.ins, aes(x=TARGET\_FLAG, y=TARGET\_AMT)) + geom\_boxplot()



### Bivariate Study

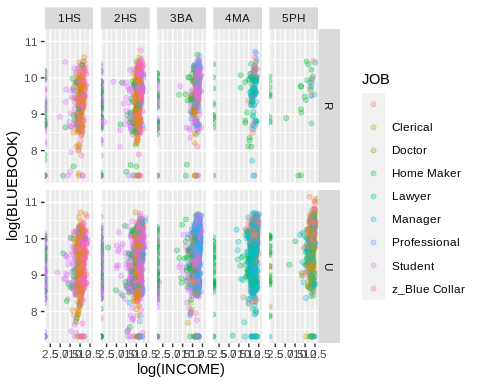
## DATA PREPARETION

#### Looking for correlations between variables

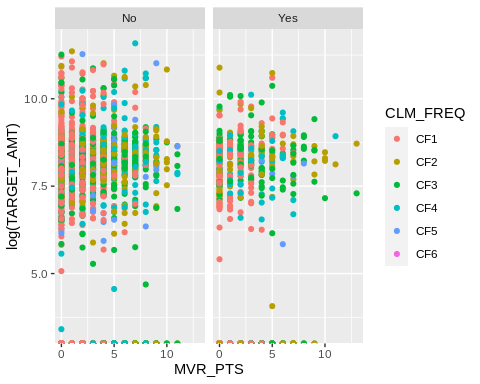
income.group <- data.frame(INCOME, EDUCATION, JOB, HOME\_VAL, CAR\_TYPE, BLUEBOOK, HOMEKIDS, RED\_CAR, URBANICITY, TRAVTIME, REVOKED, CAR\_USE, CLM\_FREQ, OLDCLAIM, MVR\_PTS)

ggplot(income.group, aes(x=log(INCOME), y=log(BLUEBOOK), color = JOB)) + geom\_point(position = position\_jitter(), alpha=.3) + facet\_grid(rows = vars(URBANICITY), cols = vars(EDUCATION))

## Warning: Removed 445 rows containing missing values (geom\_point).

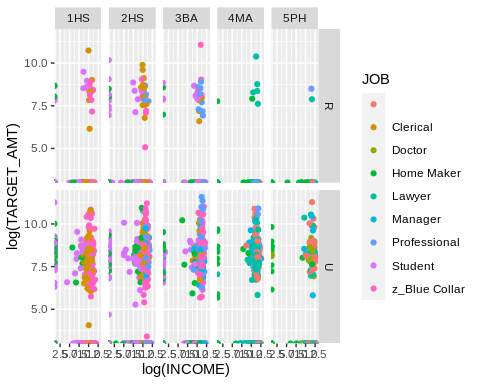
 There exist relative correlation between some categorical variables. We see that in urban like in rural area, blue collar are for les than high school or more, people with bachelor and master have professional, manager and some home maker jobs, With PhD, someone is likely to be Home Maker, Doctor or Clerical

ggplot(income.group, aes(x=MVR\_PTS, y=log(TARGET\_AMT), color = CLM\_FREQ)) + geom\_point() + facet\_grid(cols = vars(REVOKED))



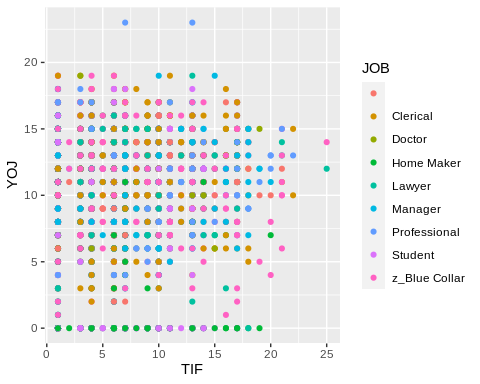
ggplot(income.group, aes(x=log(INCOME), y=log(TARGET\_AMT), color = JOB)) + geom\_point() + facet\_grid(rows = vars(URBANICITY), cols = vars(EDUCATION))

## Warning: Removed 445 rows containing missing values (geom\_point).



ggplot(cost.ins, aes(x=TIF, y=YOJ, color = JOB)) + geom\_point()

## Warning: Removed 454 rows containing missing values (geom\_point).



#### Missing Values in AGE , INCOME, HOME\_VAL, YOJ, and CAR\_AGE

The mice package in R, helps you imputing missing values with plausible data values. These plausible values are drawn from a distribution specifically designed for each missing datapoint.

Here we going to form a data frame with only varibles that have missing values

ins.missing <- data.frame(AGE, INCOME, HOME\_VAL, CAR\_AGE, YOJ)

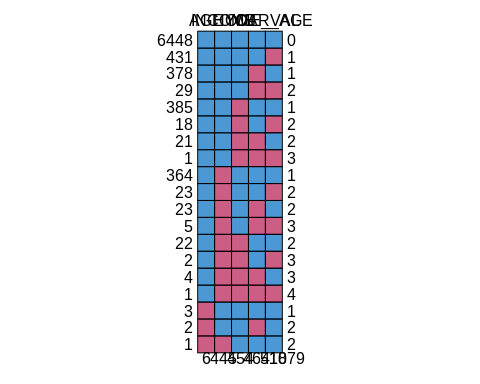
The mice package provides a nice function md.pattern() to get a better understanding of the pattern of missing data

library(mice)

##   
## Attaching package: 'mice'

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

md.pattern(ins.missing)



## AGE INCOME YOJ HOME\_VAL CAR\_AGE   
## 6448 1 1 1 1 1 0  
## 431 1 1 1 1 0 1  
## 378 1 1 1 0 1 1  
## 29 1 1 1 0 0 2  
## 385 1 1 0 1 1 1  
## 18 1 1 0 1 0 2  
## 21 1 1 0 0 1 2  
## 1 1 1 0 0 0 3  
## 364 1 0 1 1 1 1  
## 23 1 0 1 1 0 2  
## 23 1 0 1 0 1 2  
## 5 1 0 1 0 0 3  
## 22 1 0 0 1 1 2  
## 2 1 0 0 1 0 3  
## 4 1 0 0 0 1 3  
## 1 1 0 0 0 0 4  
## 3 0 1 1 1 1 1  
## 2 0 1 1 0 1 2  
## 1 0 0 1 1 1 2  
## 6 445 454 464 510 1879

There are 6448 row with no missing values. 510 missing in CAR\_AGE, 464 missing in HOME\_VAL, 454 samples miss in YOJ (Year of JOB), 445 samplea miss in INCOME, and 6 samples miss in AGE

library(VIM)

## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

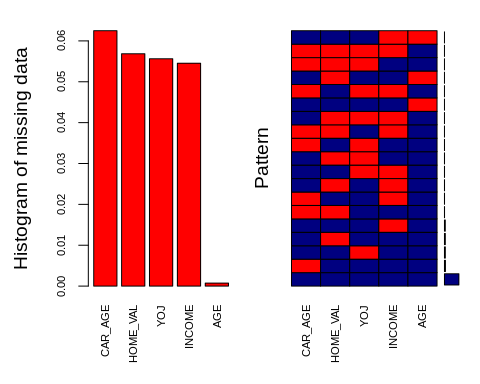
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##   
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':  
##   
## sleep

aggr\_plot <- aggr(ins.missing, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(ins.missing), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))

## Warning in plot.aggr(res, ...): not enough horizontal space to display  
## frequencies



##   
## Variables sorted by number of missings:   
## Variable Count  
## CAR\_AGE 0.062492342  
## HOME\_VAL 0.056855777  
## YOJ 0.055630437  
## INCOME 0.054527631  
## AGE 0.000735204

The plot helps us understanding that almost 90% of the samples are not missing any information, 6% are missing the CAR\_AGE, and the remaining ones show other missing patterns.

temp.ins <- mice(ins.missing,m=5,maxit=50,meth='rf',seed=123)

##   
## iter imp variable  
## 1 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 1 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 1 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 1 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 1 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 2 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 2 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 2 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 2 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 2 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 3 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 3 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 3 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 3 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 3 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 4 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 4 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 4 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 4 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 4 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 5 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 5 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 5 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 5 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 5 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 6 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 6 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 6 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 6 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 6 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 7 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 7 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 7 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 7 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 7 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 8 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 8 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 8 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 8 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 8 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 9 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 9 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 9 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 9 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 9 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 10 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 10 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 10 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 10 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 10 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 11 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 11 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 11 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 11 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 11 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 12 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 12 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 12 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 12 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 12 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 13 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 13 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 13 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 13 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 13 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 14 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 14 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 14 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 14 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 14 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 15 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 15 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 15 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 15 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 15 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 16 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 16 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 16 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 16 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 16 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 17 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 17 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 17 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 17 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 17 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 18 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 18 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 18 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 18 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 18 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 19 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 19 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 19 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 19 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 19 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 20 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 20 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 20 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 20 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 20 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 21 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 21 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 21 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 21 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 21 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 22 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 22 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 22 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 22 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 22 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 23 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 23 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 23 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 23 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 23 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 24 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 24 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 24 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 24 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 24 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 25 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 25 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 25 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 25 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 25 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 26 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 26 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 26 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 26 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 26 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 27 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 27 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 27 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 27 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 27 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 28 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 28 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 28 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 28 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 28 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 29 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 29 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 29 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 29 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 29 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 30 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 30 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 30 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 30 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 30 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 31 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 31 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 31 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 31 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 31 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 32 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 32 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 32 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 32 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 32 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 33 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 33 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 33 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 33 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 33 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 34 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 34 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 34 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 34 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 34 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 35 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 35 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 35 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 35 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 35 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 36 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 36 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 36 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 36 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 36 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 37 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 37 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 37 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 37 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 37 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 38 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 38 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 38 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 38 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 38 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 39 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 39 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 39 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 39 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 39 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 40 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 40 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 40 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 40 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 40 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 41 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 41 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 41 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 41 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 41 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 42 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 42 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 42 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 42 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 42 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 43 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 43 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 43 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 43 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 43 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 44 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 44 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 44 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 44 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 44 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 45 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 45 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 45 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 45 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 45 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 46 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 46 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 46 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 46 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 46 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 47 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 47 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 47 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 47 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 47 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 48 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 48 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 48 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 48 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 48 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 49 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 49 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 49 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 49 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 49 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 50 1 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 50 2 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 50 3 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 50 4 AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## 50 5 AGE INCOME HOME\_VAL CAR\_AGE YOJ

summary(temp.ins)

## Class: mids  
## Number of multiple imputations: 5   
## Imputation methods:  
## AGE INCOME HOME\_VAL CAR\_AGE YOJ   
## "rf" "rf" "rf" "rf" "rf"   
## PredictorMatrix:  
## AGE INCOME HOME\_VAL CAR\_AGE YOJ  
## AGE 0 1 1 1 1  
## INCOME 1 0 1 1 1  
## HOME\_VAL 1 1 0 1 1  
## CAR\_AGE 1 1 1 0 1  
## YOJ 1 1 1 1 0

completed.ins <- complete(temp.ins, 1)

summary(completed.ins)

## AGE INCOME HOME\_VAL CAR\_AGE   
## Min. :16.00 Min. : 0 Min. : 0 Min. :-3.000   
## 1st Qu.:39.00 1st Qu.: 27957 1st Qu.: 0 1st Qu.: 1.000   
## Median :45.00 Median : 53899 Median :159199 Median : 8.000   
## Mean :44.78 Mean : 61810 Mean :152235 Mean : 8.266   
## 3rd Qu.:51.00 3rd Qu.: 85734 3rd Qu.:236963 3rd Qu.:12.000   
## Max. :81.00 Max. :367030 Max. :885282 Max. :28.000   
## YOJ   
## Min. : 0.0   
## 1st Qu.: 9.0   
## Median :11.0   
## Mean :10.5   
## 3rd Qu.:13.0   
## Max. :23.0

str(completed.ins)

## 'data.frame': 8161 obs. of 5 variables:  
## $ AGE : int 60 43 35 51 50 34 54 37 34 50 ...  
## $ INCOME : num 67349 91449 16039 94160 114986 ...  
## $ HOME\_VAL: num 0 257252 124191 306251 243925 ...  
## $ CAR\_AGE : int 18 1 10 6 17 7 1 7 1 17 ...  
## $ YOJ : int 11 11 10 14 8 12 7 9 10 7 ...

costomers.insurance <- cbind(cost.ins[,-c(1, 5, 7, 8, 10, 25)], completed.ins)  
str(costomers.insurance)

## 'data.frame': 8161 obs. of 25 variables:  
## $ TARGET\_FLAG: Factor w/ 2 levels "NO","YES": 1 1 1 1 1 2 1 2 2 1 ...  
## $ TARGET\_AMT : num 0 0 0 0 0 ...  
## $ KIDSDRIV : int 0 0 0 0 0 0 0 1 0 0 ...  
## $ HOMEKIDS : Factor w/ 6 levels "HK1","HK2","HK3",..: 1 1 2 1 1 2 1 3 1 1 ...  
## $ PARENT1 : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...  
## $ MSTATUS : chr "0" "0" "1" "1" ...  
## $ SEX : chr "1" "1" "0" "1" ...  
## $ EDUCATION : chr "5PH" "2HS" "2HS" "1HS" ...  
## $ JOB : Factor w/ 9 levels "","Clerical",..: 7 9 2 9 3 9 9 9 2 7 ...  
## $ TRAVTIME : int 14 22 5 32 36 46 33 44 34 48 ...  
## $ CAR\_USE : chr "1" "0" "1" "1" ...  
## $ BLUEBOOK : num 14230 14940 4010 15440 18000 ...  
## $ TIF : int 11 1 4 7 1 1 1 1 1 7 ...  
## $ CAR\_TYPE : Factor w/ 6 levels "Minivan","Panel Truck",..: 1 1 6 1 6 4 6 5 6 5 ...  
## $ RED\_CAR : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 1 2 1 1 ...  
## $ OLDCLAIM : num 4461 0 38690 0 19217 ...  
## $ CLM\_FREQ : Factor w/ 6 levels "CF1","CF2","CF3",..: 3 1 3 1 3 1 1 2 1 1 ...  
## $ REVOKED : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 2 1 1 ...  
## $ MVR\_PTS : int 3 0 3 0 3 0 0 10 0 1 ...  
## $ URBANICITY : chr "U" "U" "U" "U" ...  
## $ AGE : int 60 43 35 51 50 34 54 37 34 50 ...  
## $ INCOME : num 67349 91449 16039 94160 114986 ...  
## $ HOME\_VAL : num 0 257252 124191 306251 243925 ...  
## $ CAR\_AGE : int 18 1 10 6 17 7 1 7 1 17 ...  
## $ YOJ : int 11 11 10 14 8 12 7 9 10 7 ...

summary(costomers.insurance)

## TARGET\_FLAG TARGET\_AMT KIDSDRIV HOMEKIDS PARENT1   
## NO :6008 Min. : 0 Min. :0.0000 HK1:5289 No :7084   
## YES:2153 1st Qu.: 0 1st Qu.:0.0000 HK2: 902 Yes:1077   
## Median : 0 Median :0.0000 HK3:1118   
## Mean : 1504 Mean :0.1711 HK4: 674   
## 3rd Qu.: 1036 3rd Qu.:0.0000 HK5: 164   
## Max. :107586 Max. :4.0000 HK6: 14   
##   
## MSTATUS SEX EDUCATION JOB   
## Length:8161 Length:8161 Length:8161 z\_Blue Collar:1825   
## Class :character Class :character Class :character Clerical :1271   
## Mode :character Mode :character Mode :character Professional :1117   
## Manager : 988   
## Lawyer : 835   
## Student : 712   
## (Other) :1413   
## TRAVTIME CAR\_USE BLUEBOOK TIF   
## Min. : 5.00 Length:8161 Min. : 1500 Min. : 1.000   
## 1st Qu.: 22.00 Class :character 1st Qu.: 9280 1st Qu.: 1.000   
## Median : 33.00 Mode :character Median :14440 Median : 4.000   
## Mean : 33.49 Mean :15710 Mean : 5.351   
## 3rd Qu.: 44.00 3rd Qu.:20850 3rd Qu.: 7.000   
## Max. :142.00 Max. :69740 Max. :25.000   
##   
## CAR\_TYPE RED\_CAR OLDCLAIM CLM\_FREQ REVOKED   
## Minivan :2145 no :5783 Min. : 0 CF1:5009 No :7161   
## Panel Truck: 676 yes:2378 1st Qu.: 0 CF2: 997 Yes:1000   
## Pickup :1389 Median : 0 CF3:1171   
## Sports Car : 907 Mean : 4037 CF4: 776   
## Van : 750 3rd Qu.: 4636 CF5: 190   
## z\_SUV :2294 Max. :57037 CF6: 18   
##   
## MVR\_PTS URBANICITY AGE INCOME   
## Min. : 0.000 Length:8161 Min. :16.00 Min. : 0   
## 1st Qu.: 0.000 Class :character 1st Qu.:39.00 1st Qu.: 27957   
## Median : 1.000 Mode :character Median :45.00 Median : 53899   
## Mean : 1.696 Mean :44.78 Mean : 61810   
## 3rd Qu.: 3.000 3rd Qu.:51.00 3rd Qu.: 85734   
## Max. :13.000 Max. :81.00 Max. :367030   
##   
## HOME\_VAL CAR\_AGE YOJ   
## Min. : 0 Min. :-3.000 Min. : 0.0   
## 1st Qu.: 0 1st Qu.: 1.000 1st Qu.: 9.0   
## Median :159199 Median : 8.000 Median :11.0   
## Mean :152235 Mean : 8.266 Mean :10.5   
## 3rd Qu.:236963 3rd Qu.:12.000 3rd Qu.:13.0   
## Max. :885282 Max. :28.000 Max. :23.0   
##

Missing Values in JOB Detecting the missing values

levels(cost.ins$JOB)

## [1] "" "Clerical" "Doctor" "Home Maker"   
## [5] "Lawyer" "Manager" "Professional" "Student"   
## [9] "z\_Blue Collar"

cost.ins$JOB[cost.ins$JOB==""]

## [1]   
## [76]   
## [151]   
## [226]   
## [301]   
## [376]   
## [451]   
## [526]   
## 9 Levels: Clerical Doctor Home Maker Lawyer Manager Professional ... z\_Blue Collar

The Education and Income can determine the job status of an individual

cost.ins$JOB[cost.ins$INDEX==112] <- "Professional"  
cost.ins$JOB[cost.ins$INDEX==112] <- "Professional"  
cost.ins$JOB[cost.ins$INDEX==112] <- "Professional"  
cost.ins$JOB[cost.ins$INDEX==112] <- "Professional"

## BUILD MODELS

### Split the data into train and test set

library(caret)

## Loading required package: lattice

set.seed(25)  
inTraining <- createDataPartition(costomers.insurance$TARGET\_AMT, p = .75, list = FALSE)  
training <- costomers.insurance[ inTraining,-1]  
testing <- costomers.insurance[-inTraining,-1]

### Build Models with the trainset using k-10 fold

#### Multiple Linear Model

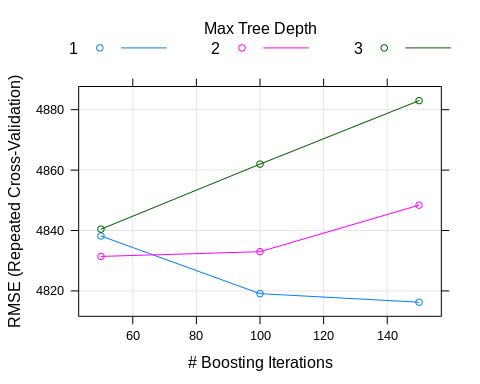
Linear regression using: 1. caret package

fitControl <- trainControl(## 10-fold CV  
 method = "repeatedcv",  
 number = 3,  
 ## repeated ten times  
 repeats = 10)  
  
  
set.seed(25)  
gbmFit1 <- train(TARGET\_AMT ~ ., data = training,   
 method = "gbm",   
 trControl = fitControl,  
 ## This last option is actually one  
 ## for gbm() that passes through  
 verbose = FALSE)  
gbmFit1

## Stochastic Gradient Boosting   
##   
## 6121 samples  
## 23 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (3 fold, repeated 10 times)   
## Summary of sample sizes: 4080, 4081, 4081, 4080, 4081, 4081, ...   
## Resampling results across tuning parameters:  
##   
## interaction.depth n.trees RMSE Rsquared MAE   
## 1 50 4838.177 0.04768719 2090.980  
## 1 100 4819.067 0.05373253 2041.509  
## 1 150 4816.242 0.05512888 2044.951  
## 2 50 4831.408 0.04967359 2063.390  
## 2 100 4832.972 0.05271478 2040.896  
## 2 150 4848.386 0.05117561 2063.607  
## 3 50 4840.435 0.04818421 2047.955  
## 3 100 4861.978 0.04845701 2055.841  
## 3 150 4883.003 0.04761540 2085.481  
##   
## Tuning parameter 'shrinkage' was held constant at a value of 0.1  
##   
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were n.trees = 150, interaction.depth =  
## 1, shrinkage = 0.1 and n.minobsinnode = 10.

Selecting models

plot(gbmFit1)



GBM model

# gbmGrid <- expand.grid(interaction.depth = c(1, 5, 9),   
# n.trees = (1:30)\*50,   
# shrinkage = 0.1,  
# n.minobsinnode = 20)  
#   
# nrow(gbmGrid)  
#   
# set.seed(25)  
# gbmFit2 <- train(TARGET\_AMT ~ ., data = training,   
# method = "gbm",   
# trControl = fitControl,   
# verbose = FALSE,   
# ## Now specify the exact models   
# ## to evaluate:  
# tuneGrid = gbmGrid)  
# gbmFit2

1. Partial Lesat Square

set.seed(25)  
fitControl <- trainControl(  
 method = "repeatedcv",  
 classProbs = FALSE,  
 repeats = 3,  
 preProc = c("center", "scale"))  
  
plsFit1 <- train(TARGET\_AMT~., data=training, method = "pls",   
 tuneLength=15, metric="RMSE",   
 trControl=fitControl)  
plsFit1

## Partial Least Squares   
##   
## 6121 samples  
## 23 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 5508, 5509, 5509, 5509, 5509, 5509, ...   
## Resampling results across tuning parameters:  
##   
## ncomp RMSE Rsquared MAE   
## 1 4867.406 0.009308241 2271.713  
## 2 4864.399 0.010357020 2254.661  
## 3 4855.370 0.014025388 2244.645  
## 4 4855.531 0.013950093 2246.348  
## 5 4850.180 0.015943579 2231.883  
## 6 4835.794 0.022555090 2204.096  
## 7 4821.088 0.028850046 2190.481  
## 8 4801.351 0.037923865 2153.921  
## 9 4795.341 0.040560343 2139.773  
## 10 4768.709 0.051832081 2091.921  
## 11 4736.043 0.065423978 2067.701  
## 12 4725.942 0.069045160 2065.656  
## 13 4723.157 0.070187205 2075.221  
## 14 4722.851 0.070515176 2072.210  
## 15 4722.980 0.070513603 2076.842  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was ncomp = 14.

1. Using step function

step.lmod <- lm(TARGET\_AMT ~ ., data = training)  
step.lmod <- step(step.lmod, trace = FALSE)

summary(step.lmod)

##   
## Call:  
## lm(formula = TARGET\_AMT ~ KIDSDRIV + PARENT1 + MSTATUS + SEX +   
## JOB + TRAVTIME + CAR\_USE + BLUEBOOK + TIF + CAR\_TYPE + OLDCLAIM +   
## CLM\_FREQ + REVOKED + MVR\_PTS + URBANICITY + INCOME + CAR\_AGE,   
## data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6202 -1748 -771 399 103199   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.218e+02 5.333e+02 -0.791 0.429029   
## KIDSDRIV 4.207e+02 1.257e+02 3.348 0.000820 \*\*\*  
## PARENT1Yes 6.773e+02 2.154e+02 3.145 0.001671 \*\*   
## MSTATUS1 -6.495e+02 1.451e+02 -4.477 7.70e-06 \*\*\*  
## SEX1 4.270e+02 1.965e+02 2.174 0.029771 \*   
## JOBClerical 2.143e+02 3.559e+02 0.602 0.547105   
## JOBDoctor -4.765e+02 4.553e+02 -1.047 0.295313   
## JOBHome Maker -5.246e+00 4.097e+02 -0.013 0.989784   
## JOBLawyer 4.261e+01 3.466e+02 0.123 0.902175   
## JOBManager -8.978e+02 3.255e+02 -2.758 0.005831 \*\*   
## JOBProfessional 2.007e+02 3.226e+02 0.622 0.533862   
## JOBStudent -4.952e+01 3.937e+02 -0.126 0.899895   
## JOBz\_Blue Collar 2.783e+02 3.235e+02 0.860 0.389766   
## TRAVTIME 1.206e+01 3.936e+00 3.064 0.002195 \*\*   
## CAR\_USE1 -6.393e+02 1.919e+02 -3.332 0.000867 \*\*\*  
## BLUEBOOK 1.808e-02 1.035e-02 1.746 0.080905 .   
## TIF -5.167e+01 1.473e+01 -3.508 0.000455 \*\*\*  
## CAR\_TYPEPanel Truck 6.108e+02 3.359e+02 1.818 0.069068 .   
## CAR\_TYPEPickup 6.412e+02 2.063e+02 3.107 0.001896 \*\*   
## CAR\_TYPESports Car 1.312e+03 2.646e+02 4.958 7.31e-07 \*\*\*  
## CAR\_TYPEVan 8.446e+02 2.555e+02 3.306 0.000951 \*\*\*  
## CAR\_TYPEz\_SUV 9.779e+02 2.185e+02 4.475 7.77e-06 \*\*\*  
## OLDCLAIM -1.616e-02 9.704e-03 -1.666 0.095855 .   
## CLM\_FREQCF2 5.579e+02 2.317e+02 2.408 0.016084 \*   
## CLM\_FREQCF3 3.723e+02 2.203e+02 1.690 0.091082 .   
## CLM\_FREQCF4 6.836e+02 2.491e+02 2.744 0.006088 \*\*   
## CLM\_FREQCF5 5.781e+02 4.266e+02 1.355 0.175431   
## CLM\_FREQCF6 2.382e+02 1.337e+03 0.178 0.858584   
## REVOKEDYes 6.602e+02 2.124e+02 3.109 0.001884 \*\*   
## MVR\_PTS 1.925e+02 3.246e+01 5.930 3.19e-09 \*\*\*  
## URBANICITYU 1.627e+03 1.711e+02 9.509 < 2e-16 \*\*\*  
## INCOME -5.300e-03 1.926e-03 -2.752 0.005950 \*\*   
## CAR\_AGE -2.072e+01 1.324e+01 -1.565 0.117732   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4786 on 6088 degrees of freedom  
## Multiple R-squared: 0.07646, Adjusted R-squared: 0.07161   
## F-statistic: 15.75 on 32 and 6088 DF, p-value: < 2.2e-16

predictor.importance <- varImp(step.lmod)  
predictor.importance

## Overall  
## KIDSDRIV 3.34767196  
## PARENT1Yes 3.14465889  
## MSTATUS1 4.47714749  
## SEX1 2.17363861  
## JOBClerical 0.60213751  
## JOBDoctor 1.04662836  
## JOBHome Maker 0.01280461  
## JOBLawyer 0.12291993  
## JOBManager 2.75813332  
## JOBProfessional 0.62215693  
## JOBStudent 0.12579882  
## JOBz\_Blue Collar 0.86010379  
## TRAVTIME 3.06379497  
## CAR\_USE1 3.33204312  
## BLUEBOOK 1.74574828  
## TIF 3.50795542  
## CAR\_TYPEPanel Truck 1.81829648  
## CAR\_TYPEPickup 3.10739638  
## CAR\_TYPESports Car 4.95804812  
## CAR\_TYPEVan 3.30631792  
## CAR\_TYPEz\_SUV 4.47519354  
## OLDCLAIM 1.66554869  
## CLM\_FREQCF2 2.40768166  
## CLM\_FREQCF3 1.68998308  
## CLM\_FREQCF4 2.74395169  
## CLM\_FREQCF5 1.35511606  
## CLM\_FREQCF6 0.17818413  
## REVOKEDYes 3.10922211  
## MVR\_PTS 5.93015622  
## URBANICITYU 9.50889820  
## INCOME 2.75150986  
## CAR\_AGE 1.56458604

Most important predictors are UBANICITY, MVR\_PTS, MSTATUS, CAR\_USE, CAR\_TYPE, TIF, TRAVTIME, KIDSDRIVKD, REVOKED, CAR\_AGE, PARENT1, INCOME, AGE, HOME\_VAL,

1. rlm from MASS package We use the Huber weigth parameter

library(MASS)  
rr.huber <- rlm(TARGET\_AMT ~ KIDSDRIV + PARENT1 + MSTATUS + JOB +   
 TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + REVOKED + MVR\_PTS +   
 URBANICITY + AGE + INCOME + HOME\_VAL + CAR\_AGE, data = training)

## Warning in rlm.default(x, y, weights, method = method, wt.method = wt.method, :  
## 'rlm' failed to converge in 20 steps

rr.huber

## Call:  
## rlm(formula = TARGET\_AMT ~ KIDSDRIV + PARENT1 + MSTATUS + JOB +   
## TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + REVOKED + MVR\_PTS +   
## URBANICITY + AGE + INCOME + HOME\_VAL + CAR\_AGE, data = training)  
## Ran 20 iterations without convergence  
##   
## Coefficients:  
## (Intercept) KIDSDRIV PARENT1Yes MSTATUS1   
## -1.233873e+02 1.444448e+02 3.310297e+02 -1.453054e+02   
## JOBClerical JOBDoctor JOBHome Maker JOBLawyer   
## 2.428367e+02 2.226210e+01 1.615710e+02 4.456314e+01   
## JOBManager JOBProfessional JOBStudent JOBz\_Blue Collar   
## -1.518386e+02 8.914443e+01 1.993392e+02 1.889387e+02   
## TRAVTIME CAR\_USE1 TIF CAR\_TYPEPanel Truck   
## 3.806003e+00 -2.805460e+02 -1.355942e+01 5.877894e+01   
## CAR\_TYPEPickup CAR\_TYPESports Car CAR\_TYPEVan CAR\_TYPEz\_SUV   
## 1.973709e+02 3.100171e+02 1.219949e+02 1.971162e+02   
## REVOKEDYes MVR\_PTS URBANICITYU AGE   
## 4.137291e+02 7.429881e+01 5.758875e+02 -1.069504e-01   
## INCOME HOME\_VAL CAR\_AGE   
## -1.208614e-03 -3.559723e-04 -2.779699e+00   
##   
## Degrees of freedom: 6121 total; 6094 residual  
## Scale estimate: 730

sqrt(mean(rr.huber$residuals^2))

## [1] 4986.62

library(leaps)  
sub.models <- regsubsets(TARGET\_AMT~., data = training, method = "exhaustive", nvmax = 32)  
rs <- summary(sub.models)  
rs$which

## (Intercept) KIDSDRIV HOMEKIDSHK2 HOMEKIDSHK3 HOMEKIDSHK4 HOMEKIDSHK5  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE  
## 2 TRUE FALSE FALSE FALSE FALSE FALSE  
## 3 TRUE FALSE FALSE FALSE FALSE FALSE  
## 4 TRUE FALSE FALSE FALSE FALSE FALSE  
## 5 TRUE FALSE FALSE FALSE FALSE FALSE  
## 6 TRUE FALSE FALSE FALSE FALSE FALSE  
## 7 TRUE TRUE FALSE FALSE FALSE FALSE  
## 8 TRUE TRUE FALSE FALSE FALSE FALSE  
## 9 TRUE TRUE FALSE FALSE FALSE FALSE  
## 10 TRUE TRUE FALSE FALSE FALSE FALSE  
## 11 TRUE TRUE FALSE FALSE FALSE FALSE  
## 12 TRUE TRUE FALSE FALSE FALSE FALSE  
## 13 TRUE TRUE FALSE FALSE FALSE FALSE  
## 14 TRUE TRUE FALSE FALSE FALSE FALSE  
## 15 TRUE TRUE FALSE FALSE FALSE FALSE  
## 16 TRUE TRUE FALSE FALSE FALSE FALSE  
## 17 TRUE TRUE FALSE FALSE FALSE FALSE  
## 18 TRUE TRUE FALSE FALSE FALSE FALSE  
## 19 TRUE TRUE FALSE FALSE FALSE FALSE  
## 20 TRUE TRUE FALSE FALSE FALSE FALSE  
## 21 TRUE TRUE FALSE FALSE FALSE FALSE  
## 22 TRUE TRUE FALSE FALSE FALSE FALSE  
## 23 TRUE TRUE FALSE FALSE FALSE FALSE  
## 24 TRUE TRUE FALSE TRUE FALSE FALSE  
## 25 TRUE TRUE FALSE FALSE FALSE FALSE  
## 26 TRUE TRUE FALSE FALSE FALSE FALSE  
## 27 TRUE TRUE FALSE TRUE FALSE FALSE  
## 28 TRUE TRUE FALSE TRUE FALSE FALSE  
## 29 TRUE TRUE FALSE TRUE FALSE FALSE  
## 30 TRUE TRUE FALSE TRUE FALSE FALSE  
## 31 TRUE TRUE FALSE TRUE FALSE FALSE  
## 32 TRUE TRUE FALSE TRUE FALSE FALSE  
## HOMEKIDSHK6 PARENT1Yes MSTATUS1 SEX1 EDUCATION2HS EDUCATION3BA EDUCATION4MA  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 4 FALSE TRUE FALSE FALSE FALSE FALSE FALSE  
## 5 FALSE TRUE FALSE FALSE FALSE FALSE FALSE  
## 6 FALSE TRUE FALSE FALSE FALSE FALSE FALSE  
## 7 FALSE FALSE TRUE FALSE FALSE FALSE FALSE  
## 8 FALSE FALSE TRUE FALSE FALSE FALSE FALSE  
## 9 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 10 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 11 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 12 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 13 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 14 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 15 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 16 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 17 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 18 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 19 FALSE TRUE TRUE FALSE FALSE FALSE FALSE  
## 20 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 21 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 22 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 23 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 24 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 25 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 26 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 27 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 28 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 29 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 30 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 31 FALSE TRUE TRUE TRUE FALSE FALSE FALSE  
## 32 FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## EDUCATION5PH JOBClerical JOBDoctor JOBHome Maker JOBLawyer JOBManager  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE FALSE FALSE  
## 4 FALSE FALSE FALSE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE FALSE FALSE TRUE  
## 6 FALSE FALSE FALSE FALSE FALSE TRUE  
## 7 FALSE FALSE FALSE FALSE FALSE TRUE  
## 8 FALSE FALSE FALSE FALSE FALSE TRUE  
## 9 FALSE FALSE FALSE FALSE FALSE TRUE  
## 10 FALSE FALSE FALSE FALSE FALSE TRUE  
## 11 FALSE FALSE FALSE FALSE FALSE TRUE  
## 12 FALSE FALSE FALSE FALSE FALSE TRUE  
## 13 FALSE FALSE FALSE FALSE FALSE TRUE  
## 14 FALSE FALSE FALSE FALSE FALSE TRUE  
## 15 FALSE FALSE FALSE FALSE FALSE TRUE  
## 16 FALSE FALSE FALSE FALSE FALSE TRUE  
## 17 FALSE FALSE FALSE FALSE FALSE TRUE  
## 18 FALSE FALSE FALSE FALSE FALSE TRUE  
## 19 FALSE FALSE FALSE FALSE FALSE TRUE  
## 20 FALSE FALSE FALSE FALSE FALSE TRUE  
## 21 FALSE FALSE FALSE FALSE FALSE TRUE  
## 22 FALSE FALSE TRUE FALSE FALSE TRUE  
## 23 FALSE FALSE TRUE FALSE FALSE TRUE  
## 24 FALSE FALSE TRUE FALSE FALSE TRUE  
## 25 FALSE FALSE TRUE FALSE FALSE TRUE  
## 26 FALSE FALSE TRUE FALSE FALSE TRUE  
## 27 FALSE FALSE TRUE FALSE FALSE TRUE  
## 28 FALSE FALSE TRUE FALSE FALSE TRUE  
## 29 FALSE FALSE TRUE FALSE FALSE TRUE  
## 30 FALSE FALSE TRUE FALSE FALSE TRUE  
## 31 FALSE TRUE TRUE FALSE FALSE TRUE  
## 32 FALSE TRUE TRUE FALSE FALSE TRUE  
## JOBProfessional JOBStudent JOBz\_Blue Collar TRAVTIME CAR\_USE1 BLUEBOOK TIF  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## 4 FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## 5 FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## 6 FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## 7 FALSE FALSE FALSE FALSE TRUE FALSE FALSE  
## 8 FALSE FALSE FALSE FALSE TRUE FALSE TRUE  
## 9 FALSE FALSE FALSE FALSE TRUE FALSE TRUE  
## 10 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 11 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 12 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 13 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 14 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 15 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 16 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 17 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 18 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 19 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 20 FALSE FALSE FALSE TRUE TRUE FALSE TRUE  
## 21 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 22 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 23 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 24 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 25 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 26 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 27 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 28 FALSE FALSE FALSE TRUE TRUE TRUE TRUE  
## 29 FALSE FALSE TRUE TRUE TRUE TRUE TRUE  
## 30 TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 31 TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## 32 TRUE FALSE TRUE TRUE TRUE TRUE TRUE  
## CAR\_TYPEPanel Truck CAR\_TYPEPickup CAR\_TYPESports Car CAR\_TYPEVan  
## 1 FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE  
## 4 FALSE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE FALSE  
## 6 FALSE FALSE FALSE FALSE  
## 7 FALSE FALSE FALSE FALSE  
## 8 FALSE FALSE FALSE FALSE  
## 9 FALSE FALSE FALSE FALSE  
## 10 FALSE FALSE FALSE FALSE  
## 11 FALSE FALSE FALSE FALSE  
## 12 FALSE FALSE TRUE FALSE  
## 13 FALSE FALSE TRUE FALSE  
## 14 FALSE FALSE TRUE TRUE  
## 15 TRUE TRUE TRUE TRUE  
## 16 TRUE TRUE TRUE TRUE  
## 17 TRUE TRUE TRUE TRUE  
## 18 TRUE TRUE TRUE TRUE  
## 19 TRUE TRUE TRUE TRUE  
## 20 TRUE TRUE TRUE TRUE  
## 21 TRUE TRUE TRUE TRUE  
## 22 TRUE TRUE TRUE TRUE  
## 23 TRUE TRUE TRUE TRUE  
## 24 TRUE TRUE TRUE TRUE  
## 25 TRUE TRUE TRUE TRUE  
## 26 TRUE TRUE TRUE TRUE  
## 27 TRUE TRUE TRUE TRUE  
## 28 TRUE TRUE TRUE TRUE  
## 29 TRUE TRUE TRUE TRUE  
## 30 TRUE TRUE TRUE TRUE  
## 31 TRUE TRUE TRUE TRUE  
## 32 TRUE TRUE TRUE TRUE  
## CAR\_TYPEz\_SUV RED\_CARyes OLDCLAIM CLM\_FREQCF2 CLM\_FREQCF3 CLM\_FREQCF4  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE FALSE FALSE FALSE  
## 4 FALSE FALSE FALSE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE FALSE FALSE FALSE  
## 6 FALSE FALSE FALSE FALSE FALSE FALSE  
## 7 FALSE FALSE FALSE FALSE FALSE FALSE  
## 8 FALSE FALSE FALSE FALSE FALSE FALSE  
## 9 FALSE FALSE FALSE FALSE FALSE FALSE  
## 10 FALSE FALSE FALSE FALSE FALSE FALSE  
## 11 FALSE FALSE FALSE FALSE FALSE FALSE  
## 12 FALSE FALSE FALSE FALSE FALSE FALSE  
## 13 FALSE FALSE FALSE FALSE FALSE FALSE  
## 14 TRUE FALSE FALSE FALSE FALSE FALSE  
## 15 TRUE FALSE FALSE FALSE FALSE FALSE  
## 16 TRUE FALSE FALSE FALSE FALSE FALSE  
## 17 TRUE FALSE FALSE FALSE FALSE FALSE  
## 18 TRUE FALSE FALSE FALSE FALSE FALSE  
## 19 TRUE FALSE FALSE FALSE FALSE TRUE  
## 20 TRUE FALSE FALSE FALSE FALSE TRUE  
## 21 TRUE FALSE FALSE FALSE FALSE TRUE  
## 22 TRUE FALSE FALSE FALSE FALSE TRUE  
## 23 TRUE FALSE FALSE TRUE FALSE TRUE  
## 24 TRUE FALSE FALSE TRUE FALSE TRUE  
## 25 TRUE FALSE TRUE TRUE TRUE TRUE  
## 26 TRUE FALSE TRUE TRUE TRUE TRUE  
## 27 TRUE FALSE TRUE TRUE TRUE TRUE  
## 28 TRUE FALSE TRUE TRUE TRUE TRUE  
## 29 TRUE FALSE TRUE TRUE TRUE TRUE  
## 30 TRUE FALSE TRUE TRUE TRUE TRUE  
## 31 TRUE FALSE TRUE TRUE TRUE TRUE  
## 32 TRUE FALSE TRUE TRUE TRUE TRUE  
## CLM\_FREQCF5 CLM\_FREQCF6 REVOKEDYes MVR\_PTS URBANICITYU AGE INCOME HOME\_VAL  
## 1 FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE  
## 2 FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE  
## 3 FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE  
## 4 FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE  
## 5 FALSE FALSE FALSE TRUE TRUE FALSE FALSE FALSE  
## 6 FALSE FALSE FALSE TRUE TRUE FALSE FALSE TRUE  
## 7 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE  
## 8 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE  
## 9 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE  
## 10 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE  
## 11 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 12 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 13 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 14 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 15 FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE  
## 16 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 17 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 18 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 19 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 20 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 21 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 22 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 23 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 24 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 25 FALSE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 26 TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 27 TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE  
## 28 TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE  
## 29 TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE  
## 30 TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE  
## 31 TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE  
## 32 TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE  
## CAR\_AGE YOJ  
## 1 FALSE FALSE  
## 2 FALSE FALSE  
## 3 FALSE FALSE  
## 4 FALSE FALSE  
## 5 FALSE FALSE  
## 6 FALSE FALSE  
## 7 FALSE FALSE  
## 8 FALSE FALSE  
## 9 FALSE FALSE  
## 10 FALSE FALSE  
## 11 FALSE FALSE  
## 12 FALSE FALSE  
## 13 TRUE FALSE  
## 14 FALSE FALSE  
## 15 FALSE FALSE  
## 16 FALSE FALSE  
## 17 TRUE FALSE  
## 18 TRUE TRUE  
## 19 TRUE TRUE  
## 20 TRUE TRUE  
## 21 TRUE TRUE  
## 22 TRUE TRUE  
## 23 TRUE TRUE  
## 24 TRUE TRUE  
## 25 TRUE TRUE  
## 26 TRUE TRUE  
## 27 TRUE TRUE  
## 28 TRUE TRUE  
## 29 TRUE TRUE  
## 30 TRUE TRUE  
## 31 TRUE TRUE  
## 32 TRUE TRUE

Examine model further - list outputs available with names() :

best.summary <- summary(sub.models)  
names(best.summary)

## [1] "which" "rsq" "rss" "adjr2" "cp" "bic" "outmat" "obj"

Explore model with lowest RSS and highest R-squared:

which.min(best.summary$rss)

## [1] 32

which.max(best.summary$adjr2)

## [1] 26

which.min(best.summary$bic)

## [1] 10

#### Binary Logistic Regression

### Split the data into train and test set

library(caret)  
set.seed(25)  
inTraining <- createDataPartition(costomers.insurance$TARGET\_AMT, p = .75, list = FALSE)  
training <- costomers.insurance[ inTraining,-2]  
testing <- costomers.insurance[-inTraining,-2]

1. glm logit and probit

logitFit <- glm(TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + JOB +   
 TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + REVOKED + MVR\_PTS +   
 URBANICITY + AGE + INCOME + HOME\_VAL + CAR\_AGE, family = binomial(link = "logit"), data = cost.ins)  
logitFit

##   
## Call: glm(formula = TARGET\_FLAG ~ KIDSDRIV + PARENT1 + MSTATUS + JOB +   
## TRAVTIME + CAR\_USE + TIF + CAR\_TYPE + REVOKED + MVR\_PTS +   
## URBANICITY + AGE + INCOME + HOME\_VAL + CAR\_AGE, family = binomial(link = "logit"),   
## data = cost.ins)  
##   
## Coefficients:  
## (Intercept) KIDSDRIV PARENT1Yes   
## -2.975e+00 3.675e-01 4.376e-01   
## MSTATUS1 JOBClerical JOBDoctor   
## -4.689e-01 4.744e-01 -1.430e-01   
## JOBHome Maker JOBLawyer JOBManager   
## 2.122e-01 7.737e-02 -6.780e-01   
## JOBProfessional JOBStudent JOBz\_Blue Collar   
## 1.958e-02 2.484e-01 3.161e-01   
## TRAVTIME CAR\_USE1 TIF   
## 1.544e-02 -7.531e-01 -5.449e-02   
## CAR\_TYPEPanel Truck CAR\_TYPEPickup CAR\_TYPESports Car   
## 4.567e-01 5.959e-01 1.051e+00   
## CAR\_TYPEVan CAR\_TYPEz\_SUV REVOKEDYes   
## 5.246e-01 8.261e-01 7.504e-01   
## MVR\_PTS URBANICITYU AGE   
## 1.354e-01 2.371e+00 -5.475e-03   
## INCOME HOME\_VAL CAR\_AGE   
## -4.817e-06 -1.343e-06 -1.427e-02   
##   
## Degrees of Freedom: 6832 Total (i.e. Null); 6806 Residual  
## (1328 observations deleted due to missingness)  
## Null Deviance: 7888   
## Residual Deviance: 6205 AIC: 6259

Selecting models 2. GBM with tuning parameters

gbmGrid <- expand.grid(interaction.depth = c(1, 5, 9),  
 n.trees = (1:30)\*50,  
 shrinkage = 0.1,  
 n.minobsinnode = 20)  
  
nrow(gbmGrid)

## [1] 90

fitControl <- trainControl(method = "repeatedcv",  
 number = 5,  
 repeats = 3,  
 ## Estimate class probabilities  
 classProbs = TRUE,  
 ## Evaluate performance using   
 ## the following function  
 summaryFunction = twoClassSummary)  
  
set.seed(25)  
gbmFit3 <- train(TARGET\_FLAG~ ., data = training,   
 method = "gbm",   
 trControl = fitControl,   
 verbose = FALSE,   
 tuneGrid = gbmGrid,  
 ## Specify which metric to optimize  
 metric = "ROC")  
gbmFit3

## Stochastic Gradient Boosting   
##   
## 6121 samples  
## 23 predictor  
## 2 classes: 'NO', 'YES'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold, repeated 3 times)   
## Summary of sample sizes: 4897, 4898, 4897, 4896, 4896, 4897, ...   
## Resampling results across tuning parameters:  
##   
## interaction.depth n.trees ROC Sens Spec   
## 1 50 0.7846826 0.9845433 0.1466396  
## 1 100 0.8022925 0.9640554 0.2711779  
## 1 150 0.8104378 0.9495596 0.3310681  
## 1 200 0.8139536 0.9419419 0.3678294  
## 1 250 0.8161578 0.9360997 0.3901345  
## 1 300 0.8171342 0.9322548 0.4014916  
## 1 350 0.8173836 0.9295912 0.4103699  
## 1 400 0.8177587 0.9284819 0.4149158  
## 1 450 0.8176976 0.9261890 0.4242037  
## 1 500 0.8178606 0.9240444 0.4262690  
## 1 550 0.8177774 0.9241181 0.4330859  
## 1 600 0.8174751 0.9243403 0.4306053  
## 1 650 0.8168240 0.9227870 0.4359704  
## 1 700 0.8164378 0.9219736 0.4370030  
## 1 750 0.8161250 0.9207163 0.4372081  
## 1 800 0.8161038 0.9184976 0.4448506  
## 1 850 0.8157490 0.9191632 0.4425815  
## 1 900 0.8159858 0.9192365 0.4440205  
## 1 950 0.8155145 0.9192375 0.4456756  
## 1 1000 0.8153702 0.9193107 0.4458833  
## 1 1050 0.8154264 0.9182756 0.4467050  
## 1 1100 0.8150268 0.9175361 0.4479466  
## 1 1150 0.8147586 0.9167227 0.4500119  
## 1 1200 0.8145428 0.9168706 0.4493927  
## 1 1250 0.8141917 0.9180548 0.4477415  
## 1 1300 0.8137164 0.9162791 0.4522874  
## 1 1350 0.8137788 0.9163529 0.4500131  
## 1 1400 0.8135270 0.9156873 0.4469172  
## 1 1450 0.8133139 0.9167231 0.4508400  
## 1 1500 0.8129239 0.9159096 0.4502227  
## 5 50 0.8147697 0.9434209 0.3649341  
## 5 100 0.8194126 0.9275194 0.4227673  
## 5 150 0.8181631 0.9194593 0.4345435  
## 5 200 0.8176155 0.9159096 0.4477639  
## 5 250 0.8162395 0.9116194 0.4533303  
## 5 300 0.8139543 0.9105094 0.4595235  
## 5 350 0.8124870 0.9086603 0.4617939  
## 5 400 0.8103730 0.9080695 0.4638655  
## 5 450 0.8084405 0.9062941 0.4673743  
## 5 500 0.8068726 0.9054811 0.4677807  
## 5 550 0.8047447 0.9021524 0.4700530  
## 5 600 0.8031804 0.9001552 0.4717042  
## 5 650 0.8014371 0.8993420 0.4721157  
## 5 700 0.7999985 0.8978628 0.4706716  
## 5 750 0.7984007 0.8972716 0.4704684  
## 5 800 0.7974701 0.8953481 0.4694364  
## 5 850 0.7970780 0.8961623 0.4665442  
## 5 900 0.7960955 0.8931293 0.4646860  
## 5 950 0.7959335 0.8933513 0.4669558  
## 5 1000 0.7951892 0.8934983 0.4690217  
## 5 1050 0.7940292 0.8909109 0.4684012  
## 5 1100 0.7935662 0.8919453 0.4712920  
## 5 1150 0.7925433 0.8900226 0.4715023  
## 5 1200 0.7918342 0.8895051 0.4688133  
## 5 1250 0.7908075 0.8889875 0.4690229  
## 5 1300 0.7899784 0.8886919 0.4692268  
## 5 1350 0.7892113 0.8867687 0.4698479  
## 5 1400 0.7890086 0.8867686 0.4714997  
## 5 1450 0.7876830 0.8858811 0.4710888  
## 5 1500 0.7868473 0.8864729 0.4710850  
## 9 50 0.8158829 0.9309955 0.4062471  
## 9 100 0.8157486 0.9178323 0.4374318  
## 9 150 0.8141147 0.9139124 0.4481741  
## 9 200 0.8111582 0.9105102 0.4510656  
## 9 250 0.8073741 0.9057772 0.4502433  
## 9 300 0.8047113 0.9017834 0.4576755  
## 9 350 0.8027659 0.8992677 0.4599504  
## 9 400 0.8004198 0.8982326 0.4649059  
## 9 450 0.7980771 0.8966801 0.4680031  
## 9 500 0.7965322 0.8939446 0.4702793  
## 9 550 0.7948396 0.8935754 0.4675897  
## 9 600 0.7924380 0.8924647 0.4715106  
## 9 650 0.7910581 0.8920957 0.4682050  
## 9 700 0.7904436 0.8911335 0.4686140  
## 9 750 0.7901280 0.8894321 0.4715023  
## 9 800 0.7889512 0.8894328 0.4661346  
## 9 850 0.7878775 0.8883229 0.4665539  
## 9 900 0.7878197 0.8888414 0.4677922  
## 9 950 0.7869631 0.8890630 0.4675846  
## 9 1000 0.7857919 0.8875099 0.4686185  
## 9 1050 0.7855878 0.8869914 0.4713042  
## 9 1100 0.7854272 0.8862523 0.4710927  
## 9 1150 0.7852166 0.8851427 0.4688255  
## 9 1200 0.7851779 0.8844778 0.4692364  
## 9 1250 0.7847167 0.8851431 0.4680012  
## 9 1300 0.7848626 0.8848467 0.4700639  
## 9 1350 0.7844656 0.8844031 0.4700620  
## 9 1400 0.7842022 0.8848473 0.4708882  
## 9 1450 0.7838938 0.8841073 0.4706857  
## 9 1500 0.7834704 0.8846985 0.4700620  
##   
## Tuning parameter 'shrinkage' was held constant at a value of 0.1  
##   
## Tuning parameter 'n.minobsinnode' was held constant at a value of 20  
## ROC was used to select the optimal model using the largest value.  
## The final values used for the model were n.trees = 100, interaction.depth =  
## 5, shrinkage = 0.1 and n.minobsinnode = 20.

whichTwoPct <- tolerance(gbmFit3$results, metric = "ROC",   
 tol = 2, maximize = FALSE)   
cat("best model within 2 pct of best:\n")

## best model within 2 pct of best:

gbmFit3$results[whichTwoPct,1:6]

## shrinkage interaction.depth n.minobsinnode n.trees ROC Sens  
## 1 0.1 1 20 50 0.7846826 0.9845433

Between Models Support Vector Machine is fit to the training set

set.seed(25)  
svmFit <- train(TARGET\_FLAG ~ ., data = training,   
 method = "svmRadial",   
 trControl = fitControl,   
 preProc = c("center", "scale"),  
 tuneLength = 8,  
 metric = "ROC")

## line search fails -1.243662 0.2075985 1.18036e-05 -6.736634e-06 -1.716852e-08 -4.038065e-10 -1.999301e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.8856834 0.5036143 1.300457e-05 -6.177812e-06 -1.248577e-08 2.541467e-09 -1.780728e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.8917007 0.5004546 1.453188e-05 -6.947213e-06 -1.406258e-08 2.864635e-09 -2.242569e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.248849 0.1937902 1.535071e-05 -8.4931e-06 -2.210199e-08 -8.934965e-10 -3.316927e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.00226 0.4328458 1.735623e-05 -8.822432e-06 -1.961619e-08 2.728158e-09 -3.645321e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.8738961 0.5101665 1.060835e-05 -5.087446e-06 -1.022003e-08 2.056164e-09 -1.188783e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.066577 0.3819111 1.230652e-05 -6.605165e-06 -1.50944e-08 1.506529e-09 -1.957104e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.9047566 0.5007584 1.663936e-05 -8.112626e-06 -1.63908e-08 3.612834e-09 -3.02042e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.9974541 0.4056774 1.203076e-05 -6.009165e-06 -1.358121e-08 1.196247e-09 -1.705808e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.015083 0.4036037 1.405987e-05 -7.21291e-06 -1.586485e-08 1.929416e-09 -2.369744e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.8804415 0.4985832 1.118072e-05 -5.450553e-06 -1.068468e-08 2.287583e-09 -1.319311e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.9904115 0.4227618 1.251524e-05 -6.353989e-06 -1.428678e-08 1.489191e-09 -1.882647e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.277279 0.2270574 2.010334e-05 -1.149765e-05 -3.006398e-08 1.229291e-10 -6.057998e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.208819 0.2769351 1.638258e-05 -9.367978e-06 -2.259721e-08 1.262052e-09 -3.820235e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.034376 0.4187783 1.335457e-05 -7.026376e-06 -1.543432e-08 2.271891e-09 -2.220818e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -0.8812915 0.4930629 1.019457e-05 -4.812849e-06 -9.795772e-09 1.747803e-09 -1.082756e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -2.551636 -1.063272 1.931778e-05 -1.302581e-05 -7.450742e-08 -3.932068e-08 -9.271344e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.443557 0.0369353 1.728571e-05 -1.007872e-05 -2.94413e-08 -3.666774e-09 -4.719573e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.18694 0.2921735 1.278084e-05 -7.073112e-06 -1.733515e-08 7.124291e-10 -2.265967e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## line search fails -1.057892 0.3728577 1.674853e-05 -8.532216e-06 -1.959993e-08 1.76049e-09 -3.432908e-13

## Warning in method$predict(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class prediction calculations failed; returning NAs

## Warning in method$prob(modelFit = modelFit, newdata = newdata, submodels =  
## param): kernlab class probability calculations failed; returning NAs

## Warning in data.frame(..., check.names = FALSE): row names were found from a  
## short variable and have been discarded

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :  
## There were missing values in resampled performance measures.

svmFit

## Support Vector Machines with Radial Basis Function Kernel   
##   
## 6121 samples  
## 23 predictor  
## 2 classes: 'NO', 'YES'   
##   
## Pre-processing: centered (45), scaled (45)   
## Resampling: Cross-Validated (5 fold, repeated 3 times)   
## Summary of sample sizes: 4897, 4898, 4897, 4896, 4896, 4897, ...   
## Resampling results across tuning parameters:  
##   
## C ROC Sens Spec   
## 0.25 0.8060189 0.9101425 0.4357235  
## 0.50 0.8077075 0.9247079 0.4076855  
## 1.00 0.8071746 0.9255220 0.4027325  
## 2.00 0.8036210 0.9299518 0.3969943  
## 4.00 0.7952786 0.9268682 0.3782626  
## 8.00 0.7870837 0.9329251 0.3527193  
## 16.00 0.7698161 0.9334358 0.3169384  
## 32.00 0.7590805 0.9405746 0.2630062  
##   
## Tuning parameter 'sigma' was held constant at a value of 0.01343115  
## ROC was used to select the optimal model using the largest value.  
## The final values used for the model were sigma = 0.01343115 and C = 0.5.

Discriminant Analysis is fit

set.seed(25)  
rdaFit <- train(TARGET\_FLAG ~ ., data = training,   
 method = "rda",   
 trControl = fitControl,   
 tuneLength = 4,  
 metric = "ROC")

## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo, :  
## There were missing values in resampled performance measures.

## Warning in train.default(x, y, weights = w, ...): missing values found in  
## aggregated results

rdaFit

## Regularized Discriminant Analysis   
##   
## 6121 samples  
## 23 predictor  
## 2 classes: 'NO', 'YES'   
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold, repeated 3 times)   
## Summary of sample sizes: 4897, 4898, 4897, 4896, 4896, 4897, ...   
## Resampling results across tuning parameters:  
##   
## gamma lambda ROC Sens Spec   
## 0.0000000 0.0000000 0.7807995 0.7787187 0.6395990  
## 0.0000000 0.3333333 0.7821815 0.7939524 0.6210122  
## 0.0000000 0.6666667 0.7909902 0.8270864 0.5823965  
## 0.0000000 1.0000000 0.8114949 0.9176841 0.4386670  
## 0.3333333 0.0000000 NaN 0.0000000 1.0000000  
## 0.3333333 0.3333333 NaN 0.0000000 1.0000000  
## 0.3333333 0.6666667 NaN 0.0000000 1.0000000  
## 0.3333333 1.0000000 NaN 0.0000000 1.0000000  
## 0.6666667 0.0000000 NaN 0.0000000 1.0000000  
## 0.6666667 0.3333333 NaN 0.0000000 1.0000000  
## 0.6666667 0.6666667 NaN 0.0000000 1.0000000  
## 0.6666667 1.0000000 NaN 0.0000000 1.0000000  
## 1.0000000 0.0000000 NaN 0.0000000 1.0000000  
## 1.0000000 0.3333333 NaN 0.0000000 1.0000000  
## 1.0000000 0.6666667 NaN 0.0000000 1.0000000  
## 1.0000000 1.0000000 NaN 0.0000000 1.0000000  
##   
## ROC was used to select the optimal model using the largest value.  
## The final values used for the model were gamma = 0 and lambda = 1.

Statistical Statements on the models perfomances

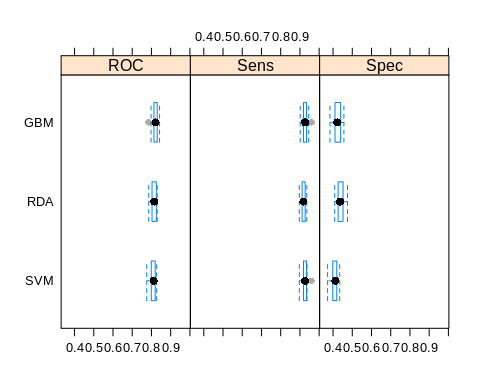
Resampling Results

resamps <- resamples(list(GBM = gbmFit3,  
 SVM = svmFit,  
 RDA = rdaFit))  
resamps

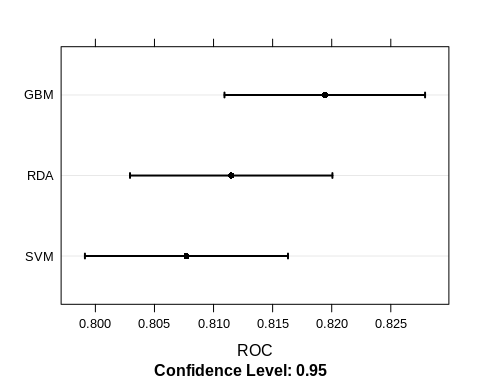
##   
## Call:  
## resamples.default(x = list(GBM = gbmFit3, SVM = svmFit, RDA = rdaFit))  
##   
## Models: GBM, SVM, RDA   
## Number of resamples: 15   
## Performance metrics: ROC, Sens, Spec   
## Time estimates for: everything, final model fit

Visualizing the resampling distribution

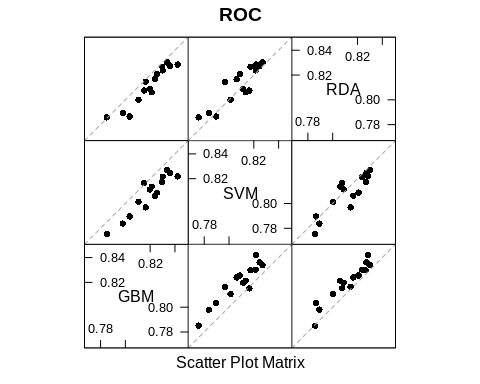
theme1 <- trellis.par.get()  
theme1$plot.symbol$col = rgb(.2, .2, .2, .4)  
theme1$plot.symbol$pch = 16  
theme1$plot.line$col = rgb(1, 0, 0, .7)  
theme1$plot.line$lwd <- 2  
trellis.par.set(theme1)  
bwplot(resamps, layout = c(3, 1))



trellis.par.set(caretTheme())  
dotplot(resamps, metric = "ROC")



splom(resamps)



Since models are fit on the same versions of the training data, it makes sense to make inferences on the differences between models. In this way we reduce the within-resample correlation that may exist. We can compute the differences, then use a simple t-test to evaluate the null hypothesis that there is no difference between models.

difValues <- diff(resamps)  
difValues

##   
## Call:  
## diff.resamples(x = resamps)  
##   
## Models: GBM, SVM, RDA   
## Metrics: ROC, Sens, Spec   
## Number of differences: 3   
## p-value adjustment: bonferroni