BA810-project

Teammates: Dongzhe Zhang, Kunpeng Huang, Chengyu Liang, Haolan Ma, Yihan Jiang, Meiling Zhang

10/10/2019

Set up

```r
library(tidyverse)
library(ggthemes)
library(randomForest)
library(gbm)
library(MASS)
library(dplyr)
library(tidyr)
library(glmnet)
library(rpart)
library(rpart.plot)
library(caret)
theme_set(theme_economist())
```

Data Cleaning

```r
original_dataset <- read_csv("original_dataset.csv")
```

1. A lot of columns contain missing values. Instead of replacing them with the median, we would take the columns that have more than 40% NA's out.

```r
# calculate the missing values proportion for each variable
na_prop <- colSums(is.na(original_dataset)) / nrow(original_dataset)
# Find the variables that have over 40% missing values
na_40 <- sort(na_prop[na_prop > 0.4], decreasing = TRUE)
# remove these columns
original_dataset <- original_dataset[ ,!names(original_dataset) %in% names(na_40)]
```

2. There are columns that we don't understanding the meaning of such as `FLAG_DOCUMENT` and `SOCIAL_CIRCLE`. Since we cannot find any additional information about them, we decided to remove these variables as well.

```r
original_dataset = original_dataset[-grep("FLAG_DOCUMENT",colnames(original_dataset))]
original_dataset = original_dataset[-grep("SOCIAL_CIRCLE",colnames(original_dataset))]
```

We also decided to remove any column that contains `CITY` in them since there are other columns that define the applicant's `REGION` and some variables that describe the characteristics of the `REGION`, using `CITY` again seems redundant and overlapping.

```
original_dataset = original_dataset[-grep("CITY", colnames(original_dataset))]
```

Because of the same reason, we decided to remove some of the columns that contain `AMT_REQ_CREDIT_BUREAU`, only keep `AMT_REQ_CREDIT_BUREAU_WEEK` represent short-term count of credit requirements and `AMT_REQ_CREDIT_BUREAU_YEAR` as long_term count of credit requirements.

```
names = c("AMT_REQ_CREDIT_BUREAU_HOUR", "AMT_REQ_CREDIT_BUREAU_DAY", "AMT_REQ_CREDIT_BUREAU_MON", "AMT_
original_dataset = original_dataset[,-which(names(original_dataset) %in% names) ]
```

3. `DAYS_EMPLOYED` represents the days that the applicant is employed until the application date, which whould be all negative in this dataset. Therefore, the value `365243` in `DAYS_EMPLOYED` column seems unreasonable and we would replace it with 0.

```
original_dataset$DAYS_EMPLOYED[which(original_dataset$DAYS_EMPLOYED == 365243)] <- 0
```

For better understanding of the data, we also need to convert `DAYS_EMPLOYED`, `DAYS_BIRTH`, `DAYS_PUBLISH` and `DAYS_REGISTRATION`, which are presented as negative in the dataset, to positive number in years.

```
original_dataset$DAYS_EMPLOYED[which(original_dataset$DAYS_EMPLOYED == 365243)] <- 0
original_dataset$DAYS_EMPLOYED = abs(original_dataset$DAYS_EMPLOYED)/365 %>% floor()
original_dataset$DAYS_BIRTH = abs(original_dataset$DAYS_BIRTH)/365 %>% floor()
original_dataset$DAYS_ID_PUBLISH = abs(original_dataset$DAYS_ID_PUBLISH)/365 %>% floor()
original_dataset$DAYS_REGISTRATION = abs(original_dataset$DAYS_REGISTRATION)/365 %>% floor()
```

4. There are some false entries in `AMT_REQ_CREDIT_BUREAU_WEEK` and `AMT_REQ_CREDIT_BUREAU_YEAR`, so we removed all observations with false entries.

```
original_dataset<-original_dataset%>% filter((is.na(AMT_REQ_CREDIT_BUREAU_WEEK)&is.na(AMT_REQ_CREDIT_BU
                              (AMT_REQ_CREDIT_BUREAU_WEEK <=AMT_REQ_CREDIT_BUREAU_YEAR))
```

Remove XNA in `CODE_GENDER`

```
original_dataset <- original_dataset %>% filter(CODE_GENDER != "XNA")
```

Set XNA in `ORGANIZATION_TYPE` to `Not_provide`

```
original_dataset[original_dataset=="XNA"] <- "Not Provided"
```

5. With columns that are left with less than 40% NA's in them, we replaced those NA's with the median of the variable.

```
ext2_median <- median(original_dataset$EXT_SOURCE_2, na.rm = TRUE)
ext3_median <- median(original_dataset$EXT_SOURCE_3, na.rm = TRUE)

original_dataset<- original_dataset%>% replace_na(list(EXT_SOURCE_2 = ext2_median,
                                  EXT_SOURCE_3 = ext3_median))

phonechange_median <- median(original_dataset$DAYS_LAST_PHONE_CHANGE, na.rm = TRUE)
original_dataset<- original_dataset%>% replace_na(list(DAYS_LAST_PHONE_CHANGE = phonechange_median))
```

```
week_median <- median(original_dataset$AMT_REQ_CREDIT_BUREAU_WEEK, na.rm = TRUE)
year_median <- median(original_dataset$AMT_REQ_CREDIT_BUREAU_YEAR, na.rm = TRUE)

original_dataset<- original_dataset%>% replace_na(list(AMT_REQ_CREDIT_BUREAU_WEEK = week_median,
                          AMT_REQ_CREDIT_BUREAU_YEAR = year_median))
```

We replaced NA in `Annuity` to 0

```
original_dataset$AMT_ANNUITY[is.na(original_dataset$AMT_ANNUITY)] <- 0
```

We replace NA in `Good Price` column to 0

```
original_dataset$AMT_GOODS_PRICE[is.na(original_dataset$AMT_GOODS_PRICE)] <- 0
```

We also removed unkown family status observations in the data.

```
unknow_status = which(is.na(original_dataset$CNT_FAM_MEMBERS))
original_dataset = original_dataset[-unknow_status,]
```

We then set other NA's as "not_provided" level

```
original_dataset[is.na(original_dataset)] <- "Not Provided"
```

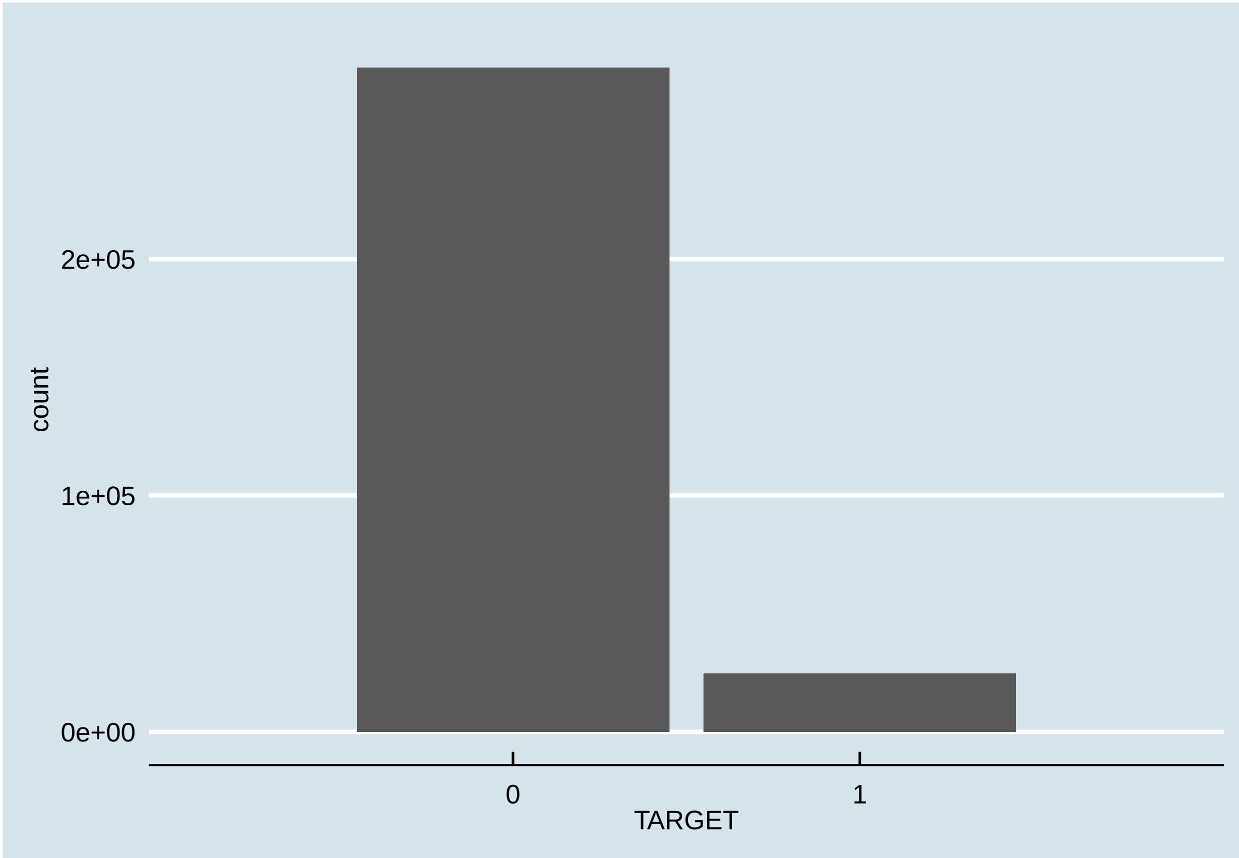
And last but not least, we factored all the columns in the dataset.

```
original_dataset <- as.data.frame(unclass(original_dataset))
```

Exploratory Data Analysis

Before we go ahead to build different models for our dataset, we need to take a look at the data that we have.

```
ggplot(original_dataset)+
  geom_bar(aes(x=TARGET,col=TARGET))+
  scale_x_discrete(limits=c(0,1))
```



From this graph we can see that the proportion of default(1) and not default(0) are highly different. Therefore, when we separate the dataset into train and test datasets, we need to make sure that the there are enough default(1) in both train and test datasets. Therefore, we would randomly select 20% from 0 and 1 as the test dataset.

```
set.seed(7)
dd_default = original_dataset %>% filter(TARGET==1)
dd_default %>%
  mutate(TRAIN = sample(c(0,1),nrow(dd_default),replace=T,prob=c(0.2,0.8))) ->dd_default

dd_not_default = original_dataset %>% filter(TARGET == 0)
dd_not_default %>%
  mutate(TRAIN = sample(c(0,1),nrow(dd_not_default),replace=T,prob=c(0.2,0.8))) ->dd_not_default

dd_clean = rbind(dd_default,dd_not_default)

application_train = dd_clean[which(dd_clean$TRAIN==1),]
application_test = dd_clean[which(dd_clean$TRAIN==0),]
```

In addition to the above dataset, we also created another dataset that has converted all the categorical variables into dummy variables in the datset. Since LASSO and Ridge would not automatically convert categorical variables, we created this dataset for LASSO and Ridge.

```
dmy <- dummyVars(formula = ~., data = application_train, fullRank = TRUE)
dummy_train <- data.frame(predict(dmy, newdata = application_train))
```

```r
dmy <- dummyVars(formula = ~., data = application_test, fullRank = TRUE)
dummy_test <- data.frame(predict(dmy, newdata = application_test))
```

In order to save time, We decided to take $\frac{1}{10}$ of `application_train` to be `subset_train`, and used it to find out the optimized forward, backoward selection and tree-based model.

```r
set.seed(7)
subset_train <- application_train[sample(1:nrow(application_train),nrow(application_train)/10),]
dummy_subset_train <- dummy_train[sample(1:nrow(application_train),nrow(application_train)/10),]
```

Linear Regression

Before we jump into Lasso and Ridge, a simple linear regression is needed for a overall understanding of the data.

```r
model_lm <- lm(TARGET~ . -SK_ID_CURR -TRAIN,data=application_train)

# Compute training MSE
yhat_lm_train <- predict(model_lm, application_train)
mse_lm_train <- mean((application_train$TARGET - yhat_lm_train)^2)

# Compute test MSE
yhat_lm_test <- predict(model_lm, application_test)
mse_lm_test <- mean((application_test$TARGET- yhat_lm_test)^2)

summary(model_lm)
```

```
##
## Call:
## lm(formula = TARGET ~ . - SK_ID_CURR - TRAIN, data = application_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4085 -0.1135 -0.0646 -0.0166  1.0895
##
## Coefficients: (1 not defined because of singularities)
##                                                     Estimate Std. Error
## (Intercept)                                        1.215e-01  2.926e-01
## NAME_CONTRACT_TYPERevolving loans                 -1.871e-02  1.980e-03
## CODE_GENDERM                                       2.583e-02  1.419e-03
## FLAG_OWN_CARY                                     -2.039e-02  1.246e-03
## FLAG_OWN_REALTYY                                   3.524e-03  1.224e-03
## CNT_CHILDREN                                       8.087e-04  8.238e-04
## AMT_INCOME_TOTAL                                   4.036e-09  2.101e-09
## AMT_CREDIT                                         1.560e-07  8.575e-09
## AMT_ANNUITY                                        6.525e-07  6.104e-08
## AMT_GOODS_PRICE                                   -1.881e-07  9.453e-09
## NAME_TYPE_SUITEFamily                             -5.258e-03  5.404e-03
## NAME_TYPE_SUITEGroup of people                    -1.022e-02  1.859e-02
## NAME_TYPE_SUITENot Provided                       -3.875e-02  9.882e-03
## NAME_TYPE_SUITEOther_A                            -7.270e-03  1.127e-02
```

```
## NAME_TYPE_SUITEOther_B                              4.391e-03  8.739e-03
## NAME_TYPE_SUITESpouse, partner                     -8.410e-03  5.911e-03
## NAME_TYPE_SUITEUnaccompanied                       -3.311e-03  5.235e-03
## NAME_INCOME_TYPECommercial associate               -1.052e-03  9.413e-02
## NAME_INCOME_TYPEMaternity leave                     3.204e-01  1.792e-01
## NAME_INCOME_TYPEPensioner                          -6.258e-02  1.372e-01
## NAME_INCOME_TYPEState servant                       1.995e-03  9.416e-02
## NAME_INCOME_TYPEStudent                            -8.751e-02  1.150e-01
## NAME_INCOME_TYPEUnemployed                          2.212e-01  1.507e-01
## NAME_INCOME_TYPEWorking                             7.016e-03  9.414e-02
## NAME_EDUCATION_TYPEHigher education                 4.034e-02  2.295e-02
## NAME_EDUCATION_TYPEIncomplete higher                4.163e-02  2.311e-02
## NAME_EDUCATION_TYPELower secondary                  6.701e-02  2.344e-02
## NAME_EDUCATION_TYPESecondary / secondary special  6.053e-02  2.294e-02
## NAME_FAMILY_STATUSMarried                          -1.238e-02  1.867e-03
## NAME_FAMILY_STATUSSeparated                        -3.154e-03  2.740e-03
## NAME_FAMILY_STATUSSingle / not married             -2.845e-03  2.238e-03
## NAME_FAMILY_STATUSWidow                            -1.065e-02  3.013e-03
## NAME_HOUSING_TYPEHouse / apartment                 -1.555e-03  8.787e-03
## NAME_HOUSING_TYPEMunicipal apartment                7.934e-03  9.217e-03
## NAME_HOUSING_TYPEOffice apartment                  -1.847e-02  1.058e-02
## NAME_HOUSING_TYPERented apartment                   6.268e-03  9.760e-03
## NAME_HOUSING_TYPEWith parents                       4.160e-03  9.108e-03
## REGION_POPULATION_RELATIVE                          1.440e-01  4.623e-02
## DAYS_BIRTH                                         -4.470e-04  7.097e-05
## DAYS_EMPLOYED                                      -1.240e-03  1.045e-04
## DAYS_REGISTRATION                                  -2.354e-04  5.984e-05
## DAYS_ID_PUBLISH                                    -1.163e-03  1.388e-04
## FLAG_MOBIL                                          8.683e-02  2.642e-01
## FLAG_EMP_PHONE                                      6.095e-02  8.868e-02
## FLAG_WORK_PHONE                                     1.342e-02  1.490e-03
## FLAG_CONT_MOBILE                                   -1.786e-02  1.238e-02
## FLAG_PHONE                                         -4.565e-03  1.268e-03
## FLAG_EMAIL                                         -6.904e-03  2.344e-03
## OCCUPATION_TYPECleaning staff                       1.419e-02  5.390e-03
## OCCUPATION_TYPECooking staff                        1.561e-02  5.034e-03
## OCCUPATION_TYPECore staff                           1.571e-03  3.668e-03
## OCCUPATION_TYPEDrivers                              1.893e-02  4.013e-03
## OCCUPATION_TYPEHigh skill tech staff                1.230e-03  4.147e-03
## OCCUPATION_TYPEHR staff                            -6.716e-03  1.287e-02
## OCCUPATION_TYPEIT staff                            -8.870e-03  1.339e-02
## OCCUPATION_TYPELaborers                             1.499e-02  3.438e-03
## OCCUPATION_TYPELow-skill Laborers                   4.675e-02  7.314e-03
## OCCUPATION_TYPEManagers                             4.623e-03  3.687e-03
## OCCUPATION_TYPEMedicine staff                       6.945e-03  5.110e-03
## OCCUPATION_TYPENot Provided                         5.274e-03  3.405e-03
## OCCUPATION_TYPEPrivate service staff               -2.429e-03  6.871e-03
## OCCUPATION_TYPERealty agents                        2.758e-03  1.156e-02
## OCCUPATION_TYPESales staff                          8.895e-03  3.570e-03
## OCCUPATION_TYPESecretaries                          1.795e-02  8.826e-03
## OCCUPATION_TYPESecurity staff                       1.772e-02  5.303e-03
## OCCUPATION_TYPEWaiters/barmen staff                 2.862e-02  8.610e-03
## CNT_FAM_MEMBERS                                            NA         NA
## REGION_RATING_CLIENT                                1.070e-02  1.327e-03
```

```
## WEEKDAY_APPR_PROCESS_STARTMONDAY                 -6.061e-03  1.864e-03
## WEEKDAY_APPR_PROCESS_STARTSATURDAY               -4.500e-03  2.088e-03
## WEEKDAY_APPR_PROCESS_STARTSUNDAY                 -4.994e-03  2.687e-03
## WEEKDAY_APPR_PROCESS_STARTTHURSDAY               -1.957e-03  1.865e-03
## WEEKDAY_APPR_PROCESS_STARTTUESDAY                 8.896e-04  1.837e-03
## WEEKDAY_APPR_PROCESS_STARTWEDNESDAY               1.482e-04  1.855e-03
## HOUR_APPR_PROCESS_START                          -2.343e-04  1.750e-04
## REG_REGION_NOT_LIVE_REGION                       -6.154e-03  6.542e-03
## REG_REGION_NOT_WORK_REGION                        4.378e-04  7.130e-03
## LIVE_REGION_NOT_WORK_REGION                      -3.662e-03  7.102e-03
## ORGANIZATION_TYPEAgriculture                     -1.651e-02  1.528e-02
## ORGANIZATION_TYPEBank                            -3.757e-02  1.525e-02
## ORGANIZATION_TYPEBusiness Entity Type 1          -2.661e-02  1.455e-02
## ORGANIZATION_TYPEBusiness Entity Type 2          -2.190e-02  1.434e-02
## ORGANIZATION_TYPEBusiness Entity Type 3          -1.516e-02  1.408e-02
## ORGANIZATION_TYPECleaning                        -5.861e-03  2.301e-02
## ORGANIZATION_TYPEConstruction                    -2.023e-04  1.450e-02
## ORGANIZATION_TYPECulture                         -2.016e-02  2.082e-02
## ORGANIZATION_TYPEElectricity                     -2.845e-02  1.699e-02
## ORGANIZATION_TYPEEmergency                       -2.709e-02  1.885e-02
## ORGANIZATION_TYPEGovernment                      -2.336e-02  1.434e-02
## ORGANIZATION_TYPEHotel                           -3.454e-02  1.690e-02
## ORGANIZATION_TYPEHousing                         -2.490e-02  1.507e-02
## ORGANIZATION_TYPEIndustry: type 1                -4.281e-03  1.677e-02
## ORGANIZATION_TYPEIndustry: type 10               -4.643e-02  3.206e-02
## ORGANIZATION_TYPEIndustry: type 11               -2.473e-02  1.516e-02
## ORGANIZATION_TYPEIndustry: type 12               -5.107e-02  2.063e-02
## ORGANIZATION_TYPEIndustry: type 13               -2.773e-02  4.064e-02
## ORGANIZATION_TYPEIndustry: type 2                -3.457e-02  1.978e-02
## ORGANIZATION_TYPEIndustry: type 3                -7.053e-03  1.498e-02
## ORGANIZATION_TYPEIndustry: type 4                -1.928e-02  1.721e-02
## ORGANIZATION_TYPEIndustry: type 5                -4.105e-02  1.863e-02
## ORGANIZATION_TYPEIndustry: type 6                -3.204e-02  3.133e-02
## ORGANIZATION_TYPEIndustry: type 7                -2.340e-02  1.624e-02
## ORGANIZATION_TYPEIndustry: type 8                 5.999e-02  6.383e-02
## ORGANIZATION_TYPEIndustry: type 9                -3.981e-02  1.496e-02
## ORGANIZATION_TYPEInsurance                       -2.231e-02  1.865e-02
## ORGANIZATION_TYPEKindergarten                    -2.524e-02  1.452e-02
## ORGANIZATION_TYPELegal Services                   1.179e-02  2.210e-02
## ORGANIZATION_TYPEMedicine                        -2.468e-02  1.449e-02
## ORGANIZATION_TYPEMilitary                        -4.905e-02  1.524e-02
## ORGANIZATION_TYPEMobile                          -2.275e-02  2.206e-02
## ORGANIZATION_TYPENot Provided                     9.562e-02  1.343e-01
## ORGANIZATION_TYPEOther                           -2.038e-02  1.423e-02
## ORGANIZATION_TYPEPolice                          -3.963e-02  1.540e-02
## ORGANIZATION_TYPEPostal                          -1.349e-02  1.545e-02
## ORGANIZATION_TYPERealtor                          1.882e-02  2.071e-02
## ORGANIZATION_TYPEReligion                        -1.683e-02  3.377e-02
## ORGANIZATION_TYPERestaurant                      -7.320e-03  1.576e-02
## ORGANIZATION_TYPESchool                          -2.401e-02  1.441e-02
## ORGANIZATION_TYPESecurity                        -2.646e-02  1.528e-02
## ORGANIZATION_TYPESecurity Ministries             -4.017e-02  1.560e-02
## ORGANIZATION_TYPESelf-employed                   -8.952e-03  1.413e-02
## ORGANIZATION_TYPEServices                        -1.866e-02  1.617e-02
```

```
## ORGANIZATION_TYPETelecom                          -1.220e-02  1.866e-02
## ORGANIZATION_TYPETrade: type 1                    -2.282e-02  2.137e-02
## ORGANIZATION_TYPETrade: type 2                    -5.442e-02  1.561e-02
## ORGANIZATION_TYPETrade: type 3                    -1.407e-02  1.492e-02
## ORGANIZATION_TYPETrade: type 4                    -5.450e-02  4.101e-02
## ORGANIZATION_TYPETrade: type 5                    -9.530e-02  4.510e-02
## ORGANIZATION_TYPETrade: type 6                    -3.458e-02  1.835e-02
## ORGANIZATION_TYPETrade: type 7                    -1.357e-02  1.444e-02
## ORGANIZATION_TYPETransport: type 1                -5.088e-02  2.527e-02
## ORGANIZATION_TYPETransport: type 2                -2.481e-02  1.541e-02
## ORGANIZATION_TYPETransport: type 3                 3.067e-02  1.653e-02
## ORGANIZATION_TYPETransport: type 4                -1.928e-02  1.461e-02
## ORGANIZATION_TYPEUniversity                       -2.430e-02  1.625e-02
## EXT_SOURCE_2                                      -1.740e-01  3.072e-03
## EXT_SOURCE_3                                      -2.053e-01  3.166e-03
## DAYS_LAST_PHONE_CHANGE                             4.334e-06  6.778e-07
## AMT_REQ_CREDIT_BUREAU_WEEK                        -6.073e-03  3.534e-03
## AMT_REQ_CREDIT_BUREAU_YEAR                         1.704e-04  3.114e-04
##                                                   t value Pr(>|t|)
## (Intercept)                                         0.415 0.678017
## NAME_CONTRACT_TYPERevolving loans                  -9.450  < 2e-16 ***
## CODE_GENDERM                                       18.202  < 2e-16 ***
## FLAG_OWN_CARY                                     -16.364  < 2e-16 ***
## FLAG_OWN_REALTYY                                    2.879 0.003993 **
## CNT_CHILDREN                                        0.982 0.326222
## AMT_INCOME_TOTAL                                    1.921 0.054686 .
## AMT_CREDIT                                         18.187  < 2e-16 ***
## AMT_ANNUITY                                        10.690  < 2e-16 ***
## AMT_GOODS_PRICE                                   -19.900  < 2e-16 ***
## NAME_TYPE_SUITEFamily                              -0.973 0.330602
## NAME_TYPE_SUITEGroup of people                     -0.550 0.582404
## NAME_TYPE_SUITENot Provided                        -3.921 8.81e-05 ***
## NAME_TYPE_SUITEOther_A                             -0.645 0.518826
## NAME_TYPE_SUITEOther_B                              0.502 0.615363
## NAME_TYPE_SUITESpouse, partner                     -1.423 0.154839
## NAME_TYPE_SUITEUnaccompanied                       -0.632 0.527102
## NAME_INCOME_TYPECommercial associate               -0.011 0.991085
## NAME_INCOME_TYPEMaternity leave                     1.788 0.073828 .
## NAME_INCOME_TYPEPensioner                          -0.456 0.648374
## NAME_INCOME_TYPEState servant                       0.021 0.983096
## NAME_INCOME_TYPEStudent                            -0.761 0.446691
## NAME_INCOME_TYPEUnemployed                          1.468 0.142142
## NAME_INCOME_TYPEWorking                             0.075 0.940585
## NAME_EDUCATION_TYPEHigher education                 1.758 0.078720 .
## NAME_EDUCATION_TYPEIncomplete higher                1.801 0.071693 .
## NAME_EDUCATION_TYPELower secondary                  2.859 0.004252 **
## NAME_EDUCATION_TYPESecondary / secondary special   2.639 0.008324 **
## NAME_FAMILY_STATUSMarried                          -6.633 3.31e-11 ***
## NAME_FAMILY_STATUSSeparated                        -1.151 0.249685
## NAME_FAMILY_STATUSSingle / not married             -1.271 0.203648
## NAME_FAMILY_STATUSWidow                            -3.535 0.000409 ***
## NAME_HOUSING_TYPEHouse / apartment                 -0.177 0.859543
## NAME_HOUSING_TYPEMunicipal apartment                0.861 0.389328
## NAME_HOUSING_TYPEOffice apartment                  -1.745 0.080909 .
```

```
## NAME_HOUSING_TYPERented apartment                  0.642 0.520755
## NAME_HOUSING_TYPEWith parents                      0.457 0.647835
## REGION_POPULATION_RELATIVE                         3.116 0.001836 **
## DAYS_BIRTH                                        -6.299 3.00e-10 ***
## DAYS_EMPLOYED                                    -11.867  < 2e-16 ***
## DAYS_REGISTRATION                                 -3.934 8.36e-05 ***
## DAYS_ID_PUBLISH                                   -8.384  < 2e-16 ***
## FLAG_MOBIL                                         0.329 0.742447
## FLAG_EMP_PHONE                                     0.687 0.491870
## FLAG_WORK_PHONE                                    9.003  < 2e-16 ***
## FLAG_CONT_MOBILE                                  -1.443 0.149127
## FLAG_PHONE                                        -3.599 0.000320 ***
## FLAG_EMAIL                                        -2.945 0.003229 **
## OCCUPATION_TYPECleaning staff                      2.632 0.008485 **
## OCCUPATION_TYPECooking staff                       3.102 0.001922 **
## OCCUPATION_TYPECore staff                          0.428 0.668458
## OCCUPATION_TYPEDrivers                             4.716 2.41e-06 ***
## OCCUPATION_TYPEHigh skill tech staff               0.297 0.766786
## OCCUPATION_TYPEHR staff                           -0.522 0.601838
## OCCUPATION_TYPEIT staff                           -0.663 0.507593
## OCCUPATION_TYPELaborers                            4.361 1.29e-05 ***
## OCCUPATION_TYPELow-skill Laborers                  6.391 1.65e-10 ***
## OCCUPATION_TYPEManagers                            1.254 0.209976
## OCCUPATION_TYPEMedicine staff                      1.359 0.174133
## OCCUPATION_TYPENot Provided                        1.549 0.121459
## OCCUPATION_TYPEPrivate service staff              -0.353 0.723734
## OCCUPATION_TYPERealty agents                       0.239 0.811463
## OCCUPATION_TYPESales staff                         2.492 0.012715 *
## OCCUPATION_TYPESecretaries                         2.034 0.041974 *
## OCCUPATION_TYPESecurity staff                      3.342 0.000831 ***
## OCCUPATION_TYPEWaiters/barmen staff                3.324 0.000887 ***
## CNT_FAM_MEMBERS                                       NA       NA
## REGION_RATING_CLIENT                               8.059 7.71e-16 ***
## WEEKDAY_APPR_PROCESS_STARTMONDAY                  -3.251 0.001149 **
## WEEKDAY_APPR_PROCESS_STARTSATURDAY                -2.156 0.031123 *
## WEEKDAY_APPR_PROCESS_STARTSUNDAY                  -1.858 0.063148 .
## WEEKDAY_APPR_PROCESS_STARTTHURSDAY                -1.049 0.294155
## WEEKDAY_APPR_PROCESS_STARTTUESDAY                  0.484 0.628246
## WEEKDAY_APPR_PROCESS_STARTWEDNESDAY                0.080 0.936342
## HOUR_APPR_PROCESS_START                           -1.339 0.180667
## REG_REGION_NOT_LIVE_REGION                        -0.941 0.346858
## REG_REGION_NOT_WORK_REGION                         0.061 0.951037
## LIVE_REGION_NOT_WORK_REGION                       -0.516 0.606043
## ORGANIZATION_TYPEAgriculture                      -1.080 0.279976
## ORGANIZATION_TYPEBank                             -2.464 0.013754 *
## ORGANIZATION_TYPEBusiness Entity Type 1           -1.829 0.067456 .
## ORGANIZATION_TYPEBusiness Entity Type 2           -1.527 0.126727
## ORGANIZATION_TYPEBusiness Entity Type 3           -1.077 0.281533
## ORGANIZATION_TYPECleaning                         -0.255 0.798964
## ORGANIZATION_TYPEConstruction                     -0.014 0.988867
## ORGANIZATION_TYPECulture                          -0.968 0.332998
## ORGANIZATION_TYPEElectricity                      -1.674 0.094085 .
## ORGANIZATION_TYPEEmergency                        -1.437 0.150665
## ORGANIZATION_TYPEGovernment                       -1.629 0.103344
```

```
## ORGANIZATION_TYPEHotel                            -2.043 0.041053 *
## ORGANIZATION_TYPEHousing                          -1.653 0.098397 .
## ORGANIZATION_TYPEIndustry: type 1                 -0.255 0.798497
## ORGANIZATION_TYPEIndustry: type 10                -1.448 0.147597
## ORGANIZATION_TYPEIndustry: type 11                -1.631 0.102831
## ORGANIZATION_TYPEIndustry: type 12                -2.475 0.013307 *
## ORGANIZATION_TYPEIndustry: type 13                -0.682 0.495066
## ORGANIZATION_TYPEIndustry: type 2                 -1.748 0.080470 .
## ORGANIZATION_TYPEIndustry: type 3                 -0.471 0.637651
## ORGANIZATION_TYPEIndustry: type 4                 -1.120 0.262635
## ORGANIZATION_TYPEIndustry: type 5                 -2.203 0.027599 *
## ORGANIZATION_TYPEIndustry: type 6                 -1.023 0.306366
## ORGANIZATION_TYPEIndustry: type 7                 -1.441 0.149450
## ORGANIZATION_TYPEIndustry: type 8                  0.940 0.347341
## ORGANIZATION_TYPEIndustry: type 9                 -2.662 0.007778 **
## ORGANIZATION_TYPEInsurance                        -1.196 0.231781
## ORGANIZATION_TYPEKindergarten                     -1.739 0.082082 .
## ORGANIZATION_TYPELegal Services                    0.534 0.593625
## ORGANIZATION_TYPEMedicine                         -1.703 0.088589 .
## ORGANIZATION_TYPEMilitary                         -3.219 0.001286 **
## ORGANIZATION_TYPEMobile                           -1.031 0.302466
## ORGANIZATION_TYPENot Provided                      0.712 0.476535
## ORGANIZATION_TYPEOther                            -1.432 0.152022
## ORGANIZATION_TYPEPolice                           -2.573 0.010070 *
## ORGANIZATION_TYPEPostal                           -0.873 0.382439
## ORGANIZATION_TYPERealtor                           0.909 0.363608
## ORGANIZATION_TYPEReligion                         -0.498 0.618220
## ORGANIZATION_TYPERestaurant                       -0.464 0.642399
## ORGANIZATION_TYPESchool                           -1.666 0.095733 .
## ORGANIZATION_TYPESecurity                         -1.731 0.083398 .
## ORGANIZATION_TYPESecurity Ministries              -2.574 0.010041 *
## ORGANIZATION_TYPESelf-employed                    -0.634 0.526272
## ORGANIZATION_TYPEServices                         -1.154 0.248402
## ORGANIZATION_TYPETelecom                          -0.654 0.513295
## ORGANIZATION_TYPETrade: type 1                    -1.068 0.285696
## ORGANIZATION_TYPETrade: type 2                    -3.487 0.000489 ***
## ORGANIZATION_TYPETrade: type 3                    -0.943 0.345723
## ORGANIZATION_TYPETrade: type 4                    -1.329 0.183858
## ORGANIZATION_TYPETrade: type 5                    -2.113 0.034587 *
## ORGANIZATION_TYPETrade: type 6                    -1.885 0.059464 .
## ORGANIZATION_TYPETrade: type 7                    -0.940 0.347319
## ORGANIZATION_TYPETransport: type 1                -2.013 0.044069 *
## ORGANIZATION_TYPETransport: type 2                -1.609 0.107535
## ORGANIZATION_TYPETransport: type 3                 1.855 0.063616 .
## ORGANIZATION_TYPETransport: type 4                -1.320 0.186855
## ORGANIZATION_TYPEUniversity                       -1.495 0.134888
## EXT_SOURCE_2                                     -56.619  < 2e-16 ***
## EXT_SOURCE_3                                     -64.846  < 2e-16 ***
## DAYS_LAST_PHONE_CHANGE                             6.395 1.61e-10 ***
## AMT_REQ_CREDIT_BUREAU_WEEK                        -1.719 0.085678 .
## AMT_REQ_CREDIT_BUREAU_YEAR                         0.547 0.584152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2641 on 244336 degrees of freedom
## Multiple R-squared:  0.06098,    Adjusted R-squared:  0.06045
## F-statistic:   115 on 138 and 244336 DF,  p-value: < 2.2e-16
```

```r
print(paste("MSE of training dataset is", signif(mse_lm_train,4 )))
```

```
## [1] "MSE of training dataset is 0.06973"
```

```r
print(paste("MSE of testing dataset is", signif(mse_lm_test,4 )))
```

```
## [1] "MSE of testing dataset is 0.06983"
```

We select out the top 10 predictors both negative or positive affect the default probability.

```r
topcof <- sort(model_lm$coefficients, decreasing = TRUE)
topcof[1:10]
```

```
##                  NAME_INCOME_TYPEMaternity leave
##                                       0.32041182
##                       NAME_INCOME_TYPEUnemployed
##                                       0.22121048
##                       REGION_POPULATION_RELATIVE
##                                       0.14402552
##                                      (Intercept)
##                                       0.12147195
##                    ORGANIZATION_TYPENot Provided
##                                       0.09561599
##                                       FLAG_MOBIL
##                                       0.08682527
##               NAME_EDUCATION_TYPELower secondary
##                                       0.06701318
##                                   FLAG_EMP_PHONE
##                                       0.06095192
## NAME_EDUCATION_TYPESecondary / secondary special
##                                       0.06053418
##                 ORGANIZATION_TYPEIndustry: type 8
##                                       0.05998960
```

```r
leastcof <- sort(model_lm$coefficients)
leastcof[1:10]
```

```
##                       EXT_SOURCE_3                        EXT_SOURCE_2
##                        -0.20532510                         -0.17395978
##     ORGANIZATION_TYPETrade: type 5             NAME_INCOME_TYPEStudent
##                        -0.09529889                         -0.08751180
##          NAME_INCOME_TYPEPensioner      ORGANIZATION_TYPETrade: type 4
##                        -0.06258153                         -0.05450202
##     ORGANIZATION_TYPETrade: type 2 ORGANIZATION_TYPEIndustry: type 12
##                        -0.05442453                         -0.05106923
## ORGANIZATION_TYPETransport: type 1          ORGANIZATION_TYPEMilitary
##                        -0.05088271                         -0.04905336
```

Lasso & Ridge

```
c_names <- colnames(dummy_train)
c_names <- c_names[!c_names %in% c("SK_ID_CURR", "TARGET")]

loopformula <- "TARGET ~ NAME_CONTRACT_TYPE.Revolving.loans"

for (name in c_names[2:length(c_names)]) {
  loopformula <- paste0(loopformula, "+", name, sep = "")
}

f_all <- as.formula(loopformula)
```

Set x_test, x_train, y_test, x_train

```
x1_train <- model.matrix(f_all, dummy_train)[ , -1]
y1_train <- dummy_train$TARGET

x1_test <- model.matrix(f_all, dummy_test)[ ,-1]
y1_test <- dummy_test$TARGET
```

```
## run lasso regression
fit_lasso <- cv.glmnet(x1_train, y1_train, alpha = 1, nfolds = 10)

# compute MSE train
yhat_lasso_train <- predict(fit_lasso, x1_train, s = fit_lasso$lambda.min)
mse_lasso_train <- mean((y1_train - yhat_lasso_train)^2)

# compute MSE test
yhat_lasso_test <- predict(fit_lasso, x1_test, s = fit_lasso$lambda.min)
mse_lasso_test <- mean((y1_test - yhat_lasso_test)^2)

#find out the variables with values after lasso regression
temp <- coef(fit_lasso)
temp2 <- coef(fit_lasso)
temp2 <- as.data.frame(summary(temp2))
cbind ( as.vector(temp@Dimnames[[1]]) [temp2$i], temp2$x)
```

```
##       [,1]
##  [1,] "(Intercept)"
##  [2,] "NAME_CONTRACT_TYPE.Revolving.loans"
##  [3,] "CODE_GENDER.M"
##  [4,] "FLAG_OWN_CAR.Y"
##  [5,] "NAME_INCOME_TYPE.Pensioner"
##  [6,] "NAME_INCOME_TYPE.Working"
##  [7,] "NAME_EDUCATION_TYPE.Higher.education"
##  [8,] "NAME_EDUCATION_TYPE.Secondary...secondary.special"
##  [9,] "NAME_FAMILY_STATUS.Married"
## [10,] "DAYS_BIRTH"
## [11,] "DAYS_EMPLOYED"
## [12,] "DAYS_ID_PUBLISH"
## [13,] "OCCUPATION_TYPE.Drivers"
```

```
## [14,] "OCCUPATION_TYPE.Laborers"
## [15,] "OCCUPATION_TYPE.Low.skill.Laborers"
## [16,] "OCCUPATION_TYPE.Not.Provided"
## [17,] "REGION_RATING_CLIENT"
## [18,] "ORGANIZATION_TYPE.Self.employed"
## [19,] "EXT_SOURCE_2"
## [20,] "EXT_SOURCE_3"
## [21,] "DAYS_LAST_PHONE_CHANGE"
##       [,2]
##  [1,] "0.288724079096263"
##  [2,] "-0.0140550841700229"
##  [3,] "0.0191084914810265"
##  [4,] "-0.00770509844951862"
##  [5,] "-0.00324257302957579"
##  [6,] "0.00738896989120393"
##  [7,] "-0.0110803055887317"
##  [8,] "0.00674652560949894"
##  [9,] "-0.00176886762709733"
## [10,] "-0.000494752528821249"
## [11,] "-0.000678001319344054"
## [12,] "-0.000586731527513708"
## [13,] "0.000762263483909344"
## [14,] "0.00204599445986182"
## [15,] "0.000978572403982065"
## [16,] "-0.000125528569967539"
## [17,] "0.00163144980343066"
## [18,] "0.00280220667008627"
## [19,] "-0.174085342822593"
## [20,] "-0.190254109184574"
## [21,] "9.67575690940942e-07"
```

```r
## run ridge regression
fit_Ridge <- cv.glmnet(x1_train, y1_train, alpha = 0, nfolds = 10)

# compute MSE train
yhat_Ridge_train <- predict(fit_Ridge, x1_train, s = fit_Ridge$lambda.min)
mse_Ridge_train <- mean((y1_train - yhat_Ridge_train)^2)

# compute MSE test
yhat_Ridge_test <- predict(fit_Ridge, x1_test, s = fit_Ridge$lambda.min)
mse_Ridge_test <- mean((y1_test - yhat_Ridge_test)^2)

#output the coefficients of ridge regression
coef(fit_Ridge)
```

```
## 141 x 1 sparse Matrix of class "dgCMatrix"
##                                                      1
## (Intercept)                               1.780161e-01
## NAME_CONTRACT_TYPE.Revolving.loans       -1.808433e-02
## CODE_GENDER.M                             1.419388e-02
## FLAG_OWN_CAR.Y                           -1.038149e-02
## FLAG_OWN_REALTY.Y                         1.701225e-03
## CNT_CHILDREN                              9.241472e-04
## AMT_INCOME_TOTAL                          2.559212e-09
```

```
## AMT_CREDIT                                        -4.438328e-10
## AMT_ANNUITY                                        1.590475e-07
## AMT_GOODS_PRICE                                   -7.926554e-09
## NAME_TYPE_SUITE.Family                            -1.277680e-03
## NAME_TYPE_SUITE.Group.of.people                   -2.587620e-03
## NAME_TYPE_SUITE.Not.Provided                      -1.481284e-02
## NAME_TYPE_SUITE.Other_A                           -6.816076e-05
## NAME_TYPE_SUITE.Other_B                            6.631568e-03
## NAME_TYPE_SUITE.Spouse..partner                   -2.073149e-03
## NAME_TYPE_SUITE.Unaccompanied                      1.246239e-03
## NAME_INCOME_TYPE.Commercial.associate             -2.298222e-03
## NAME_INCOME_TYPE.Maternity.leave                   1.805031e-01
## NAME_INCOME_TYPE.Pensioner                        -4.108065e-03
## NAME_INCOME_TYPE.State.servant                    -3.655299e-03
## NAME_INCOME_TYPE.Student                          -5.326077e-02
## NAME_INCOME_TYPE.Unemployed                        1.601872e-01
## NAME_INCOME_TYPE.Working                           5.000788e-03
## NAME_EDUCATION_TYPE.Higher.education              -1.006264e-02
## NAME_EDUCATION_TYPE.Incomplete.higher             -3.976496e-03
## NAME_EDUCATION_TYPE.Lower.secondary                1.333357e-02
## NAME_EDUCATION_TYPE.Secondary...secondary.special  9.050814e-03
## NAME_FAMILY_STATUS.Married                        -6.006931e-03
## NAME_FAMILY_STATUS.Separated                       8.388648e-04
## NAME_FAMILY_STATUS.Single...not.married            3.633268e-03
## NAME_FAMILY_STATUS.Widow                          -4.744421e-03
## NAME_HOUSING_TYPE.House...apartment               -3.701050e-03
## NAME_HOUSING_TYPE.Municipal.apartment              3.966633e-03
## NAME_HOUSING_TYPE.Office.apartment                -1.299485e-02
## NAME_HOUSING_TYPE.Rented.apartment                 6.220613e-03
## NAME_HOUSING_TYPE.With.parents                     5.434189e-03
## REGION_POPULATION_RELATIVE                        -3.702326e-02
## DAYS_BIRTH                                        -4.239185e-04
## DAYS_EMPLOYED                                     -9.202727e-04
## DAYS_REGISTRATION                                 -2.652099e-04
## DAYS_ID_PUBLISH                                   -9.912208e-04
## FLAG_MOBIL                                         5.561171e-02
## FLAG_EMP_PHONE                                     4.078703e-03
## FLAG_WORK_PHONE                                    6.733474e-03
## FLAG_CONT_MOBILE                                  -7.054702e-03
## FLAG_PHONE                                        -4.069566e-03
## FLAG_EMAIL                                        -2.657803e-03
## OCCUPATION_TYPE.Cleaning.staff                     4.433214e-03
## OCCUPATION_TYPE.Cooking.staff                      6.715919e-03
## OCCUPATION_TYPE.Core.staff                        -5.856410e-03
## OCCUPATION_TYPE.Drivers                            1.071898e-02
## OCCUPATION_TYPE.High.skill.tech.staff             -6.519871e-03
## OCCUPATION_TYPE.HR.staff                          -1.180548e-02
## OCCUPATION_TYPE.IT.staff                          -1.054472e-02
## OCCUPATION_TYPE.Laborers                           7.300892e-03
## OCCUPATION_TYPE.Low.skill.Laborers                 3.244601e-02
## OCCUPATION_TYPE.Managers                          -3.983101e-03
## OCCUPATION_TYPE.Medicine.staff                    -2.900269e-03
## OCCUPATION_TYPE.Not.Provided                      -2.873919e-03
## OCCUPATION_TYPE.Private.service.staff             -7.433499e-03
```

```
## OCCUPATION_TYPE.Realty.agents                     -1.865841e-03
## OCCUPATION_TYPE.Sales.staff                        1.937240e-03
## OCCUPATION_TYPE.Secretaries                        3.742092e-03
## OCCUPATION_TYPE.Security.staff                     7.621442e-03
## OCCUPATION_TYPE.Waiters.barmen.staff               1.435788e-02
## CNT_FAM_MEMBERS                                    2.579486e-04
## REGION_RATING_CLIENT                               8.796142e-03
## WEEKDAY_APPR_PROCESS_START.MONDAY                 -2.827456e-03
## WEEKDAY_APPR_PROCESS_START.SATURDAY               -1.950784e-03
## WEEKDAY_APPR_PROCESS_START.SUNDAY                 -2.053290e-03
## WEEKDAY_APPR_PROCESS_START.THURSDAY               -1.111987e-04
## WEEKDAY_APPR_PROCESS_START.TUESDAY                 1.612295e-03
## WEEKDAY_APPR_PROCESS_START.WEDNESDAY               1.032917e-03
## HOUR_APPR_PROCESS_START                           -5.099896e-04
## REG_REGION_NOT_LIVE_REGION                        -1.027382e-03
## REG_REGION_NOT_WORK_REGION                         2.561968e-04
## LIVE_REGION_NOT_WORK_REGION                       -2.347886e-04
## ORGANIZATION_TYPE.Agriculture                      3.954326e-03
## ORGANIZATION_TYPE.Bank                            -1.467827e-02
## ORGANIZATION_TYPE.Business.Entity.Type.1          -3.664011e-03
## ORGANIZATION_TYPE.Business.Entity.Type.2          -1.612245e-03
## ORGANIZATION_TYPE.Business.Entity.Type.3           3.612358e-03
## ORGANIZATION_TYPE.Cleaning                         9.515487e-03
## ORGANIZATION_TYPE.Construction                     1.404942e-02
## ORGANIZATION_TYPE.Culture                         -4.818801e-03
## ORGANIZATION_TYPE.Electricity                     -7.184950e-03
## ORGANIZATION_TYPE.Emergency                       -4.730684e-03
## ORGANIZATION_TYPE.Government                      -3.322578e-03
## ORGANIZATION_TYPE.Hotel                           -9.445274e-03
## ORGANIZATION_TYPE.Housing                         -3.990723e-03
## ORGANIZATION_TYPE.Industry..type.1                 1.117403e-02
## ORGANIZATION_TYPE.Industry..type.10               -1.862045e-02
## ORGANIZATION_TYPE.Industry..type.11               -2.971521e-03
## ORGANIZATION_TYPE.Industry..type.12               -2.185123e-02
## ORGANIZATION_TYPE.Industry..type.13                5.346823e-03
## ORGANIZATION_TYPE.Industry..type.2                -8.091275e-03
## ORGANIZATION_TYPE.Industry..type.3                 8.663727e-03
## ORGANIZATION_TYPE.Industry..type.4                 2.569753e-03
## ORGANIZATION_TYPE.Industry..type.5                -1.146545e-02
## ORGANIZATION_TYPE.Industry..type.6                -8.498841e-03
## ORGANIZATION_TYPE.Industry..type.7                -1.850135e-03
## ORGANIZATION_TYPE.Industry..type.8                 4.543858e-02
## ORGANIZATION_TYPE.Industry..type.9                -1.304792e-02
## ORGANIZATION_TYPE.Insurance                       -5.841556e-03
## ORGANIZATION_TYPE.Kindergarten                    -3.938307e-03
## ORGANIZATION_TYPE.Legal.Services                   1.409998e-02
## ORGANIZATION_TYPE.Medicine                        -4.423616e-03
## ORGANIZATION_TYPE.Military                        -1.667574e-02
## ORGANIZATION_TYPE.Mobile                          -2.261618e-03
## ORGANIZATION_TYPE.Not.Provided                    -4.038548e-03
## ORGANIZATION_TYPE.Other                           -1.312812e-03
## ORGANIZATION_TYPE.Police                          -1.293786e-02
## ORGANIZATION_TYPE.Postal                           3.471622e-03
## ORGANIZATION_TYPE.Realtor                          1.886949e-02
```

```
## ORGANIZATION_TYPE.Religion                          -2.823543e-03
## ORGANIZATION_TYPE.Restaurant                         1.037058e-02
## ORGANIZATION_TYPE.School                            -4.006558e-03
## ORGANIZATION_TYPE.Security                          -2.753282e-04
## ORGANIZATION_TYPE.Security.Ministries               -1.325527e-02
## ORGANIZATION_TYPE.Self.employed                      6.863255e-03
## ORGANIZATION_TYPE.Services                          -3.116234e-03
## ORGANIZATION_TYPE.Telecom                            3.546755e-03
## ORGANIZATION_TYPE.Trade..type.1                     -7.012512e-04
## ORGANIZATION_TYPE.Trade..type.2                     -1.959838e-02
## ORGANIZATION_TYPE.Trade..type.3                      4.587065e-03
## ORGANIZATION_TYPE.Trade..type.4                     -2.147321e-02
## ORGANIZATION_TYPE.Trade..type.5                     -4.611936e-02
## ORGANIZATION_TYPE.Trade..type.6                     -1.266336e-02
## ORGANIZATION_TYPE.Trade..type.7                      3.319974e-03
## ORGANIZATION_TYPE.Transport..type.1                 -1.918728e-02
## ORGANIZATION_TYPE.Transport..type.2                 -3.959918e-03
## ORGANIZATION_TYPE.Transport..type.3                  3.229248e-02
## ORGANIZATION_TYPE.Transport..type.4                  1.171884e-03
## ORGANIZATION_TYPE.University                        -6.268338e-03
## EXT_SOURCE_2                                        -1.099722e-01
## EXT_SOURCE_3                                        -1.281947e-01
## DAYS_LAST_PHONE_CHANGE                               5.314895e-06
## AMT_REQ_CREDIT_BUREAU_WEEK                          -2.534715e-03
## AMT_REQ_CREDIT_BUREAU_YEAR                           6.034002e-04
## TRAIN                                                .
```

Forward Selection

After the lasso and ridge regression, we also want to see the best predictors through forward and backward selection. First, we would start with the simplest model, which only contains the intercept.

```r
null <- lm(TARGET ~ 1, data = dummy_subset_train)
full <- lm(TARGET ~ . -SK_ID_CURR -TRAIN, data = dummy_subset_train)

forward.lm <- step(null, scope=list(lower=null, upper=full),
                   direction="forward")

summary(forward.lm)
#In order to save time and notebook sapce and make the outcome more clear, We didn't run the code again
```

Call: lm(formula = TARGET ~ EXT_SOURCE_2 + EXT_SOURCE_3 + CODE_GENDER.M + NAME_EDUCATION_TYPE.Higher.education + DAYS_BIRTH + FLAG_OWN_CAR.Y + NAME_CONTRACT_TYPE.Revolving.loans + NAME_INCOME_TYPE.Working + DAYS_EMPLOYED + DAYS_ID_PUBLISH + OCCUPATION_TYPE.High.skill.tech.staff + OCCUPATION_TYPE.Low.skill.Laborers + FLAG_WORK_PHONE + NAME_INCOME_TYPE.Commercial.associate + REGION_RATING_CLIENT + ORGANIZATION_TYPE.Construction + NAME_EDUCATION_TYPE.Incomplete.higher + NAME_HOUSING_TYPE.With.parents + WEEKDAY_APPR_PROCESS_START.SUNDAY + NAME_TYPE_SUITE.Unaccompanied + AMT_ANNUITY + AMT_GOODS_PRICE + AMT_CREDIT + ORGANIZATION_TYPE.Realtor + AMT_REQ_CREDIT_BUREAU_WEEK + WEEKDAY_APPR_PROCESS_STA + ORGANIZATION_TYPE.Industry..type.13 + OCCUPATION_TYPE.Cooking.staff + NAME_TYPE_SUITE.Other_B + ORGANIZATION_TYPE.Mobile + ORGANIZATION_TYPE.School + FLAG_PHONE + ORGANIZATION_TYPE.Security + ORGANIZATION_TYPE.Transport..type.3 + ORGANIZATION_TYPE.Bank

+ DAYS_LAST_PHONE_CHANGE + ORGANIZATION_TYPE.Housing + ORGANIZATION_TYPE.Emergency + ORGANIZATION_TYPE.Industry..type.7 + LIVE_REGION_NOT_WORK_REGION + OCCUPATION_TYPE.Laborers + ORGANIZATION_TYPE.Cleaning + ORGANIZATION_TYPE.Transport..type.2 + NAME_FAMILY_STATUS.Single. . . not.married, data = dummy_subset_train)

Residuals: Min 1Q Median 3Q Max -0.41207 -0.11806 -0.06565 -0.01302 1.08794

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.791e-01 1.600e-02 17.448 < 2e-16 ***EXT_SOURCE_2 -1.848e-01 9.779e-03 -18.897 < 2e-16*** EXT_SOURCE_3 -2.059e-01 1.003e-02 -20.541 < 2e-16 ***CODE_GENDER.M 3.109e-02 4.094e-03 7.595 3.19e-14*** NAME_EDUCATION_TYPE.Higher.education -2.650e-02 4.270e-03 -6.205 5.56e-10 ***DAYS_BIRTH -6.135e-04 1.851e-04 -3.314 0.000922*** FLAG_OWN_CAR.Y -2.468e-02 3.953e-03 -6.243 4.37e-10 ***NAME_CONTRACT_TYPE.Revolving.loans -1.998e-02 6.181e-03 -3.233 0.001226*** *NAME_INCOME_TYPE.Working 2.059e-02 5.240e-03 3.930 8.52e-05* **DAYS_EMPLOYED -1.425e-03 2.922e-04 -4.876 1.09e-06** *DAYS_ID_PUBLISH -1.386e-03 4.348e-04 -3.187 0.001440* ***OCCUPATION_TYPE.High.skill.tech.staff -2.981e-02 9.039e-03 -3.299 0.000973*** OCCUPATION_TYPE.Low.skill.Laborers 7.073e-02 2.117e-02 3.341 0.000836 ***FLAG_WORK_PHONE 2.222e-02 4.678e-03 4.750 2.04e-06*** NAME_INCOME_TYPE.Commercial.associate 1.280e-02 5.837e-03 2.194 0.028257 *
REGION_RATING_CLIENT 1.178e-02 3.629e-03 3.247 0.001169 ** ORGANIZATION_TYPE.Construction 2.693e-02 1.146e-02 2.349 0.018819 *
NAME_EDUCATION_TYPE.Incomplete.higher -2.338e-02 9.817e-03 -2.381 0.017262 *
NAME_HOUSING_TYPE.With.parents 2.007e-02 8.089e-03 2.481 0.013120 *
WEEKDAY_APPR_PROCESS_START.SUNDAY -1.991e-02 7.738e-03 -2.573 0.010096 *
NAME_TYPE_SUITE.Unaccompanied 1.260e-02 4.482e-03 2.812 0.004928 ** AMT_ANNUITY 7.920e-07 1.927e-07 4.111 3.96e-05 ***AMT_GOODS_PRICE -2.060e-07 3.008e-08 -6.849 7.60e-12***
AMT_CREDIT 1.710e-07 2.731e-08 6.261 3.90e-10 ** *ORGANIZATION_TYPE.Realtor 1.169e-01 4.886e-02 2.392 0.016758*
AMT_REQ_CREDIT_BUREAU_WEEK -2.747e-02 1.143e-02 -2.403 0.016282 *
WEEKDAY_APPR_PROCESS_START.MONDAY -1.053e-02 4.680e-03 -2.250 0.024479 *
ORGANIZATION_TYPE.Industry..type.13 2.658e-01 1.195e-01 2.223 0.026204 *
OCCUPATION_TYPE.Cooking.staff 3.033e-02 1.256e-02 2.416 0.015719 *
NAME_TYPE_SUITE.Other_B 4.968e-02 2.426e-02 2.048 0.040582 *
ORGANIZATION_TYPE.Mobile -1.132e-01 5.463e-02 -2.073 0.038210 *
ORGANIZATION_TYPE.School -2.123e-02 1.039e-02 -2.043 0.041046 *
FLAG_PHONE -7.783e-03 4.032e-03 -1.930 0.053595 .
ORGANIZATION_TYPE.Security -2.893e-02 1.663e-02 -1.739 0.081999 .
ORGANIZATION_TYPE.Transport..type.3 5.085e-02 2.774e-02 1.833 0.066768 .
ORGANIZATION_TYPE.Bank -3.393e-02 1.949e-02 -1.741 0.081768 .
DAYS_LAST_PHONE_CHANGE 3.768e-06 2.141e-06 1.760 0.078379 .
ORGANIZATION_TYPE.Housing -3.085e-02 1.727e-02 -1.787 0.073950 .
ORGANIZATION_TYPE.Emergency -6.476e-02 3.908e-02 -1.657 0.097519 .
ORGANIZATION_TYPE.Industry..type.7 -4.438e-02 2.609e-02 -1.701 0.089016 .
LIVE_REGION_NOT_WORK_REGION -1.508e-02 9.015e-03 -1.673 0.094360 .
OCCUPATION_TYPE.Laborers 7.505e-03 4.908e-03 1.529 0.126204
ORGANIZATION_TYPE.Cleaning 9.569e-02 5.977e-02 1.601 0.109436
ORGANIZATION_TYPE.Transport..type.2 3.007e-02 2.076e-02 1.449 0.147480
NAME_FAMILY_STATUS.Single. . . not.married 7.167e-03 4.974e-03 1.441 0.149600
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.267 on 24402 degrees of freedom Multiple R-squared: 0.06976, Adjusted R-squared: 0.06808 F-statistic: 41.59 on 44 and 24402 DF, p-value: < 2.2e-16

```r
fwd_names <- names(forward.lm$coefficients)
fwd_loop <- "TARGET ~ "
```

```r
for (name in fwd_names[2: length(fwd_names)]) {
  fwd_loop <- paste0(fwd_loop, "+", name, sep = "")
}

fwd_all <- as.formula(fwd_loop)
fwd <- lm(fwd_all, data = dummy_train)
```

Compute training and test MSE

```r
# Compute training MSE
yhat_fwd_train <- predict(fwd)
mse_fwd_train <- mean((dummy_train$TARGET- yhat_fwd_train)^2)

# Compute test MSE
yhat_fwd_test <- predict(fwd, dummy_test)
mse_fwd_test <- mean((application_test$TARGET- yhat_fwd_test)^2)

print(paste("MSE of training dataset is", signif(mse_fwd_train,4 )))
print(paste("MSE of testing dataset is", signif(mse_fwd_test,4 )))
```

We reuse the MSE from our previous process.

```r
mse_fwd_train = 0.06986
mse_fwd_test = 0.06988

print(paste("MSE of training dataset is", signif(mse_fwd_train,4 )))
```

```
## [1] "MSE of training dataset is 0.06986"
```

```r
print(paste("MSE of testing dataset is", signif(mse_fwd_test,4 )))
```

```
## [1] "MSE of testing dataset is 0.06988"
```

Backward Selection

```r
backward.lm <- stepAIC(full, scope=list(lower=null, upper=full),
                      direction="backward")
```

Step: AIC=-64521.53 TARGET ~ NAME_CONTRACT_TYPE.Revolving.loans + CODE_GENDER.M + FLAG_OWN_CAR.Y + AMT_CREDIT + AMT_ANNUITY + AMT_GOODS_PRICE + NAME_TYPE_SUITE.Family + NAME_TYPE_SUITE.Other_B + NAME_TYPE_SUITE.Spouse..partner + NAME_INCOME_TYPE.Commercial.associate + NAME_INCOME_TYPE.State.servant + NAME_EDUCATION_TYPE.Lower.secondary + NAME_EDUCATION_TYPE.Secondary...secondary.special + NAME_FAMILY_STATUS.Widow + NAME_HOUSING_TYPE.House...apartment + DAYS_BIRTH + DAYS_EMPLOYED + DAYS_ID_PUBLISH + FLAG_WORK_PHONE + FLAG_PHONE + OCCUPATION_TYPE.Cooking.staff + OCCUPATION_TYPE.High.skill.tech.staff + OCCUPATION_TYPE.Laborers + OCCUPATION_TYPE.Low.skill.Laborers + REGION_RATING_CLIENT + WEEKDAY_APPR_PROCESS_START.MONDAY + WEEKDAY_APPR_PROCESS_START.SUNDAY + LIVE_REGION_NOT_WORK_REGION + ORGANIZATION_TYPE.Business.Entity.Type.3

+ ORGANIZATION_TYPE.Cleaning + ORGANIZATION_TYPE.Construction + ORGANIZATION_TYPE.Industry..type.1 + ORGANIZATION_TYPE.Industry..type.13 + ORGANIZATION_TYPE.Insurance + ORGANIZATION_TYPE.Legal.Services + ORGANIZATION_TYPE.Medicine + ORGANIZATION_TYPE.Mobile + ORGANIZATION_TYPE.Other + ORGANIZATION_TYPE.Realtor + ORGANIZATION_TYPE.Self.employed + ORGANIZATION_TYPE.Transport..type.2 + ORGANIZATION_TYPE.Transport..type.3 + ORGANIZATION_TYPE.Transport..type.4 + EXT_SOURCE_2 + EXT_SOURCE_3 + DAYS_LAST_PHONE_CHANGE + AMT_REQ_CREDIT_BUREAU_WEEK

```
                                               Df Sum of Sq    RSS    AIC
```

1739.0 -64522 - ORGANIZATION_TYPE.Insurance 1 0.1432 1739.2 -64522 - ORGANIZATION_TYPE.Medicine 1 0.1440 1739.2 -64522 - ORGANIZATION_TYPE.Legal.Services 1 0.1522 1739.2 -64521 - NAME_FAMILY_STATUS.Widow 1 0.1541 1739.2 -64521 - NAME_EDUCATION_TYPE.Lower.secondary 1 0.1747 1739.2 -64521 - ORGANIZATION_TYPE.Transport..type.4 1 0.1836 1739.2 -64521 - NAME_TYPE_SUITE.Other_B 1 0.1845 1739.2 -64521 - NAME_TYPE_SUITE.Spouse..partner 1 0.1962 1739.2 -64521 - ORGANIZATION_TYPE.Industry..type.1 1 0.2001 1739.2 -64521 - LIVE_REGION_NOT_WORK_REGION 1 0.2050 1739.3 -64521 - DAYS_LAST_PHONE_CHANGE 1 0.2260 1739.3 -64520 - ORGANIZATION_TYPE.Mobile 1 0.2264 1739.3 -64520 - OCCUPATION_TYPE.Laborers 1 0.2336 1739.3 -64520 - ORGANIZATION_TYPE.Cleaning 1 0.2423 1739.3 -64520 - NAME_INCOME_TYPE.Commercial.associate 1 0.2425 1739.3 -64520 - FLAG_PHONE 1 0.2955 1739.3 -64519 - ORGANIZATION_TYPE.Transport..type.2 1 0.2978 1739.3 -64519 - NAME_HOUSING_TYPE.House…apartment 1 0.3338 1739.4 -64519 - ORGANIZATION_TYPE.Other 1 0.3468 1739.4 -64519 - WEEKDAY_APPR_PROCESS_START.MONDAY 1 0.3579 1739.4 -64518 - ORGANIZATION_TYPE.Industry..type.13 1 0.3803 1739.4 -64518 - ORGANIZATION_TYPE.Transport..type.3 1 0.4040 1739.5 -64518 - AMT_REQ_CREDIT_BUREAU_WEEK 1 0.4042 1739.5 -64518 - NAME_INCOME_TYPE.State.servant 1 0.4368 1739.5 -64517 - NAME_TYPE_SUITE.Family 1 0.4387 1739.5 -64517 - OCCUPATION_TYPE.Cooking.staff 1 0.4624 1739.5 -64517 - WEEKDAY_APPR_PROCESS_START.SUNDAY 1 0.4753 1739.5 -64517 - ORGANIZATION_TYPE.Realtor 1 0.5320 1739.6 -64516 - OCCUPATION_TYPE.High.skill.tech.staff 1 0.6525 1739.7 -64514 - REGION_RATING_CLIENT 1 0.7530 1739.8 -64513 - NAME_CONTRACT_TYPE.Revolving.loans 1 0.7700 1739.8 -64513 - DAYS_ID_PUBLISH 1 0.7701 1739.8 -64513 - OCCUPATION_TYPE.Low.skill.Laborers 1 0.8017 1739.8 -64512 - ORGANIZATION_TYPE.Self.employed 1 0.8642 1739.9 -64511 - ORGANIZATION_TYPE.Construction 1 0.8981 1740.0 -64511 - ORGANIZATION_TYPE.Business.Entity.Type.3 1 1.0409 1740.1 -64509 - AMT_ANNUITY 1 1.1490 1740.2 -64507 - DAYS_EMPLOYED 1 1.2949 1740.3 -64505 - DAYS_BIRTH 1 1.3641 1740.4 -64504 - FLAG_WORK_PHONE 1 1.7107 1740.8 -64499 - AMT_CREDIT 1 2.7980 1741.8 -64484 - NAME_EDUCATION_TYPE.Secondary…secondary.special 1 2.9545 1742.0 -64482 - FLAG_OWN_CAR.Y 1 2.9712 1742.0 -64482 - AMT_GOODS_PRICE 1 3.3587 1742.4 -64476 - CODE_GENDER.M 1 3.8714 1742.9 -64469 - EXT_SOURCE_2 1 25.4149 1764.5 -64169 - EXT_SOURCE_3 1 30.0509 1769.1 -64105

```r
## Backward Stepwise Regression
#####

bck_names <- names(backward.lm$coefficients)
bck_loop <- "TARGET ~ "

for (name in bck_names[2: length(bck_names)]) {
  bck_loop <- paste0(bck_loop, "+", name, sep = "")
}

bck_all <- as.formula(bck_loop)

bck <- lm(bck_all, data = dummy_train)
```

Compute training and test MSE

```r
# Compute training MSE
yhat_bck_train <- predict(bck)
mse_bck_train <- mean((dummy_train$TARGET- yhat_bck_train)^2)

# Compute test MSE
yhat_bck_test <- predict(bck, dummy_test)
mse_bck_test <- mean((dummy_test$TARGET- yhat_bck_test)^2)

print(paste("MSE of training dataset is", signif(mse_bck_train,4 )))
print(paste("MSE of testing dataset is", signif(mse_bck_test,4 )))
```

```r
mse_bck_train = 0.06985
mse_bck_test = 0.06987

print(paste("MSE of training dataset is", signif(mse_bck_train,4 )))
```

```
## [1] "MSE of training dataset is 0.06985"
```

```r
print(paste("MSE of testing dataset is", signif(mse_bck_test,4 )))
```

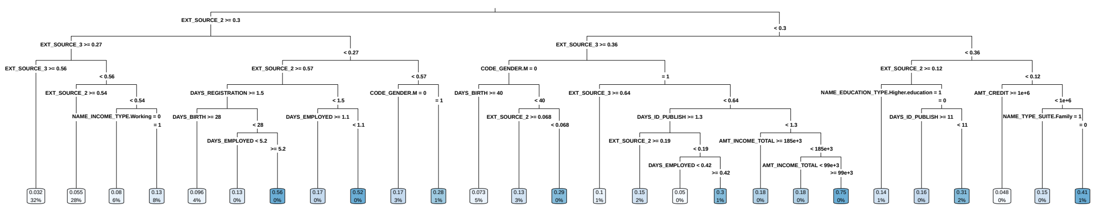
```
## [1] "MSE of testing dataset is 0.06987"
```

Decision Tree

```r
f1 <- as.formula(TARGET ~ . -SK_ID_CURR -TRAIN)

fit.tree <- rpart(f1, dummy_subset_train,
                  control = rpart.control(cp = 0.001))

rpart.plot(fit.tree, type = 3)
```



```r
yhat.train.tree <- predict(fit.tree, dummy_train)
mse.train.tree <- mean((dummy_train$TARGET - yhat.train.tree)^2)
mse.train.tree
```

```
## [1] 0.07149059
```

```r
yhat.test.tree <- predict(fit.tree, dummy_test)
mse.test.tree <- mean((dummy_test$TARGET - yhat.test.tree)^2)
mse.test.tree
```

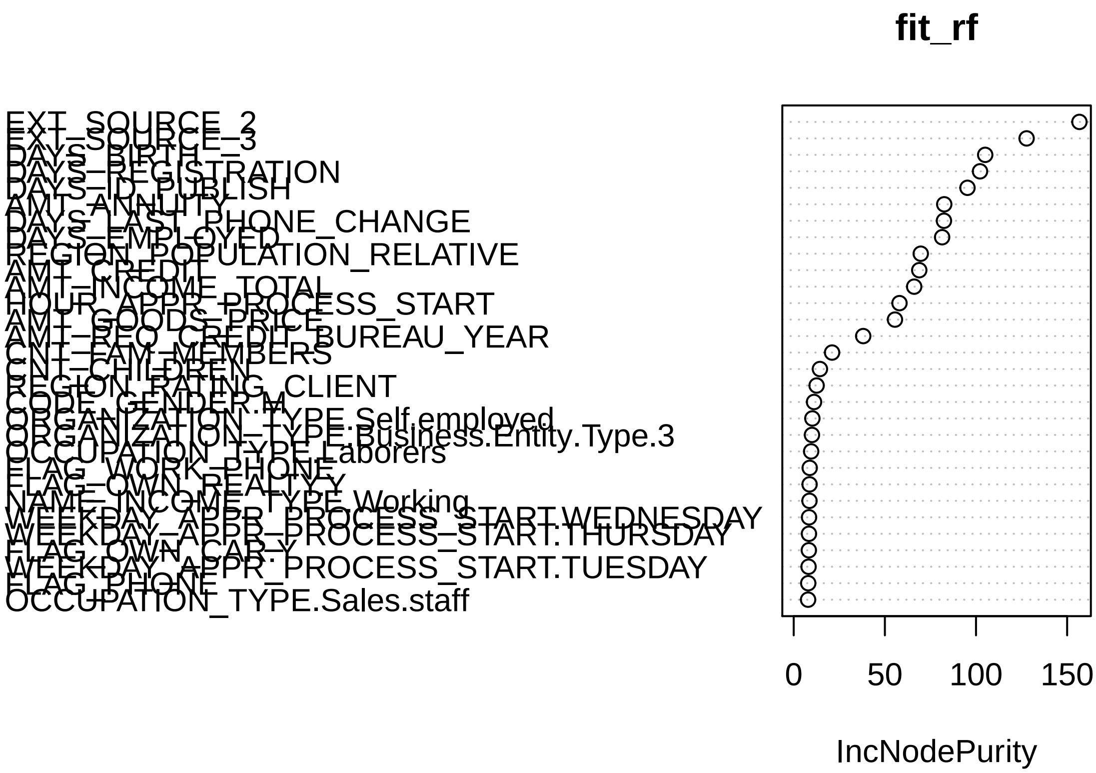
```
## [1] 0.07192651
```

Random Forest

```r
fit_rf <- randomForest(f1, dummy_subset_train, ntree = 200, do.trace = F)
```

```
## Warning in randomForest.default(m, y, ...): The response has five or fewer
## unique values. Are you sure you want to do regression?
```

```r
## Check which variables are most predictive using a variable importance plot.
varImpPlot(fit_rf)
```




```r
## Predictions and compute a train MSE.
yhat_rf_train <- predict(fit_rf, dummy_train)
mse_rf_train <- mean((yhat_rf_train - dummy_train$TARGET) ^ 2)
print(mse_rf_train)
```

```
## [1] 0.06531608
```

```r
## Predictions and compute the MSE's.
yhat_rf_test <- predict(fit_rf, dummy_test)
mse_rf_test <- mean((yhat_rf_test - dummy_test$TARGET) ^ 2)
print(mse_rf_test)
```

```
## [1] 0.07107114
```

Boosting

Here we tried to optimize the model by tuning the parameters through K-fold cross validations, the best model would have lowest RMSE in validation dataset.
In order to save time, we only choose to tune the `interaction.depth` parameters, set other parameters in the function as constant. We also randomly selected $\frac{1}{10}$ `application_train` to be `subset_train`, and used it to find out the optimized model, then apply it to the complete dataset.

```r
f2 <- as.formula(TARGET ~ . - SK_ID_CURR - TRAIN)
```

```r
#Because it was extremely time-comsuming to train the model with such large sample size, so we decided

fitControl <- trainControl(## 5-fold CV
                           method = "repeatedcv",
                           number = 5,
                           ## repeated five times
                           repeats = 5)

gbmGrid <-  expand.grid(interaction.depth = 1:5,
                        n.trees = 200,
                        shrinkage = 0.01,
                        n.minobsinnode = 10)

set.seed(7)
gbmFit <- train(f2, data = subset_train,
                 method = "gbm",
                 trControl = fitControl,
                 verbose = FALSE,
                tuneGrid = gbmGrid)

gbmFit
```

As the result, best performed model has `interactin.depth = 4`.
Then we applied it on the complete `application_train` dataset.

```r
fit_btree <- gbm(f2,
data = application_train,
distribution = "gaussian",
n.trees = 200,
interaction.depth = 4,
shrinkage = 0.01)
```

```r
relative.influence(fit_btree)
```

```
## n.trees not given. Using 200 trees.
```

```
##         NAME_CONTRACT_TYPE                CODE_GENDER
##                    0.00000                  373.00914
##               FLAG_OWN_CAR            FLAG_OWN_REALTY
##                   64.23079                    0.00000
##               CNT_CHILDREN           AMT_INCOME_TOTAL
##                    0.00000                    0.00000
##                 AMT_CREDIT                AMT_ANNUITY
##                    0.00000                    0.00000
##            AMT_GOODS_PRICE            NAME_TYPE_SUITE
##                    0.00000                    0.00000
##           NAME_INCOME_TYPE        NAME_EDUCATION_TYPE
##                    0.00000                  392.07649
##         NAME_FAMILY_STATUS          NAME_HOUSING_TYPE
##                    0.00000                    0.00000
```

```
##  REGION_POPULATION_RELATIVE                  DAYS_BIRTH
##                     0.00000                   265.95191
##               DAYS_EMPLOYED           DAYS_REGISTRATION
##                   226.40463                     0.00000
##             DAYS_ID_PUBLISH                  FLAG_MOBIL
##                     0.00000                     0.00000
##              FLAG_EMP_PHONE             FLAG_WORK_PHONE
##                     0.00000                     0.00000
##            FLAG_CONT_MOBILE                  FLAG_PHONE
##                     0.00000                     0.00000
##                  FLAG_EMAIL             OCCUPATION_TYPE
##                     0.00000                   954.63684
##             CNT_FAM_MEMBERS        REGION_RATING_CLIENT
##                     0.00000                     0.00000
##  WEEKDAY_APPR_PROCESS_START     HOUR_APPR_PROCESS_START
##                     0.00000                     0.00000
##  REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION
##                     0.00000                     0.00000
## LIVE_REGION_NOT_WORK_REGION           ORGANIZATION_TYPE
##                     0.00000                  1091.58904
##                EXT_SOURCE_2                EXT_SOURCE_3
##                 11269.66411                 11331.69974
##       DAYS_LAST_PHONE_CHANGE  AMT_REQ_CREDIT_BUREAU_WEEK
##                     0.00000                     0.00000
##  AMT_REQ_CREDIT_BUREAU_YEAR
##                     0.00000
```

```
yhat_btree <- predict(fit_btree, application_train, n.trees = 200)
mse_btree <- mean((yhat_btree - application_train$TARGET) ^ 2)

yhat_btree_test <- predict(fit_btree, application_test, n.trees = 200)
mse_btree_test <- mean((yhat_btree_test - application_test$TARGET) ^ 2)

print(paste("MSE of training dataset is", signif(mse_btree,4 )))
```

```
## [1] "MSE of training dataset is 0.0702"
```

```
print(paste("MSE of testing dataset is", signif(mse_btree_test,4 )))
```

```
## [1] "MSE of testing dataset is 0.07029"
```

```
mse_result <- tibble(Model = c("Linear Regression", "Forward Selection", "Backward Selection",
                               "Ridge", "Lasso", "Decision Trees",
                               "Random Forest", "Boosting Trees"),
                     MSE_Train= c(signif(0.06972772,6), signif(0.06986146,6), signif(0.06985078,6),
                                  signif(0.06976723,6), signif(0.06973808,6), signif(0.07149059,6),
                                  signif(0.06535254,6), signif(0.07020926,6)),
                     MSE_Test = c(signif(0.06982647,6), signif(0.06988122,6), signif(0.06987248,6),
                                  signif(0.06986277,6), signif(0.06982388,6), signif(0.07192651,6),
                                  signif(0.07106714,6), signif(0.070292,6)))
mse_tidy <- gather(mse_result, type, mse, -Model)
```

```r
ggplot(mse_tidy, aes(x=Model, y=mse, fill=type)) +
  geom_histogram(stat = "identity", position = "dodge") +
   geom_hline(yintercept = 0.06982388, linetype="dashed") +
  coord_cartesian(ylim = c(0.065, 0.072)) +
  theme(axis.text.x = element_text(angle = 50, vjust = 0.65))
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```