Data Mining and Risk Analysis - Credit Score System

Sunny Jing

2/6/2017

Problem Description

Through this project, we will build a credit score system based on data from previously available credit accounts, and determine a cut-off score at which the bank should offer credit to a given customer in different senarios. The focus of the project is not to build a good predictive model, but rather to show, as a bank or a credit agency, how to determine a cutoff score given a set of predictions or scores of the potential borrowers.

In determining a cut-off credit score, we consider the following two senarios:

Scenario 1: Company targets the default rate for the newly approved cards to be 10% (risk driven)

Scenario 2: Company is expecting 100,000 new applications and target the new approvals 50,000 (market share driven)

Data Cleaning and Preparation

```
datapath<-"/Users/dongpingjing/Dropbox/Data Mining/Project Folder/Revised Raw Data Set"
original<-read.csv(file=paste(datapath,"LoanStats3b.csv",sep="/"))

# Overview of the data
str(original)
```

```
## 'data.frame':    136942 obs. of  111 variables:
##  $ id                         : Factor w/ 136942 levels "1000848","1004665",..: 31 39 35 41 37 28
##  $ member_id                  : int   11981072 12001069 11999781 12001118 12000897 11971241 1208673
##  $ loan_amnt                  : int   12000 27600 12000 28000 27050 12000 11100 9750 12000 4800 ..
##  $ funded_amnt                : int   12000 27600 12000 28000 27050 12000 11100 9750 12000 4800 ..
##  $ funded_amnt_inv            : num   12000 27600 12000 28000 27050 ...
##  $ term                       : Factor w/ 2 levels " 36 months"," 60 months": 1 2 1 1 1 1 1 1 1 1
##  $ int_rate                   : num   0.11 0.2 0.076 0.076 0.11 0.12 0.15 0.14 0.135 0.11 ...
##  $ installment                : num   393 731 374 873 885 ...
##  $ grade                      : Factor w/ 7 levels "A","B","C","D",..: 2 4 1 1 2 2 3 3 2 2 ...
##  $ sub_grade                  : Factor w/ 35 levels "A1","A2","A3",..: 7 20 3 3 7 8 13 11 10 7 ..
##  $ emp_title                  : Factor w/ 92176 levels ""," 	  MOTHER  THERESA  HOME",..: 64604 7
##  $ emp_length                 : Factor w/ 12 levels "< 1 year","1 year",..: 6 8 5 7 3 3 3 2 3 4 .
##  $ home_ownership             : Factor w/ 4 levels "MORTGAGE","OTHER",..: 4 1 1 1 3 1 1 4 4 1 ...
##  $ annual_inc                 : num   60000 73000 96500 325000 55000 130000 90000 26000 40000 39600
##  $ verification_status        : Factor w/ 3 levels "Not Verified",..: 1 2 1 2 3 2 1 1 2 2 ...
##  $ issue_d                    : Factor w/ 23 levels "12-Apr","12-Aug",..: 14 14 14 14 14 14 14 14
##  $ loan_status                : Factor w/ 7 levels "Charged Off",..: 4 2 4 4 4 4 4 4 4 4 ...
##  $ pymnt_plan                 : Factor w/ 1 level "n": 1 1 1 1 1 1 1 1 1 1 ...
##  $ url                        : Factor w/ 136942 levels "https://lendingclub.com/browse/loanDetai
##  $ desc                       : Factor w/ 66318 levels ""," 	 Loan for purchase of grand piano. P
##  $ purpose                    : Factor w/ 13 levels "car","credit_card",..: 3 3 3 3 3 3 9 3 3 4 .
##  $ title                      : Factor w/ 34952 levels ""," 	credit_card",..: 25067 7160 11132 27
##  $ zip_code                   : Factor w/ 834 levels "010xx","011xx",..: 247 674 654 784 431 681 
##  $ addr_state                 : Factor w/ 48 levels "AK","AL","AR",..: 27 6 41 5 22 6 32 5 30 41 
```

1

```
##  $ dti                        : num  4.62 23.13 12.61 18.55 22.87 ...
##  $ delinq_2yrs                : int  0 1 0 0 0 0 1 0 0 0 ...
##  $ earliest_cr_line           : Factor w/ 599 levels "1-Apr","1-Aug",..: 109 378 46 510 545 513 7
##  $ inq_last_6mths             : int  1 1 0 1 0 1 0 0 0 2 ...
##  $ mths_since_last_delinq      : int  NA 7 NA NA NA NA 16 NA 53 NA ...
##  $ mths_since_last_record      : int  NA NA NA NA NA NA NA NA 33 NA ...
##  $ open_acc                   : int  15 10 17 15 14 9 9 12 7 3 ...
##  $ pub_rec                    : int  0 0 0 0 0 0 0 0 2 0 ...
##  $ revol_bal                  : int  7137 27003 13248 29581 36638 10805 6619 7967 5572 4136 ...
##  $ revol_util                 : Factor w/ 1061 levels "",," -    "," 0.001 ",..: 242 830 559 548 614
##  $ total_acc                  : int  18 24 30 31 27 19 12 28 32 8 ...
##  $ initial_list_status         : Factor w/ 2 levels "f","w": 1 1 1 2 2 1 1 1 2 2 ...
##  $ out_prncp                  : num  0 14362 0 0 0 ...
##  $ out_prncp_inv              : num  0 14362 0 0 0 ...
##  $ total_pymnt                : num  13989 26308 13398 29151 31753 ...
##  $ total_pymnt_inv            : num  13989 26308 13398 29151 31753 ...
##  $ total_rec_prncp            : num  12000 13238 12000 28000 27050 ...
##  $ total_rec_int              : num  1989 13070 1398 1151 4703 ...
##  $ total_rec_late_fee          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ recoveries                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ collection_recovery_fee     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ last_pymnt_d               : Factor w/ 61 levels "","12-Apr","12-Aug",..: 49 61 55 26 54 61 52
##  $ last_pymnt_amnt            : num  3776 731 2927 23916 6074 ...
##  $ next_pymnt_d               : Factor w/ 4 levels "","17-Feb","17-Jan",..: 1 2 1 1 1 1 1 1 1 1 .
##  $ last_credit_pull_d          : Factor w/ 62 levels "","12-Apr","12-Aug",..: 62 62 56 27 59 62 51
##  $ collections_12_mths_ex_med  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog : int  NA 7 NA NA NA NA 16 NA 53 NA ...
##  $ policy_code                : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type           : Factor w/ 1 level "INDIVIDUAL": 1 1 1 1 1 1 1 1 1 1 ...
##  $ annual_inc_joint            : logi  NA NA NA NA NA NA ...
##  $ dti_joint                  : logi  NA NA NA NA NA NA ...
##  $ verification_status_joint   : logi  NA NA NA NA NA NA ...
##  $ acc_now_delinq             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt               : int  0 0 0 0 0 0 0 0 15386 0 ...
##  $ tot_cur_bal                : int  7137 241609 200314 799592 114834 327264 353402 14123 13605 4
##  $ open_acc_6m                : logi  NA NA NA NA NA NA ...
##  $ open_il_6m                 : logi  NA NA NA NA NA NA ...
##  $ open_il_12m                : logi  NA NA NA NA NA NA ...
##  $ open_il_24m                : logi  NA NA NA NA NA NA ...
##  $ mths_since_rcnt_il          : logi  NA NA NA NA NA NA ...
##  $ total_bal_il               : logi  NA NA NA NA NA NA ...
##  $ il_util                    : logi  NA NA NA NA NA NA ...
##  $ open_rv_12m                : logi  NA NA NA NA NA NA ...
##  $ open_rv_24m                : logi  NA NA NA NA NA NA ...
##  $ max_bal_bc                 : logi  NA NA NA NA NA NA ...
##  $ all_util                   : logi  NA NA NA NA NA NA ...
##  $ total_rev_hi_lim            : int  29700 32600 23800 54200 59900 16200 10000 15100 8100 25700 .
##  $ inq_fi                     : logi  NA NA NA NA NA NA ...
##  $ total_cu_tl                : logi  NA NA NA NA NA NA ...
##  $ inq_last_12m               : logi  NA NA NA NA NA NA ...
##  $ acc_open_past_24mths        : int  8 2 4 6 3 4 2 2 4 0 ...
##  $ avg_cur_bal                : int  476 24161 11783 53306 9570 36362 39267 1177 2268 1379 ...
##  $ bc_open_to_buy             : int  15216 4853 2441 13901 16473 3567 1016 1752 1428 21564 ...
##  $ bc_util                    : num  15.9 74.7 83.5 67.1 53.9 93 74.6 75.7 79.6 16.1 ...
```

```
##  $ chargeoff_within_12_mths    : int   0 1 0 0 0 0 0 0 0 0 ...
##  $ delinq_amnt                 : int   0 0 0 0 0 0 0 0 0 0 ...
##  $ mo_sin_old_il_acct          : int   NA 173 123 125 117 173 NA 67 124 104 ...
##  $ mo_sin_old_rev_tl_op        : int   48 294 118 229 326 193 150 83 182 220 ...
##  $ mo_sin_rcnt_rev_tl_op       : int   1 4 10 5 16 4 11 12 1 25 ...
##  $ mo_sin_rcnt_tl              : int   1 4 9 2 6 4 11 12 1 25 ...
##  $ mort_acc                    : int   0 4 1 5 4 3 1 0 0 0 ...
##  $ mths_since_recent_bc        : int   1 4 10 5 16 85 11 12 11 25 ...
##  $ mths_since_recent_bc_dlq    : int   NA 7 NA NA NA NA 35 NA 53 NA ...
##  $ mths_since_recent_inq       : int   3 6 10 3 8 4 11 20 17 3 ...
##  $ mths_since_recent_revol_delinq: int  NA 7 NA NA NA NA 35 NA 53 NA ...
##  $ num_accts_ever_120_pd       : int   0 3 0 0 0 0 1 0 6 0 ...
##  $ num_actv_bc_tl              : int   4 5 4 4 2 3 4 6 2 2 ...
##  $ num_actv_rev_tl             : int   7 7 5 5 4 5 8 7 2 2 ...
##  $ num_bc_sats                 : int   8 5 4 6 4 4 4 6 3 3 ...
##  $ num_bc_tl                   : int   10 11 10 8 8 4 4 11 14 4 ...
##  $ num_il_tl                   : int   0 4 15 11 8 8 0 8 8 1 ...
##  $ num_op_rev_tl               : int   15 7 8 9 10 5 8 9 6 3 ...
##  $ num_rev_accts               : int   18 16 14 15 15 8 11 20 24 7 ...
##  $ num_rev_tl_bal_gt_0         : int   7 7 5 5 4 5 8 7 2 2 ...
##  $ num_sats                    : int   15 10 17 15 14 9 9 12 7 3 ...
##   [list output truncated]
```

We are limiting the study to the accounts that were granted in 2014 for a period of 36 months, so we will remove all the 60 months accounts.

```
# Keep only 36 month term data:
original<-original[original$term==" 36 months",]
original$term<-NULL

loan.data<-original
colnames(loan.data)
```

```
##   [1] "id"                    "member_id"
##   [3] "loan_amnt"             "funded_amnt"
##   [5] "funded_amnt_inv"       "int_rate"
##   [7] "installment"           "grade"
##   [9] "sub_grade"             "emp_title"
##  [11] "emp_length"            "home_ownership"
##  [13] "annual_inc"            "verification_status"
##  [15] "issue_d"               "loan_status"
##  [17] "pymnt_plan"            "url"
##  [19] "desc"                  "purpose"
##  [21] "title"                 "zip_code"
##  [23] "addr_state"            "dti"
##  [25] "delinq_2yrs"           "earliest_cr_line"
##  [27] "inq_last_6mths"        "mths_since_last_delinq"
##  [29] "mths_since_last_record" "open_acc"
##  [31] "pub_rec"               "revol_bal"
##  [33] "revol_util"            "total_acc"
##  [35] "initial_list_status"   "out_prncp"
##  [37] "out_prncp_inv"         "total_pymnt"
##  [39] "total_pymnt_inv"       "total_rec_prncp"
##  [41] "total_rec_int"         "total_rec_late_fee"
##  [43] "recoveries"            "collection_recovery_fee"
```

```
##  [45] "last_pymnt_d"                "last_pymnt_amnt"
##  [47] "next_pymnt_d"                "last_credit_pull_d"
##  [49] "collections_12_mths_ex_med"  "mths_since_last_major_derog"
##  [51] "policy_code"                 "application_type"
##  [53] "annual_inc_joint"            "dti_joint"
##  [55] "verification_status_joint"   "acc_now_delinq"
##  [57] "tot_coll_amt"                "tot_cur_bal"
##  [59] "open_acc_6m"                 "open_il_6m"
##  [61] "open_il_12m"                 "open_il_24m"
##  [63] "mths_since_rcnt_il"          "total_bal_il"
##  [65] "il_util"                     "open_rv_12m"
##  [67] "open_rv_24m"                 "max_bal_bc"
##  [69] "all_util"                    "total_rev_hi_lim"
##  [71] "inq_fi"                      "total_cu_tl"
##  [73] "inq_last_12m"                "acc_open_past_24mths"
##  [75] "avg_cur_bal"                 "bc_open_to_buy"
##  [77] "bc_util"                     "chargeoff_within_12_mths"
##  [79] "delinq_amnt"                 "mo_sin_old_il_acct"
##  [81] "mo_sin_old_rev_tl_op"        "mo_sin_rcnt_rev_tl_op"
##  [83] "mo_sin_rcnt_tl"              "mort_acc"
##  [85] "mths_since_recent_bc"        "mths_since_recent_bc_dlq"
##  [87] "mths_since_recent_inq"       "mths_since_recent_revol_delinq"
##  [89] "num_accts_ever_120_pd"       "num_actv_bc_tl"
##  [91] "num_actv_rev_tl"             "num_bc_sats"
##  [93] "num_bc_tl"                   "num_il_tl"
##  [95] "num_op_rev_tl"               "num_rev_accts"
##  [97] "num_rev_tl_bal_gt_0"         "num_sats"
##  [99] "num_tl_120dpd_2m"            "num_tl_30dpd"
## [101] "num_tl_90g_dpd_24m"          "num_tl_op_past_12m"
## [103] "pct_tl_nvr_dlq"              "percent_bc_gt_75"
## [105] "pub_rec_bankruptcies"        "tax_liens"
## [107] "tot_hi_cred_lim"             "total_bal_ex_mort"
## [109] "total_bc_limit"              "total_il_high_credit_limit"
```

This study is done at the end of 2016. Therefore, any account granted in 2013 with term period of 36 months should have been paid in full. In defining the response variable for the model study, we assgin all the accounts granted in 2013 that are not paid in full to be 1, as "defaulted", and those that are paid in full to be 0, as "not defaulted"

```r
# Response varible:
loan.data$defaulted[loan.data$loan_status=="Fully Paid"]<-0
loan.data$defaulted[is.na(loan.data$defaulted)]<-1
loan.data$defaulted<-factor(loan.data$defaulted)
loan.data$loan_status<-NULL
table(loan.data$defaulted)
```

```
##
##     0     1
## 91194 13487
```

In data cleaning, we remove all the variables that don't carry any meaningful value, such as id, url... We also remove all the variables that are compeletly blank, or have unified value. Since the model will be used to predict credit score for any new customer, we will not have information such as loan amount, interest rate, etc., we will remove these variables that are associated with an existing account.

```
# Remove text value variables:
loan.data$url<-NULL
loan.data$emp_title<-NULL
loan.data$desc<-NULL
loan.data$zip_code<-NULL
loan.data$addr_state<-NULL
loan.data$title<-NULL   # too messy

# Remove data variables and ID:
loan.data$id<-NULL
loan.data$member_id<-NULL
loan.data$issue_d<-NULL
loan.data$earliest_cr_line<-NULL
loan.data$last_pymnt_d<-NULL
loan.data$next_pymnt_d<-NULL #Some fully paid account still have scheduled next payment in Feb 2017.
loan.data$last_credit_pull_d<-NULL
loan.data$grade<-NULL
loan.data$sub_grade<-NULL

# variables to remove because of blank:
blank<-apply(loan.data,2,function(x){sum(is.na(x))})
(blank.col<-colnames(loan.data[,blank==nrow(loan.data)]))
```

```
##  [1] "annual_inc_joint"        "dti_joint"
##  [3] "verification_status_joint" "open_acc_6m"
##  [5] "open_il_6m"              "open_il_12m"
##  [7] "open_il_24m"             "mths_since_rcnt_il"
##  [9] "total_bal_il"            "il_util"
## [11] "open_rv_12m"             "open_rv_24m"
## [13] "max_bal_bc"              "all_util"
## [15] "inq_fi"                  "total_cu_tl"
## [17] "inq_last_12m"
```

```
loan.data[,blank.col]<-NULL

# Make sure class right (revol_util as a percentange value is recognized as a categorical value, so we
loan.data$revol_util<-as.numeric(levels(loan.data$revol_util)[loan.data$revol_util])
```

```
## Warning: NAs introduced by coercion
```

```
head(loan.data$revol_util)
```

```
## [1] 0.240 0.557 0.546 0.612 0.670 0.662
```

```
# variables to remove because of unified value:
one.value<-apply(loan.data,2,function(x){length(unique(x))})
(unified.col<-colnames(loan.data[,one.value==1]))
```

```
## [1] "pymnt_plan"        "policy_code"        "application_type"
```

```
loan.data[,unified.col]<-NULL

# Remove variables regarding current credit card infomation:

loan.data[,c("loan_amnt",
            "funded_amnt",
            "funded_amnt_inv",
```

```
            "int_rate",
            "installment",
            "out_prncp",
            "out_prncp_inv",
            "total_pymnt",
            "total_pymnt_inv",
            "total_rec_prncp",
            "total_rec_int",
            "total_rec_late_fee",
            "recoveries",
            "collection_recovery_fee",
            "last_pymnt_amnt")]<-NULL
```

Finding and imputing for missing values:

In inputing for missing values, we take two approaches, 99 quantile imputing and 0 imputing, depanding on the variable meanings. We impute 99 quantile for any variables that represent length of time. For example. variable "mths_since_last_delinq" means number of months since the borrower's last delinquent. When this variable is blank, it normally means that the borrower did not have a delinquent. Therefore, we need to impute a large enough value for its missing value to avoid distorting the meaning of the value. For variables that represent amount, we impute 0.

```
missing<-apply(loan.data,2,function(x){sum(is.na(x))})
colnames.missing<-colnames(loan.data[,missing!=0])
colnames.missing<-colnames.missing[order(colnames.missing)]
colnames.missing
```

```
##  [1] "avg_cur_bal"                "bc_open_to_buy"
##  [3] "bc_util"                    "mo_sin_old_il_acct"
##  [5] "mo_sin_old_rev_tl_op"       "mo_sin_rcnt_rev_tl_op"
##  [7] "mo_sin_rcnt_tl"             "mths_since_last_delinq"
##  [9] "mths_since_last_major_derog" "mths_since_last_record"
## [11] "mths_since_recent_bc"       "mths_since_recent_bc_dlq"
## [13] "mths_since_recent_inq"      "mths_since_recent_revol_delinq"
## [15] "num_accts_ever_120_pd"      "num_actv_bc_tl"
## [17] "num_actv_rev_tl"            "num_bc_tl"
## [19] "num_il_tl"                  "num_op_rev_tl"
## [21] "num_rev_accts"              "num_rev_tl_bal_gt_0"
## [23] "num_tl_120dpd_2m"           "num_tl_30dpd"
## [25] "num_tl_90g_dpd_24m"         "num_tl_op_past_12m"
## [27] "pct_tl_nvr_dlq"             "percent_bc_gt_75"
## [29] "revol_util"                 "tot_coll_amt"
## [31] "tot_cur_bal"                "tot_hi_cred_lim"
## [33] "total_il_high_credit_limit" "total_rev_hi_lim"
```
```
# fill in variables that represent number of months with the 99 quantile of the data:
(colname<-colnames.missing[c(4:14,27)])
```

```
##  [1] "mo_sin_old_il_acct"          "mo_sin_old_rev_tl_op"
##  [3] "mo_sin_rcnt_rev_tl_op"       "mo_sin_rcnt_tl"
##  [5] "mths_since_last_delinq"      "mths_since_last_major_derog"
##  [7] "mths_since_last_record"      "mths_since_recent_bc"
##  [9] "mths_since_recent_bc_dlq"    "mths_since_recent_inq"
## [11] "mths_since_recent_revol_delinq" "pct_tl_nvr_dlq"
```

```r
for (i in colname){
  loan.data[ which(is.na(loan.data[,i])),i]<-quantile(loan.data[which(!is.na(loan.data[,i])),i], 0.99)
}

#fill in variables that represent amounts with 0:
(colname<-colnames.missing[-c(4:14,27)])
```

```
##  [1] "avg_cur_bal"             "bc_open_to_buy"
##  [3] "bc_util"                 "num_accts_ever_120_pd"
##  [5] "num_actv_bc_tl"          "num_actv_rev_tl"
##  [7] "num_bc_tl"               "num_il_tl"
##  [9] "num_op_rev_tl"           "num_rev_accts"
## [11] "num_rev_tl_bal_gt_0"     "num_tl_120dpd_2m"
## [13] "num_tl_30dpd"            "num_tl_90g_dpd_24m"
## [15] "num_tl_op_past_12m"      "percent_bc_gt_75"
## [17] "revol_util"              "tot_coll_amt"
## [19] "tot_cur_bal"             "tot_hi_cred_lim"
## [21] "total_il_high_credit_limit" "total_rev_hi_lim"
```

```r
for (i in colname){
  loan.data[ which(is.na(loan.data[,i])),i] <-0
}

# Variables left:
colnames(loan.data)
```

```
##  [1] "emp_length"                    "home_ownership"
##  [3] "annual_inc"                    "verification_status"
##  [5] "purpose"                       "dti"
##  [7] "delinq_2yrs"                   "inq_last_6mths"
##  [9] "mths_since_last_delinq"        "mths_since_last_record"
## [11] "open_acc"                      "pub_rec"
## [13] "revol_bal"                     "revol_util"
## [15] "total_acc"                     "initial_list_status"
## [17] "collections_12_mths_ex_med"    "mths_since_last_major_derog"
## [19] "acc_now_delinq"                "tot_coll_amt"
## [21] "tot_cur_bal"                   "total_rev_hi_lim"
## [23] "acc_open_past_24mths"          "avg_cur_bal"
## [25] "bc_open_to_buy"                "bc_util"
## [27] "chargeoff_within_12_mths"      "delinq_amnt"
## [29] "mo_sin_old_il_acct"            "mo_sin_old_rev_tl_op"
## [31] "mo_sin_rcnt_rev_tl_op"         "mo_sin_rcnt_tl"
## [33] "mort_acc"                      "mths_since_recent_bc"
## [35] "mths_since_recent_bc_dlq"      "mths_since_recent_inq"
## [37] "mths_since_recent_revol_delinq" "num_accts_ever_120_pd"
## [39] "num_actv_bc_tl"                "num_actv_rev_tl"
## [41] "num_bc_sats"                   "num_bc_tl"
## [43] "num_il_tl"                     "num_op_rev_tl"
## [45] "num_rev_accts"                 "num_rev_tl_bal_gt_0"
## [47] "num_sats"                      "num_tl_120dpd_2m"
## [49] "num_tl_30dpd"                  "num_tl_90g_dpd_24m"
## [51] "num_tl_op_past_12m"            "pct_tl_nvr_dlq"
## [53] "percent_bc_gt_75"             "pub_rec_bankruptcies"
## [55] "tax_liens"                    "tot_hi_cred_lim"
```

```
## [57] "total_bal_ex_mort"          "total_bc_limit"
## [59] "total_il_high_credit_limit"  "defaulted"
```

Select the varaibles for logistic model:

categorical variables

```r
cat<-NULL
n=ncol(loan.data)
for (i in 1:n){
  cat[i]<-class(loan.data[,i])
}
cat.var<-loan.data[,cat=="factor"]

# Try Tree regression:
library("caret")
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```r
library("rpart")
tree<-rpart(defaulted~.,data=cat.var,control=rpart.control(minsplit=30L,cp=0,xval=10L,maxsurrogate = 0))
tree$cptable
```

```
##   CP nsplit rel error xerror xstd
## 1  0      0         1      0    0
```

```r
# Conduct step wise analysis:
logistic.cat<-glm(defaulted~.,data=cat.var,family=binomial(link=logit))
summary(logistic.cat)
```

```
##
## Call:
## glm(formula = defaulted ~ ., family = binomial(link = logit),
##     data = cat.var)
##
## Deviance Residuals:
##     Min      1Q   Median      3Q      Max
## -0.9150  -0.5614  -0.4932  -0.4533   2.2741
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -2.410840   0.098474 -24.482  < 2e-16
## emp_length1 year        -0.090107   0.049932  -1.805  0.07114
## emp_length10+ years     -0.046213   0.038153  -1.211  0.22580
## emp_length2 years       -0.034609   0.045506  -0.761  0.44693
## emp_length3 years       -0.025435   0.047161  -0.539  0.58966
## emp_length4 years       -0.029978   0.050462  -0.594  0.55247
## emp_length5 years       -0.022126   0.047188  -0.469  0.63914
## emp_length6 years        0.045806   0.048953   0.936  0.34942
## emp_length7 years        0.058339   0.050512   1.155  0.24811
## emp_length8 years        0.020540   0.054301   0.378  0.70524
## emp_length9 years       -0.004914   0.059288  -0.083  0.93394
## emp_lengthn/a            0.386075   0.051064   7.561 4.01e-14
## home_ownershipOWN        0.240516   0.034800   6.911 4.80e-12
```

```
## home_ownershipRENT                      0.406747   0.020433  19.906  < 2e-16
## verification_statusSource Verified  0.048983   0.025576   1.915  0.05547
## verification_statusVerified          0.093359   0.021572   4.328 1.51e-05
## purposecredit_card                   0.063133   0.093758   0.673  0.50072
## purposedebt_consolidation            0.277160   0.092360   3.001  0.00269
## purposehome_improvement              0.216345   0.100658   2.149  0.03161
## purposehouse                         0.199790   0.148175   1.348  0.17755
## purposemajor_purchase               -0.006499   0.114605  -0.057  0.95478
## purposemedical                       0.574879   0.126411   4.548 5.42e-06
## purposemoving                        0.277807   0.138970   1.999  0.04560
## purposeother                         0.564975   0.098575   5.731 9.96e-09
## purposerenewable_energy              0.754606   0.315484   2.392  0.01676
## purposesmall_business                0.870562   0.110085   7.908 2.61e-15
## purposevacation                      0.423029   0.142851   2.961  0.00306
## purposewedding                       0.187168   0.132116   1.417  0.15657
## initial_list_statusw                -0.038949   0.023532  -1.655  0.09789
##
## (Intercept)                         ***
## emp_length1 year                    .
## emp_length10+ years
## emp_length2 years
## emp_length3 years
## emp_length4 years
## emp_length5 years
## emp_length6 years
## emp_length7 years
## emp_length8 years
## emp_length9 years
## emp_lengthn/a                       ***
## home_ownershipOWN                   ***
## home_ownershipRENT                  ***
## verification_statusSource Verified .
## verification_statusVerified         ***
## purposecredit_card
## purposedebt_consolidation           **
## purposehome_improvement             *
## purposehouse
## purposemajor_purchase
## purposemedical                      ***
## purposemoving                       *
## purposeother                        ***
## purposerenewable_energy             *
## purposesmall_business               ***
## purposevacation                     **
## purposewedding
## initial_list_statusw                .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80431  on 104680  degrees of freedom
## Residual deviance: 79538  on 104652  degrees of freedom
## AIC: 79596
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
stepwise.cat<-suppressWarnings(step(logistic.cat))
```

```
## Start:  AIC=79596.24
## defaulted ~ emp_length + home_ownership + verification_status +
##     purpose + initial_list_status
##
##                         Df Deviance   AIC
## <none>                        79538 79596
## - initial_list_status  1     79541 79597
## - verification_status  2     79557 79611
## - emp_length          11     79651 79687
## - purpose             12     79835 79869
## - home_ownership       2     79940 79994
```

```
summary(stepwise.cat)
```

```
##
## Call:
## glm(formula = defaulted ~ emp_length + home_ownership + verification_status +
##     purpose + initial_list_status, family = binomial(link = logit),
##     data = cat.var)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9150  -0.5614  -0.4932  -0.4533   2.2741
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -2.410840   0.098474 -24.482  < 2e-16
## emp_length1 year                -0.090107   0.049932  -1.805  0.07114
## emp_length10+ years             -0.046213   0.038153  -1.211  0.22580
## emp_length2 years               -0.034609   0.045506  -0.761  0.44693
## emp_length3 years               -0.025435   0.047161  -0.539  0.58966
## emp_length4 years               -0.029978   0.050462  -0.594  0.55247
## emp_length5 years               -0.022126   0.047188  -0.469  0.63914
## emp_length6 years                0.045806   0.048953   0.936  0.34942
## emp_length7 years                0.058339   0.050512   1.155  0.24811
## emp_length8 years                0.020540   0.054301   0.378  0.70524
## emp_length9 years               -0.004914   0.059288  -0.083  0.93394
## emp_lengthn/a                    0.386075   0.051064   7.561 4.01e-14
## home_ownershipOWN                0.240516   0.034800   6.911 4.80e-12
## home_ownershipRENT               0.406747   0.020433  19.906  < 2e-16
## verification_statusSource Verified  0.048983   0.025576   1.915  0.05547
## verification_statusVerified      0.093359   0.021572   4.328 1.51e-05
## purposecredit_card               0.063133   0.093758   0.673  0.50072
## purposedebt_consolidation        0.277160   0.092360   3.001  0.00269
## purposehome_improvement          0.216345   0.100658   2.149  0.03161
## purposehouse                     0.199790   0.148175   1.348  0.17755
## purposemajor_purchase           -0.006499   0.114605  -0.057  0.95478
## purposemedical                   0.574879   0.126411   4.548 5.42e-06
## purposemoving                    0.277807   0.138970   1.999  0.04560
## purposeother                     0.564975   0.098575   5.731 9.96e-09
```

```
## purposerenewable_energy                 0.754606   0.315484    2.392  0.01676
## purposesmall_business                    0.870562   0.110085    7.908 2.61e-15
## purposevacation                          0.423029   0.142851    2.961  0.00306
## purposewedding                           0.187168   0.132116    1.417  0.15657
## initial_list_statusw                    -0.038949   0.023532   -1.655  0.09789
##
## (Intercept)                           ***
## emp_length1 year                        .
## emp_length10+ years
## emp_length2 years
## emp_length3 years
## emp_length4 years
## emp_length5 years
## emp_length6 years
## emp_length7 years
## emp_length8 years
## emp_length9 years
## emp_lengthn/a                         ***
## home_ownershipOWN                     ***
## home_ownershipRENT                    ***
## verification_statusSource Verified .
## verification_statusVerified           ***
## purposecredit_card
## purposedebt_consolidation             **
## purposehome_improvement               *
## purposehouse
## purposemajor_purchase
## purposemedical                        ***
## purposemoving                         *
## purposeother                          ***
## purposerenewable_energy               *
## purposesmall_business                 ***
## purposevacation                       **
## purposewedding
## initial_list_statusw                    .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80431  on 104680  degrees of freedom
## Residual deviance: 79538  on 104652  degrees of freedom
## AIC: 79596
##
## Number of Fisher Scoring iterations: 4
```

```r
# according to the stepwise model, we decide that the employment length does not matter as much as wheth
levels(loan.data$emp_length)
```

```
##  [1] "< 1 year"  "1 year"    "10+ years" "2 years"   "3 years"
##  [6] "4 years"   "5 years"   "6 years"   "7 years"   "8 years"
## [11] "9 years"   "n/a"
```

```r
levels(loan.data$emp_length)[c(1:11)]<-"Employed"
table(loan.data$emp_length)
```

```
##
## Employed          n/a
##     99869      4812
```

Numeric variables

```
# Checking correlation between numeric variables:
num.var<-loan.data[,cat!="factor"]
correlation<-round(cor(num.var),3)

logistic.full<-glm(defaulted~.,data=loan.data,family=binomial(link=logit))
```

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

```
summary(logistic.full)
```

```
##
## Call:
## glm(formula = defaulted ~ ., family = binomial(link = logit),
##     data = loan.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3794  -0.5701  -0.4688  -0.3566   3.6796
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.783e+00  2.278e-01 -12.217  < 2e-16
## emp_lengthn/a                   3.138e-01  4.161e-02   7.541 4.66e-14
## home_ownershipOWN               1.118e-01  3.703e-02   3.018 0.002541
## home_ownershipRENT              2.562e-01  2.461e-02  10.411  < 2e-16
## annual_inc                     -4.448e-06  4.016e-07 -11.078  < 2e-16
## verification_statusSource Verified 1.176e-01  2.616e-02   4.497 6.89e-06
## verification_statusVerified     1.075e-01  2.308e-02   4.657 3.20e-06
## purposecredit_card             -8.703e-02  9.548e-02  -0.911 0.362040
## purposedebt_consolidation       1.009e-01  9.399e-02   1.074 0.282934
## purposehome_improvement         1.754e-01  1.023e-01   1.715 0.086425
## purposehouse                    2.328e-01  1.509e-01   1.543 0.122722
## purposemajor_purchase          -1.490e-02  1.164e-01  -0.128 0.898094
## purposemedical                  4.350e-01  1.286e-01   3.381 0.000722
## purposemoving                   1.536e-01  1.412e-01   1.087 0.276819
## purposeother                    4.028e-01  1.002e-01   4.019 5.84e-05
## purposerenewable_energy         7.966e-01  3.223e-01   2.472 0.013453
## purposesmall_business           8.849e-01  1.122e-01   7.889 3.05e-15
## purposevacation                 1.723e-01  1.451e-01   1.187 0.235200
## purposewedding                  7.536e-02  1.342e-01   0.561 0.574507
## dti                             1.607e-02  1.639e-03   9.807  < 2e-16
## delinq_2yrs                     6.720e-02  1.849e-02   3.634 0.000279
## inq_last_6mths                  1.344e-01  1.130e-02  11.898  < 2e-16
## mths_since_last_delinq         -1.384e-03  6.301e-04  -2.197 0.028007
## mths_since_last_record          1.577e-03  8.418e-04   1.873 0.061061
## open_acc                        1.382e-02  5.393e-03   2.563 0.010381
## pub_rec                         1.201e-01  9.299e-02   1.292 0.196403
## revol_bal                      -1.569e-06  1.324e-06  -1.185 0.235936
## revol_util                      7.193e-01  7.034e-02  10.227  < 2e-16
## total_acc                      -1.076e-02  1.999e-03  -5.381 7.39e-08
```

```
## initial_list_statusw                  -1.776e-02  2.409e-02  -0.737 0.461020
## collections_12_mths_ex_med              9.310e-02  1.664e-01   0.560 0.575789
## mths_since_last_major_derog            -7.669e-04  6.938e-04  -1.105 0.269002
## acc_now_delinq                          1.783e-02  2.736e-01   0.065 0.948058
## tot_coll_amt                            1.280e-05  1.096e-05   1.167 0.243026
## tot_cur_bal                             1.492e-06  6.972e-07   2.140 0.032318
## total_rev_hi_lim                        1.899e-06  6.552e-07   2.899 0.003740
## acc_open_past_24mths                    7.349e-02  4.814e-03  15.267  < 2e-16
## avg_cur_bal                            -4.323e-06  1.836e-06  -2.355 0.018506
## bc_open_to_buy                         -1.036e-05  2.645e-06  -3.915 9.04e-05
## bc_util                                -3.548e-03  8.480e-04  -4.184 2.87e-05
## chargeoff_within_12_mths               -4.888e-02  1.300e-01  -0.376 0.706925
## delinq_amnt                             3.534e-06  1.750e-05   0.202 0.839959
## mo_sin_old_il_acct                      3.466e-04  1.702e-04   2.037 0.041675
## mo_sin_old_rev_tl_op                   -4.598e-04  1.271e-04  -3.618 0.000297
## mo_sin_rcnt_rev_tl_op                   1.253e-03  9.493e-04   1.320 0.186950
## mo_sin_rcnt_tl                         -2.407e-03  1.524e-03  -1.579 0.114343
## mort_acc                               -2.150e-03  6.547e-03  -0.328 0.742573
## mths_since_recent_bc                   -2.120e-03  4.292e-04  -4.940 7.83e-07
## mths_since_recent_bc_dlq                2.672e-04  6.975e-04   0.383 0.701682
## mths_since_recent_inq                  -7.580e-03  1.767e-03  -4.289 1.80e-05
## mths_since_recent_revol_delinq         -3.766e-04  7.288e-04  -0.517 0.605349
## num_accts_ever_120_pd                  -2.398e-02  1.406e-02  -1.706 0.088014
## num_actv_bc_tl                         -5.600e-02  1.170e-02  -4.786 1.70e-06
## num_actv_rev_tl                        -2.556e-02  6.011e-02  -0.425 0.670659
## num_bc_sats                             4.240e-02  9.034e-03   4.693 2.69e-06
## num_bc_tl                               9.666e-03  5.658e-03   1.708 0.087586
## num_il_tl                              -6.840e-03  2.479e-03  -2.759 0.005798
## num_op_rev_tl                          -6.187e-03  7.456e-03  -0.830 0.406665
## num_rev_accts                          -4.389e-03  4.515e-03  -0.972 0.330999
## num_rev_tl_bal_gt_0                     7.013e-02  5.984e-02   1.172 0.241200
## num_sats                               -2.118e-02  5.940e-03  -3.566 0.000363
## num_tl_120dpd_2m                        3.213e-01  4.658e-01   0.690 0.490347
## num_tl_30dpd                            6.654e-01  2.727e-01   2.440 0.014676
## num_tl_90g_dpd_24m                      6.394e-02  3.113e-02   2.054 0.039959
## num_tl_op_past_12m                      3.336e-02  8.244e-03   4.047 5.19e-05
## pct_tl_nvr_dlq                          1.702e-03  1.838e-03   0.926 0.354401
## percent_bc_gt_75                        3.730e-03  5.093e-04   7.325 2.39e-13
## pub_rec_bankruptcies                   -1.117e-01  9.416e-02  -1.186 0.235571
## tax_liens                              -9.310e-02  9.725e-02  -0.957 0.338394
## tot_hi_cred_lim                        -1.413e-06  6.463e-07  -2.186 0.028817
## total_bal_ex_mort                      -1.886e-07  5.600e-07  -0.337 0.736303
## total_bc_limit                         -2.166e-06  1.743e-06  -1.243 0.213840
## total_il_high_credit_limit             -5.705e-07  5.688e-07  -1.003 0.315910
##
## (Intercept)                            ***
## emp_lengthn/a                          ***
## home_ownershipOWN                      **
## home_ownershipRENT                     ***
## annual_inc                             ***
## verification_statusSource Verified ***
## verification_statusVerified            ***
## purposecredit_card
## purposedebt_consolidation
```

```
## purposehome_improvement            .
## purposehouse
## purposemajor_purchase
## purposemedical                     ***
## purposemoving
## purposeother                       ***
## purposerenewable_energy            *
## purposesmall_business              ***
## purposevacation
## purposewedding
## dti                                ***
## delinq_2yrs                        ***
## inq_last_6mths                     ***
## mths_since_last_delinq             *
## mths_since_last_record             .
## open_acc                           *
## pub_rec
## revol_bal
## revol_util                         ***
## total_acc                          ***
## initial_list_statusw
## collections_12_mths_ex_med
## mths_since_last_major_derog
## acc_now_delinq
## tot_coll_amt
## tot_cur_bal                        *
## total_rev_hi_lim                   **
## acc_open_past_24mths               ***
## avg_cur_bal                        *
## bc_open_to_buy                     ***
## bc_util                            ***
## chargeoff_within_12_mths
## delinq_amnt
## mo_sin_old_il_acct                 *
## mo_sin_old_rev_tl_op               ***
## mo_sin_rcnt_rev_tl_op
## mo_sin_rcnt_tl
## mort_acc
## mths_since_recent_bc               ***
## mths_since_recent_bc_dlq
## mths_since_recent_inq              ***
## mths_since_recent_revol_delinq
## num_accts_ever_120_pd              .
## num_actv_bc_tl                     ***
## num_actv_rev_tl
## num_bc_sats                        ***
## num_bc_tl                          .
## num_il_tl                          **
## num_op_rev_tl
## num_rev_accts
## num_rev_tl_bal_gt_0
## num_sats                           ***
## num_tl_120dpd_2m
## num_tl_30dpd                       *
```

```
## num_tl_90g_dpd_24m                         *
## num_tl_op_past_12m                         ***
## pct_tl_nvr_dlq
## percent_bc_gt_75                           ***
## pub_rec_bankruptcies
## tax_liens
## tot_hi_cred_lim                            *
## total_bal_ex_mort
## total_bc_limit
## total_il_high_credit_limit
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80431  on 104680  degrees of freedom
## Residual deviance: 76645  on 104608  degrees of freedom
## AIC: 76791
##
## Number of Fisher Scoring iterations: 5
```

logistic.full$coefficients

```
##                     (Intercept)                         emp_lengthn/a
##                   -2.782622e+00                          3.137576e-01
##                home_ownershipOWN                     home_ownershipRENT
##                    1.117861e-01                          2.561667e-01
##                      annual_inc   verification_statusSource Verified
##                   -4.448469e-06                          1.176388e-01
##        verification_statusVerified                     purposecredit_card
##                    1.074740e-01                         -8.703258e-02
##          purposedebt_consolidation                purposehome_improvement
##                    1.009244e-01                          1.754196e-01
##                    purposehouse                     purposemajor_purchase
##                    2.328492e-01                         -1.490239e-02
##                  purposemedical                            purposemoving
##                    4.349597e-01                          1.535618e-01
##                   purposeother                    purposerenewable_energy
##                    4.027606e-01                          7.965919e-01
##           purposesmall_business                         purposevacation
##                    8.848634e-01                          1.722917e-01
##                 purposewedding                                       dti
##                    7.535714e-02                          1.607499e-02
##                    delinq_2yrs                          inq_last_6mths
##                    6.719955e-02                          1.344314e-01
##         mths_since_last_delinq              mths_since_last_record
##                   -1.384397e-03                          1.576739e-03
##                       open_acc                                 pub_rec
##                    1.382240e-02                          1.201310e-01
##                       revol_bal                               revol_util
##                   -1.569427e-06                          7.193319e-01
##                       total_acc                      initial_list_statusw
##                   -1.075960e-02                         -1.775893e-02
##        collections_12_mths_ex_med         mths_since_last_major_derog
##                    9.309681e-02                         -7.668658e-04
```

```
##              acc_now_delinq                          tot_coll_amt
##                1.782612e-02                          1.279968e-05
##                 tot_cur_bal                       total_rev_hi_lim
##                1.492249e-06                          1.899520e-06
##         acc_open_past_24mths                           avg_cur_bal
##                7.348942e-02                         -4.323344e-06
##               bc_open_to_buy                               bc_util
##               -1.035620e-05                         -3.547533e-03
##       chargeoff_within_12_mths                          delinq_amnt
##               -4.887663e-02                          3.534344e-06
##             mo_sin_old_il_acct                     mo_sin_old_rev_tl_op
##                3.465726e-04                         -4.598118e-04
##           mo_sin_rcnt_rev_tl_op                         mo_sin_rcnt_tl
##                1.252686e-03                         -2.407025e-03
##                     mort_acc                     mths_since_recent_bc
##               -2.150444e-03                         -2.119874e-03
##        mths_since_recent_bc_dlq                  mths_since_recent_inq
##                2.671764e-04                         -7.579797e-03
##    mths_since_recent_revol_delinq                 num_accts_ever_120_pd
##               -3.765850e-04                         -2.398050e-02
##                 num_actv_bc_tl                        num_actv_rev_tl
##               -5.599653e-02                         -2.556268e-02
##                  num_bc_sats                               num_bc_tl
##                4.239773e-02                          9.666119e-03
##                    num_il_tl                             num_op_rev_tl
##               -6.840423e-03                         -6.187165e-03
##                 num_rev_accts                        num_rev_tl_bal_gt_0
##               -4.388886e-03                          7.012798e-02
##                     num_sats                          num_tl_120dpd_2m
##               -2.118232e-02                          3.213212e-01
##                  num_tl_30dpd                         num_tl_90g_dpd_24m
##                6.653541e-01                          6.394225e-02
##              num_tl_op_past_12m                          pct_tl_nvr_dlq
##                3.336380e-02                          1.702183e-03
##                percent_bc_gt_75                      pub_rec_bankruptcies
##                3.730439e-03                         -1.116805e-01
##                     tax_liens                           tot_hi_cred_lim
##               -9.310460e-02                         -1.412854e-06
##               total_bal_ex_mort                          total_bc_limit
##               -1.885872e-07                         -2.166333e-06
##         total_il_high_credit_limit
##               -5.704914e-07
```

```r
# Remove other insignificant variables
to.remove<-names(summary(logistic.full)$coefficients[,1][round(summary(logistic.full)$coefficients,2)[,
to.remove
```

```
## [1] "mths_since_last_delinq"       "mths_since_last_record"
## [3] "revol_bal"                    "mths_since_last_major_derog"
## [5] "tot_coll_amt"                 "tot_cur_bal"
## [7] "avg_cur_bal"                  "delinq_amnt"
## [9] "mo_sin_old_il_acct"           "mo_sin_rcnt_rev_tl_op"
## [11] "mo_sin_rcnt_tl"              "mort_acc"
## [13] "mths_since_recent_bc_dlq"    "mths_since_recent_revol_delinq"
## [15] "num_rev_accts"               "pct_tl_nvr_dlq"
```

```
## [17] "tot_hi_cred_lim"              "total_bal_ex_mort"
## [19] "total_bc_limit"              "total_il_high_credit_limit"
```

```
model.data<-loan.data[,-which(colnames(loan.data) %in% to.remove)]
colnames(model.data)
```

```
##  [1] "emp_length"                "home_ownership"
##  [3] "annual_inc"                "verification_status"
##  [5] "purpose"                   "dti"
##  [7] "delinq_2yrs"               "inq_last_6mths"
##  [9] "open_acc"                  "pub_rec"
## [11] "revol_util"                "total_acc"
## [13] "initial_list_status"       "collections_12_mths_ex_med"
## [15] "acc_now_delinq"            "total_rev_hi_lim"
## [17] "acc_open_past_24mths"      "bc_open_to_buy"
## [19] "bc_util"                   "chargeoff_within_12_mths"
## [21] "mo_sin_old_rev_tl_op"      "mths_since_recent_bc"
## [23] "mths_since_recent_inq"     "num_accts_ever_120_pd"
## [25] "num_actv_bc_tl"            "num_actv_rev_tl"
## [27] "num_bc_sats"               "num_bc_tl"
## [29] "num_il_tl"                 "num_op_rev_tl"
## [31] "num_rev_tl_bal_gt_0"       "num_sats"
## [33] "num_tl_120dpd_2m"          "num_tl_30dpd"
## [35] "num_tl_90g_dpd_24m"        "num_tl_op_past_12m"
## [37] "percent_bc_gt_75"          "pub_rec_bankruptcies"
## [39] "tax_liens"                 "defaulted"
```

Split data into train and test:

```
# Take 63.2% for training and the rest for testing:
n=nrow(model.data)
set.seed(2345341)
rownumbers<-sample(seq_len(nrow(model.data)),size=n*0.632)
train<-model.data[rownumbers,]
test<-model.data[-rownumbers,]
```

Logistic Regression:

```
log.model.full<-glm(defaulted~.,data=train,family=binomial(link=logit))
```

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

```
summary(log.model.full)
```

```
##
## Call:
## glm(formula = defaulted ~ ., family = binomial(link = logit),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8003  -0.5705  -0.4714  -0.3615   3.7738
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -2.430e+00  1.373e-01 -17.695  < 2e-16
## emp_lengthn/a                   3.291e-01  5.182e-02   6.351 2.13e-10
```

```
## home_ownershipOWN                   1.143e-01  4.497e-02   2.541 0.011040
## home_ownershipRENT                  2.591e-01  2.711e-02   9.558  < 2e-16
## annual_inc                         -5.209e-06  4.451e-07 -11.704  < 2e-16
## verification_statusSource Verified  1.271e-01  3.274e-02   3.883 0.000103
## verification_statusVerified         8.614e-02  2.886e-02   2.985 0.002838
## purposecredit_card                 -1.842e-01  1.203e-01  -1.531 0.125665
## purposedebt_consolidation           2.208e-02  1.184e-01   0.187 0.852013
## purposehome_improvement             8.353e-02  1.287e-01   0.649 0.516357
## purposehouse                        1.967e-01  1.870e-01   1.052 0.292921
## purposemajor_purchase              -1.621e-02  1.450e-01  -0.112 0.911015
## purposemedical                      2.186e-01  1.646e-01   1.328 0.184177
## purposemoving                       2.249e-02  1.791e-01   0.126 0.900024
## purposeother                        2.842e-01  1.262e-01   2.251 0.024360
## purposerenewable_energy             8.699e-01  3.914e-01   2.222 0.026258
## purposesmall_business               7.959e-01  1.408e-01   5.654 1.57e-08
## purposevacation                     8.542e-02  1.797e-01   0.475 0.634533
## purposewedding                     -3.286e-02  1.690e-01  -0.194 0.845814
## dti                                 1.409e-02  1.882e-03   7.486 7.08e-14
## delinq_2yrs                         1.029e-01  1.983e-02   5.190 2.11e-07
## inq_last_6mths                      1.356e-01  1.403e-02   9.659  < 2e-16
## open_acc                            1.752e-02  6.471e-03   2.707 0.006793
## pub_rec                             5.734e-02  1.095e-01   0.524 0.600460
## revol_util                          7.076e-01  8.616e-02   8.212  < 2e-16
## total_acc                          -1.087e-02  2.138e-03  -5.083 3.71e-07
## initial_list_statusw               -1.160e-02  2.996e-02  -0.387 0.698695
## collections_12_mths_ex_med          1.979e-01  2.008e-01   0.986 0.324271
## acc_now_delinq                     -3.982e-01  3.636e-01  -1.095 0.273482
## total_rev_hi_lim                    2.125e-07  6.827e-07   0.311 0.755546
## acc_open_past_24mths                7.367e-02  5.973e-03  12.333  < 2e-16
## bc_open_to_buy                     -1.548e-05  1.959e-06  -7.903 2.73e-15
## bc_util                            -5.032e-03  1.030e-03  -4.883 1.04e-06
## chargeoff_within_12_mths           -1.270e-01  1.796e-01  -0.707 0.479525
## mo_sin_old_rev_tl_op               -3.794e-04  1.538e-04  -2.466 0.013651
## mths_since_recent_bc               -2.195e-03  4.926e-04  -4.457 8.31e-06
## mths_since_recent_inq              -7.557e-03  2.192e-03  -3.448 0.000565
## num_accts_ever_120_pd              -5.216e-03  1.390e-02  -0.375 0.707393
## num_actv_bc_tl                     -5.842e-02  1.366e-02  -4.275 1.91e-05
## num_actv_rev_tl                     1.122e-02  7.663e-02   0.146 0.883620
## num_bc_sats                         5.901e-02  1.063e-02   5.551 2.83e-08
## num_bc_tl                           6.820e-04  4.047e-03   0.169 0.866182
## num_il_tl                          -5.881e-03  2.622e-03  -2.243 0.024876
## num_op_rev_tl                      -4.839e-03  7.813e-03  -0.619 0.535724
## num_rev_tl_bal_gt_0                 3.578e-02  7.635e-02   0.469 0.639328
## num_sats                           -2.795e-02  7.092e-03  -3.941 8.12e-05
## num_tl_120dpd_2m                    6.350e-01  5.987e-01   1.061 0.288828
## num_tl_30dpd                        1.001e+00  3.411e-01   2.935 0.003339
## num_tl_90g_dpd_24m                  6.174e-02  3.672e-02   1.681 0.092689
## num_tl_op_past_12m                  3.072e-02  9.571e-03   3.210 0.001329
## percent_bc_gt_75                    4.227e-03  6.378e-04   6.627 3.43e-11
## pub_rec_bankruptcies               -1.471e-01  1.175e-01  -1.252 0.210579
## tax_liens                          -2.166e-02  1.336e-01  -0.162 0.871214
##
## (Intercept)                        ***
## emp_lengthn/a                      ***
```

```
## home_ownershipOWN                     *
## home_ownershipRENT                    ***
## annual_inc                            ***
## verification_statusSource Verified    ***
## verification_statusVerified           **
## purposecredit_card
## purposedebt_consolidation
## purposehome_improvement
## purposehouse
## purposemajor_purchase
## purposemedical
## purposemoving
## purposeother                          *
## purposerenewable_energy               *
## purposesmall_business                 ***
## purposevacation
## purposewedding
## dti                                   ***
## delinq_2yrs                           ***
## inq_last_6mths                        ***
## open_acc                              **
## pub_rec
## revol_util                            ***
## total_acc                             ***
## initial_list_statusw
## collections_12_mths_ex_med
## acc_now_delinq
## total_rev_hi_lim
## acc_open_past_24mths                  ***
## bc_open_to_buy                        ***
## bc_util                               ***
## chargeoff_within_12_mths
## mo_sin_old_rev_tl_op                  *
## mths_since_recent_bc                  ***
## mths_since_recent_inq                 ***
## num_accts_ever_120_pd
## num_actv_bc_tl                        ***
## num_actv_rev_tl
## num_bc_sats                           ***
## num_bc_tl
## num_il_tl                             *
## num_op_rev_tl
## num_rev_tl_bal_gt_0
## num_sats                              ***
## num_tl_120dpd_2m
## num_tl_30dpd                          **
## num_tl_90g_dpd_24m                    .
## num_tl_op_past_12m                    **
## percent_bc_gt_75                      ***
## pub_rec_bankruptcies
## tax_liens
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 50963  on 66157  degrees of freedom
## Residual deviance: 48647  on 66105  degrees of freedom
## AIC: 48753
##
## Number of Fisher Scoring iterations: 5
```

```r
stepwise<-suppressWarnings(step(log.model.full))
```

```
## Start:  AIC=48752.51
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##      purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##      pub_rec + revol_util + total_acc + initial_list_status +
##      collections_12_mths_ex_med + acc_now_delinq + total_rev_hi_lim +
##      acc_open_past_24mths + bc_open_to_buy + bc_util + chargeoff_within_12_mths +
##      mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##      num_accts_ever_120_pd + num_actv_bc_tl + num_actv_rev_tl +
##      num_bc_sats + num_bc_tl + num_il_tl + num_op_rev_tl + num_rev_tl_bal_gt_0 +
##      num_sats + num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m +
##      num_tl_op_past_12m + percent_bc_gt_75 + pub_rec_bankruptcies +
##      tax_liens
##
##                              Df Deviance    AIC
## - num_actv_rev_tl             1    48647 48751
## - tax_liens                   1    48647 48751
## - num_bc_tl                   1    48647 48751
## - total_rev_hi_lim            1    48647 48751
## - num_accts_ever_120_pd       1    48647 48751
## - initial_list_status         1    48647 48751
## - num_rev_tl_bal_gt_0         1    48647 48751
## - pub_rec                     1    48647 48751
## - num_op_rev_tl               1    48647 48751
## - chargeoff_within_12_mths    1    48647 48751
## - collections_12_mths_ex_med  1    48647 48751
## - num_tl_120dpd_2m            1    48648 48752
## - acc_now_delinq              1    48648 48752
## - pub_rec_bankruptcies        1    48648 48752
## <none>                             48647 48753
## - num_tl_90g_dpd_24m          1    48649 48753
## - num_il_tl                   1    48652 48756
## - mo_sin_old_rev_tl_op        1    48653 48757
## - open_acc                    1    48654 48758
## - num_tl_30dpd                1    48654 48758
## - num_tl_op_past_12m          1    48657 48761
## - mths_since_recent_inq       1    48658 48762
## - verification_status         2    48664 48766
## - num_sats                    1    48662 48766
## - num_actv_bc_tl              1    48665 48769
## - mths_since_recent_bc        1    48667 48771
## - bc_util                     1    48670 48774
## - delinq_2yrs                 1    48672 48776
## - total_acc                   1    48673 48777
## - num_bc_sats                 1    48677 48781
## - emp_length                  1    48685 48789
```

```
## - percent_bc_gt_75           1    48691 48795
## - dti                        1    48703 48807
## - bc_open_to_buy             1    48713 48817
## - revol_util                1    48714 48818
## - home_ownership            2    48739 48841
## - inq_last_6mths            1    48737 48841
## - acc_open_past_24mths       1    48795 48899
## - annual_inc                1    48799 48903
## - purpose                   12    48832 48914
##
## Step:  AIC=48750.53
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + initial_list_status +
##     collections_12_mths_ex_med + acc_now_delinq + total_rev_hi_lim +
##     acc_open_past_24mths + bc_open_to_buy + bc_util + chargeoff_within_12_mths +
##     mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_accts_ever_120_pd + num_actv_bc_tl + num_bc_sats + num_bc_tl +
##     num_il_tl + num_op_rev_tl + num_rev_tl_bal_gt_0 + num_sats +
##     num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies + tax_liens
##
##                              Df Deviance    AIC
## - tax_liens                   1    48647 48749
## - num_bc_tl                   1    48647 48749
## - total_rev_hi_lim            1    48647 48749
## - num_accts_ever_120_pd       1    48647 48749
## - initial_list_status         1    48647 48749
## - pub_rec                     1    48647 48749
## - num_op_rev_tl               1    48647 48749
## - chargeoff_within_12_mths    1    48647 48749
## - collections_12_mths_ex_med  1    48647 48749
## - num_tl_120dpd_2m            1    48648 48750
## - acc_now_delinq              1    48648 48750
## - pub_rec_bankruptcies        1    48648 48750
## <none>                             48647 48751
## - num_tl_90g_dpd_24m          1    48649 48751
## - num_il_tl                   1    48652 48754
## - mo_sin_old_rev_tl_op        1    48653 48755
## - open_acc                    1    48654 48756
## - num_tl_30dpd                1    48654 48756
## - num_tl_op_past_12m          1    48657 48759
## - mths_since_recent_inq       1    48659 48761
## - verification_status         2    48664 48764
## - num_sats                    1    48662 48764
## - num_actv_bc_tl              1    48665 48767
## - mths_since_recent_bc        1    48667 48769
## - num_rev_tl_bal_gt_0         1    48670 48772
## - bc_util                     1    48670 48772
## - delinq_2yrs                 1    48672 48774
## - total_acc                   1    48673 48775
## - num_bc_sats                 1    48677 48779
## - emp_length                  1    48685 48787
## - percent_bc_gt_75            1    48691 48793
```

```
## - dti                          1    48703 48805
## - bc_open_to_buy               1    48713 48815
## - revol_util                   1    48714 48816
## - home_ownership               2    48739 48839
## - inq_last_6mths               1    48737 48839
## - acc_open_past_24mths         1    48795 48897
## - annual_inc                   1    48800 48902
## - purpose                     12    48832 48912
##
## Step:  AIC=48748.56
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + initial_list_status +
##     collections_12_mths_ex_med + acc_now_delinq + total_rev_hi_lim +
##     acc_open_past_24mths + bc_open_to_buy + bc_util + chargeoff_within_12_mths +
##     mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_accts_ever_120_pd + num_actv_bc_tl + num_bc_sats + num_bc_tl +
##     num_il_tl + num_op_rev_tl + num_rev_tl_bal_gt_0 + num_sats +
##     num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies
##
##                               Df Deviance    AIC
## - num_bc_tl                    1    48647 48747
## - total_rev_hi_lim             1    48647 48747
## - num_accts_ever_120_pd        1    48647 48747
## - initial_list_status          1    48647 48747
## - num_op_rev_tl                1    48647 48747
## - pub_rec                      1    48647 48747
## - chargeoff_within_12_mths     1    48647 48747
## - collections_12_mths_ex_med   1    48647 48747
## - num_tl_120dpd_2m             1    48648 48748
## - acc_now_delinq               1    48648 48748
## <none>                              48647 48749
## - num_tl_90g_dpd_24m           1    48649 48749
## - pub_rec_bankruptcies         1    48650 48750
## - num_il_tl                    1    48652 48752
## - mo_sin_old_rev_tl_op         1    48653 48753
## - open_acc                     1    48654 48754
## - num_tl_30dpd                 1    48654 48754
## - num_tl_op_past_12m           1    48657 48757
## - mths_since_recent_inq        1    48659 48759
## - verification_status          2    48664 48762
## - num_sats                     1    48662 48762
## - num_actv_bc_tl               1    48665 48765
## - mths_since_recent_bc         1    48667 48767
## - num_rev_tl_bal_gt_0          1    48670 48770
## - bc_util                      1    48670 48770
## - delinq_2yrs                  1    48672 48772
## - total_acc                    1    48673 48773
## - num_bc_sats                  1    48677 48777
## - emp_length                   1    48685 48785
## - percent_bc_gt_75             1    48691 48791
## - dti                          1    48703 48803
## - bc_open_to_buy               1    48713 48813
```

```
## - revol_util                      1      48714 48814
## - home_ownership                  2      48739 48837
## - inq_last_6mths                  1      48737 48837
## - acc_open_past_24mths            1      48795 48895
## - annual_inc                      1      48800 48900
## - purpose                         12     48833 48911
##
## Step:  AIC=48746.59
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + initial_list_status +
##     collections_12_mths_ex_med + acc_now_delinq + total_rev_hi_lim +
##     acc_open_past_24mths + bc_open_to_buy + bc_util + chargeoff_within_12_mths +
##     mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_accts_ever_120_pd + num_actv_bc_tl + num_bc_sats + num_il_tl +
##     num_op_rev_tl + num_rev_tl_bal_gt_0 + num_sats + num_tl_120dpd_2m +
##     num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies
##
##                                 Df Deviance    AIC
## - total_rev_hi_lim               1      48647 48745
## - num_accts_ever_120_pd          1      48647 48745
## - initial_list_status            1      48647 48745
## - num_op_rev_tl                  1      48647 48745
## - pub_rec                        1      48647 48745
## - chargeoff_within_12_mths       1      48647 48745
## - collections_12_mths_ex_med     1      48648 48746
## - num_tl_120dpd_2m               1      48648 48746
## - acc_now_delinq                 1      48648 48746
## <none>                                  48647 48747
## - num_tl_90g_dpd_24m             1      48649 48747
## - pub_rec_bankruptcies           1      48650 48748
## - num_il_tl                      1      48652 48750
## - mo_sin_old_rev_tl_op           1      48653 48751
## - open_acc                       1      48654 48752
## - num_tl_30dpd                   1      48654 48752
## - num_tl_op_past_12m             1      48657 48755
## - mths_since_recent_inq          1      48659 48757
## - verification_status            2      48664 48760
## - num_sats                       1      48662 48760
## - num_actv_bc_tl                 1      48666 48764
## - mths_since_recent_bc           1      48667 48765
## - num_rev_tl_bal_gt_0            1      48670 48768
## - bc_util                        1      48670 48768
## - delinq_2yrs                    1      48672 48770
## - num_bc_sats                    1      48678 48776
## - total_acc                      1      48680 48778
## - emp_length                     1      48685 48783
## - percent_bc_gt_75               1      48691 48789
## - dti                            1      48703 48801
## - bc_open_to_buy                 1      48713 48811
## - revol_util                     1      48715 48813
## - inq_last_6mths                 1      48737 48835
## - home_ownership                 2      48739 48835
```

```
## - acc_open_past_24mths          1    48795 48893
## - annual_inc                     1    48800 48898
## - purpose                       12    48833 48909
##
## Step:  AIC=48744.68
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + initial_list_status +
##     collections_12_mths_ex_med + acc_now_delinq + acc_open_past_24mths +
##     bc_open_to_buy + bc_util + chargeoff_within_12_mths + mo_sin_old_rev_tl_op +
##     mths_since_recent_bc + mths_since_recent_inq + num_accts_ever_120_pd +
##     num_actv_bc_tl + num_bc_sats + num_il_tl + num_op_rev_tl +
##     num_rev_tl_bal_gt_0 + num_sats + num_tl_120dpd_2m + num_tl_30dpd +
##     num_tl_90g_dpd_24m + num_tl_op_past_12m + percent_bc_gt_75 +
##     pub_rec_bankruptcies
##
##                               Df Deviance    AIC
## - num_accts_ever_120_pd        1    48647 48743
## - initial_list_status          1    48647 48743
## - num_op_rev_tl                1    48647 48743
## - pub_rec                      1    48647 48743
## - chargeoff_within_12_mths     1    48647 48743
## - collections_12_mths_ex_med   1    48648 48744
## - num_tl_120dpd_2m             1    48648 48744
## - acc_now_delinq               1    48648 48744
## <none>                              48647 48745
## - num_tl_90g_dpd_24m           1    48649 48745
## - pub_rec_bankruptcies         1    48650 48746
## - num_il_tl                    1    48652 48748
## - mo_sin_old_rev_tl_op         1    48653 48749
## - open_acc                     1    48654 48750
## - num_tl_30dpd                 1    48655 48751
## - num_tl_op_past_12m           1    48657 48753
## - mths_since_recent_inq        1    48659 48755
## - verification_status          2    48664 48758
## - num_sats                     1    48662 48758
## - num_actv_bc_tl               1    48666 48762
## - mths_since_recent_bc         1    48667 48763
## - num_rev_tl_bal_gt_0          1    48670 48766
## - bc_util                      1    48670 48766
## - delinq_2yrs                  1    48672 48768
## - num_bc_sats                  1    48678 48774
## - total_acc                    1    48680 48776
## - emp_length                   1    48685 48781
## - percent_bc_gt_75             1    48691 48787
## - dti                          1    48704 48800
## - revol_util                   1    48715 48811
## - bc_open_to_buy               1    48727 48823
## - inq_last_6mths               1    48737 48833
## - home_ownership               2    48739 48833
## - acc_open_past_24mths         1    48795 48891
## - annual_inc                   1    48807 48903
## - purpose                     12    48833 48907
##
```

```
## Step:  AIC=48742.83
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + initial_list_status +
##     collections_12_mths_ex_med + acc_now_delinq + acc_open_past_24mths +
##     bc_open_to_buy + bc_util + chargeoff_within_12_mths + mo_sin_old_rev_tl_op +
##     mths_since_recent_bc + mths_since_recent_inq + num_actv_bc_tl +
##     num_bc_sats + num_il_tl + num_op_rev_tl + num_rev_tl_bal_gt_0 +
##     num_sats + num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m +
##     num_tl_op_past_12m + percent_bc_gt_75 + pub_rec_bankruptcies
##
##                              Df Deviance    AIC
## - initial_list_status         1    48647  48741
## - num_op_rev_tl               1    48647  48741
## - pub_rec                     1    48647  48741
## - chargeoff_within_12_mths    1    48647  48741
## - collections_12_mths_ex_med  1    48648  48742
## - num_tl_120dpd_2m            1    48648  48742
## - acc_now_delinq              1    48648  48742
## <none>                             48647  48743
## - num_tl_90g_dpd_24m          1    48649  48743
## - pub_rec_bankruptcies        1    48650  48744
## - num_il_tl                   1    48652  48746
## - mo_sin_old_rev_tl_op        1    48653  48747
## - open_acc                    1    48654  48748
## - num_tl_30dpd                1    48655  48749
## - num_tl_op_past_12m          1    48657  48751
## - mths_since_recent_inq       1    48659  48753
## - verification_status         2    48664  48756
## - num_sats                    1    48662  48756
## - num_actv_bc_tl              1    48666  48760
## - mths_since_recent_bc        1    48667  48761
## - num_rev_tl_bal_gt_0         1    48670  48764
## - bc_util                     1    48670  48764
## - delinq_2yrs                 1    48672  48766
## - num_bc_sats                 1    48679  48773
## - total_acc                   1    48681  48775
## - emp_length                  1    48685  48779
## - percent_bc_gt_75            1    48691  48785
## - dti                         1    48704  48798
## - revol_util                  1    48715  48809
## - bc_open_to_buy              1    48728  48822
## - home_ownership              2    48739  48831
## - inq_last_6mths              1    48737  48831
## - acc_open_past_24mths        1    48796  48890
## - annual_inc                  1    48807  48901
## - purpose                    12    48833  48905
##
## Step:  AIC=48740.99
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + collections_12_mths_ex_med +
##     acc_now_delinq + acc_open_past_24mths + bc_open_to_buy +
##     bc_util + chargeoff_within_12_mths + mo_sin_old_rev_tl_op +
```

```
##      mths_since_recent_bc + mths_since_recent_inq + num_actv_bc_tl +
##      num_bc_sats + num_il_tl + num_op_rev_tl + num_rev_tl_bal_gt_0 +
##      num_sats + num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m +
##      num_tl_op_past_12m + percent_bc_gt_75 + pub_rec_bankruptcies
##
##                              Df Deviance   AIC
## - num_op_rev_tl               1    48647 48739
## - pub_rec                     1    48647 48739
## - chargeoff_within_12_mths    1    48648 48740
## - collections_12_mths_ex_med  1    48648 48740
## - num_tl_120dpd_2m            1    48648 48740
## - acc_now_delinq              1    48648 48740
## <none>                             48647 48741
## - num_tl_90g_dpd_24m          1    48650 48742
## - pub_rec_bankruptcies        1    48650 48742
## - num_il_tl                   1    48652 48744
## - mo_sin_old_rev_tl_op        1    48653 48745
## - open_acc                    1    48654 48746
## - num_tl_30dpd                1    48655 48747
## - num_tl_op_past_12m          1    48657 48749
## - mths_since_recent_inq       1    48659 48751
## - verification_status         2    48664 48754
## - num_sats                    1    48663 48755
## - num_actv_bc_tl              1    48666 48758
## - mths_since_recent_bc        1    48668 48760
## - num_rev_tl_bal_gt_0         1    48670 48762
## - bc_util                     1    48671 48763
## - delinq_2yrs                 1    48673 48765
## - num_bc_sats                 1    48679 48771
## - total_acc                   1    48681 48773
## - emp_length                  1    48685 48777
## - percent_bc_gt_75            1    48691 48783
## - dti                         1    48705 48797
## - revol_util                  1    48715 48807
## - bc_open_to_buy              1    48728 48820
## - inq_last_6mths              1    48738 48830
## - home_ownership              2    48740 48830
## - acc_open_past_24mths        1    48796 48888
## - annual_inc                  1    48807 48899
## - purpose                    12    48833 48903
##
## Step:  AIC=48739.28
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##      purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##      pub_rec + revol_util + total_acc + collections_12_mths_ex_med +
##      acc_now_delinq + acc_open_past_24mths + bc_open_to_buy +
##      bc_util + chargeoff_within_12_mths + mo_sin_old_rev_tl_op +
##      mths_since_recent_bc + mths_since_recent_inq + num_actv_bc_tl +
##      num_bc_sats + num_il_tl + num_rev_tl_bal_gt_0 + num_sats +
##      num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##      percent_bc_gt_75 + pub_rec_bankruptcies
##
##                              Df Deviance   AIC
## - pub_rec                     1    48648 48738
```

```
## - chargeoff_within_12_mths     1     48648 48738
## - collections_12_mths_ex_med   1     48648 48738
## - num_tl_120dpd_2m             1     48648 48738
## - acc_now_delinq               1     48649 48739
## <none>                               48647 48739
## - num_tl_90g_dpd_24m           1     48650 48740
## - pub_rec_bankruptcies         1     48650 48740
## - num_il_tl                    1     48652 48742
## - mo_sin_old_rev_tl_op         1     48654 48744
## - num_tl_30dpd                 1     48655 48745
## - open_acc                     1     48655 48745
## - num_tl_op_past_12m           1     48657 48747
## - mths_since_recent_inq        1     48659 48749
## - verification_status          2     48665 48753
## - num_actv_bc_tl               1     48666 48756
## - num_sats                     1     48667 48757
## - mths_since_recent_bc         1     48668 48758
## - bc_util                      1     48671 48761
## - delinq_2yrs                  1     48673 48763
## - num_rev_tl_bal_gt_0          1     48679 48769
## - num_bc_sats                  1     48679 48769
## - total_acc                    1     48683 48773
## - emp_length                   1     48686 48776
## - percent_bc_gt_75             1     48692 48782
## - dti                          1     48706 48796
## - revol_util                   1     48718 48808
## - bc_open_to_buy               1     48729 48819
## - home_ownership               2     48740 48828
## - inq_last_6mths               1     48738 48828
## - acc_open_past_24mths         1     48797 48887
## - annual_inc                   1     48808 48898
## - purpose                     12     48834 48902
##
## Step:  AIC=48737.75
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     revol_util + total_acc + collections_12_mths_ex_med + acc_now_delinq +
##     acc_open_past_24mths + bc_open_to_buy + bc_util + chargeoff_within_12_mths +
##     mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_actv_bc_tl + num_bc_sats + num_il_tl + num_rev_tl_bal_gt_0 +
##     num_sats + num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m +
##     num_tl_op_past_12m + percent_bc_gt_75 + pub_rec_bankruptcies
##
##                               Df Deviance   AIC
## - chargeoff_within_12_mths     1     48648 48736
## - collections_12_mths_ex_med   1     48649 48737
## - num_tl_120dpd_2m             1     48649 48737
## - acc_now_delinq               1     48649 48737
## <none>                               48648 48738
## - num_tl_90g_dpd_24m           1     48650 48738
## - pub_rec_bankruptcies         1     48653 48741
## - num_il_tl                    1     48653 48741
## - mo_sin_old_rev_tl_op         1     48654 48742
## - num_tl_30dpd                 1     48656 48744
```

```
## - open_acc                         1     48656 48744
## - num_tl_op_past_12m               1     48658 48746
## - mths_since_recent_inq            1     48660 48748
## - verification_status              2     48666 48752
## - num_actv_bc_tl                   1     48666 48754
## - num_sats                         1     48668 48756
## - mths_since_recent_bc             1     48668 48756
## - bc_util                          1     48672 48760
## - delinq_2yrs                      1     48673 48761
## - num_rev_tl_bal_gt_0              1     48679 48767
## - num_bc_sats                      1     48680 48768
## - total_acc                        1     48684 48772
## - emp_length                       1     48686 48774
## - percent_bc_gt_75                 1     48692 48780
## - dti                              1     48707 48795
## - revol_util                       1     48719 48807
## - bc_open_to_buy                   1     48730 48818
## - home_ownership                   2     48740 48826
## - inq_last_6mths                   1     48738 48826
## - acc_open_past_24mths             1     48797 48885
## - annual_inc                       1     48808 48896
## - purpose                         12     48834 48900
##
## Step:  AIC=48736.32
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     revol_util + total_acc + collections_12_mths_ex_med + acc_now_delinq +
##     acc_open_past_24mths + bc_open_to_buy + bc_util + mo_sin_old_rev_tl_op +
##     mths_since_recent_bc + mths_since_recent_inq + num_actv_bc_tl +
##     num_bc_sats + num_il_tl + num_rev_tl_bal_gt_0 + num_sats +
##     num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies
##
##                              Df Deviance    AIC
## - collections_12_mths_ex_med  1     48649 48735
## - num_tl_120dpd_2m            1     48649 48735
## - acc_now_delinq              1     48650 48736
## <none>                              48648 48736
## - num_tl_90g_dpd_24m          1     48651 48737
## - pub_rec_bankruptcies        1     48653 48739
## - num_il_tl                   1     48653 48739
## - mo_sin_old_rev_tl_op        1     48655 48741
## - open_acc                    1     48656 48742
## - num_tl_30dpd                1     48656 48742
## - num_tl_op_past_12m          1     48658 48744
## - mths_since_recent_inq       1     48660 48746
## - verification_status         2     48666 48750
## - num_actv_bc_tl              1     48667 48753
## - num_sats                    1     48668 48754
## - mths_since_recent_bc        1     48669 48755
## - bc_util                     1     48672 48758
## - delinq_2yrs                 1     48674 48760
## - num_rev_tl_bal_gt_0         1     48680 48766
## - num_bc_sats                 1     48680 48766
```

28

```
## - total_acc                    1      48685 48771
## - emp_length                   1      48687 48773
## - percent_bc_gt_75             1      48693 48779
## - dti                          1      48707 48793
## - revol_util                   1      48719 48805
## - bc_open_to_buy               1      48730 48816
## - home_ownership               2      48740 48824
## - inq_last_6mths               1      48739 48825
## - acc_open_past_24mths         1      48798 48884
## - annual_inc                   1      48808 48894
## - purpose                     12      48835 48899
##
## Step:  AIC=48735.15
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##       purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##       revol_util + total_acc + acc_now_delinq + acc_open_past_24mths +
##       bc_open_to_buy + bc_util + mo_sin_old_rev_tl_op + mths_since_recent_bc +
##       mths_since_recent_inq + num_actv_bc_tl + num_bc_sats + num_il_tl +
##       num_rev_tl_bal_gt_0 + num_sats + num_tl_120dpd_2m + num_tl_30dpd +
##       num_tl_90g_dpd_24m + num_tl_op_past_12m + percent_bc_gt_75 +
##       pub_rec_bankruptcies
##
##                              Df Deviance    AIC
## - num_tl_120dpd_2m            1      48650 48734
## - acc_now_delinq              1      48650 48734
## <none>                               48649 48735
## - num_tl_90g_dpd_24m          1      48652 48736
## - pub_rec_bankruptcies        1      48654 48738
## - num_il_tl                   1      48654 48738
## - mo_sin_old_rev_tl_op        1      48656 48740
## - num_tl_30dpd                1      48657 48741
## - open_acc                    1      48657 48741
## - num_tl_op_past_12m          1      48659 48743
## - mths_since_recent_inq       1      48661 48745
## - verification_status         2      48667 48749
## - num_actv_bc_tl              1      48668 48752
## - num_sats                    1      48669 48753
## - mths_since_recent_bc        1      48670 48754
## - bc_util                     1      48673 48757
## - delinq_2yrs                 1      48675 48759
## - num_rev_tl_bal_gt_0         1      48680 48764
## - num_bc_sats                 1      48681 48765
## - total_acc                   1      48686 48770
## - emp_length                  1      48688 48772
## - percent_bc_gt_75            1      48694 48778
## - dti                         1      48708 48792
## - revol_util                  1      48720 48804
## - bc_open_to_buy              1      48731 48815
## - home_ownership              2      48741 48823
## - inq_last_6mths              1      48740 48824
## - acc_open_past_24mths        1      48799 48883
## - annual_inc                  1      48809 48893
## - purpose                    12      48836 48898
##
```

```
## Step:  AIC=48734.16
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     revol_util + total_acc + acc_now_delinq + acc_open_past_24mths +
##     bc_open_to_buy + bc_util + mo_sin_old_rev_tl_op + mths_since_recent_bc +
##     mths_since_recent_inq + num_actv_bc_tl + num_bc_sats + num_il_tl +
##     num_rev_tl_bal_gt_0 + num_sats + num_tl_30dpd + num_tl_90g_dpd_24m +
##     num_tl_op_past_12m + percent_bc_gt_75 + pub_rec_bankruptcies
##
##                             Df Deviance   AIC
## - acc_now_delinq            1    48651 48733
## <none>                           48650 48734
## - num_tl_90g_dpd_24m        1    48653 48735
## - pub_rec_bankruptcies      1    48655 48737
## - num_il_tl                 1    48655 48737
## - mo_sin_old_rev_tl_op      1    48657 48739
## - num_tl_30dpd              1    48657 48739
## - open_acc                  1    48658 48740
## - num_tl_op_past_12m        1    48660 48742
## - mths_since_recent_inq     1    48662 48744
## - verification_status       2    48668 48748
## - num_actv_bc_tl            1    48669 48751
## - num_sats                  1    48670 48752
## - mths_since_recent_bc      1    48671 48753
## - bc_util                   1    48674 48756
## - delinq_2yrs               1    48675 48757
## - num_rev_tl_bal_gt_0       1    48682 48764
## - num_bc_sats               1    48682 48764
## - total_acc                 1    48687 48769
## - emp_length                1    48689 48771
## - percent_bc_gt_75          1    48695 48777
## - dti                       1    48709 48791
## - revol_util                1    48721 48803
## - bc_open_to_buy            1    48732 48814
## - home_ownership            2    48742 48822
## - inq_last_6mths            1    48741 48823
## - acc_open_past_24mths      1    48800 48882
## - annual_inc                1    48810 48892
## - purpose                  12    48837 48897
##
## Step:  AIC=48732.6
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     revol_util + total_acc + acc_open_past_24mths + bc_open_to_buy +
##     bc_util + mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_actv_bc_tl + num_bc_sats + num_il_tl + num_rev_tl_bal_gt_0 +
##     num_sats + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies
##
##                             Df Deviance   AIC
## <none>                           48651 48733
## - num_tl_90g_dpd_24m        1    48653 48733
## - pub_rec_bankruptcies      1    48655 48735
## - num_il_tl                 1    48655 48735
```

```
## - mo_sin_old_rev_tl_op    1    48657 48737
## - open_acc                1    48659 48739
## - num_tl_op_past_12m       1    48660 48740
## - num_tl_30dpd             1    48662 48742
## - mths_since_recent_inq    1    48663 48743
## - verification_status      2    48668 48746
## - num_actv_bc_tl           1    48669 48749
## - num_sats                 1    48670 48750
## - mths_since_recent_bc     1    48671 48751
## - bc_util                  1    48675 48755
## - delinq_2yrs              1    48676 48756
## - num_rev_tl_bal_gt_0      1    48682 48762
## - num_bc_sats              1    48682 48762
## - total_acc                1    48687 48767
## - emp_length               1    48689 48769
## - percent_bc_gt_75         1    48695 48775
## - dti                      1    48709 48789
## - revol_util               1    48722 48802
## - bc_open_to_buy           1    48733 48813
## - home_ownership           2    48743 48821
## - inq_last_6mths           1    48741 48821
## - acc_open_past_24mths     1    48800 48880
## - annual_inc               1    48810 48890
## - purpose                 12    48837 48895
```

`stepwise$formula`

```
## defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     revol_util + total_acc + acc_open_past_24mths + bc_open_to_buy +
##     bc_util + mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_actv_bc_tl + num_bc_sats + num_il_tl + num_rev_tl_bal_gt_0 +
##     num_sats + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies
```

`anova(stepwise,log.model.full,test="Chisq")`

```
## Analysis of Deviance Table
##
## Model 1: defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     revol_util + total_acc + acc_open_past_24mths + bc_open_to_buy +
##     bc_util + mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_actv_bc_tl + num_bc_sats + num_il_tl + num_rev_tl_bal_gt_0 +
##     num_sats + num_tl_30dpd + num_tl_90g_dpd_24m + num_tl_op_past_12m +
##     percent_bc_gt_75 + pub_rec_bankruptcies
## Model 2: defaulted ~ emp_length + home_ownership + annual_inc + verification_status +
##     purpose + dti + delinq_2yrs + inq_last_6mths + open_acc +
##     pub_rec + revol_util + total_acc + initial_list_status +
##     collections_12_mths_ex_med + acc_now_delinq + total_rev_hi_lim +
##     acc_open_past_24mths + bc_open_to_buy + bc_util + chargeoff_within_12_mths +
##     mo_sin_old_rev_tl_op + mths_since_recent_bc + mths_since_recent_inq +
##     num_accts_ever_120_pd + num_actv_bc_tl + num_actv_rev_tl +
##     num_bc_sats + num_bc_tl + num_il_tl + num_op_rev_tl + num_rev_tl_bal_gt_0 +
##     num_sats + num_tl_120dpd_2m + num_tl_30dpd + num_tl_90g_dpd_24m +
```

```
##     num_tl_op_past_12m + percent_bc_gt_75 + pub_rec_bankruptcies +
##     tax_liens
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     66117       48651
## 2     66105       48647 12   4.0865   0.9818
```

```
log.model<-stepwise
```

VIF:

- Variable Group 1: Number of Accounts, open_acc, total_acc, mort_acc, num_actv_bc_tl, num_actv_rev_tl, num_bc_sats, num_bc_tl, num_il_tl, num_op_rev_tl, num_rev_accts,num_rev_tl_bal_gt_0, num_sats.

- Variable Group 2: Age (= length of credit history), earliest_cr_line, mo_sin_old_il_acct, mo_sin_old_rev_tl_op,

- Variable Group 3: Balance, revol_bal, tot_cur_bal, avg_cur_bal, total_bal_ex_mort,

- Variable Group 4: Derogatory (~ bad) records; delinq_2yrs, mths_since_last_delinq, mths_since_last_record, pub_rec, collections_12_mths_ex_med, mths_since_last_major_derog, acc_now_delinq, tot_coll_amt, chargeoff_within_12_mths, delinq_amnt, mths_since_recent_bc_dlq, mths_since_recent_revol_delinq num_accts_ever_120_pd, num_tl_120dpd_2m, num_tl_30dpd, num_tl_90g_dpd_24m, num_tl_op_past_12m, pct_tl_nvr_dlq, pub_rec_bankruptcies, tax_liens.

- Variable Group 5: Inquiries; inq_last_6mths,mths_since_recent_inq.

- Variable Group 6: Number of Credit Lines/Limits; total_rev_hi_lim, tot_hi_cred_lim, total_bc_limit, total_il_high_credit_limit.

- Variable Group 7: New (=Recently Opened) Accounts; acc_open_past_24mths, mo_sin_rcnt_rev_tl_op, mo_sin_rcnt_tl,mths_since_recent_bc.

- Variable Group 8: Utilization; revol_util, bc_util, bc_open_to_buy,percent_bc_gt_75.

- Variable Group 9: Internal Variables;

- They appear to be internal variables not available at application but created later to include future information relative to application.; out_prncp, out_prncp_inv, policy_code, recoveries, collection_recovery_fee.

```
library("car")
vif(log.model)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## emp_length           1.111772  1        1.054406
## home_ownership       1.201194  2        1.046895
## annual_inc           1.451257  1        1.204681
## verification_status  1.204474  2        1.047609
## purpose              1.164395 12        1.006362
## dti                  1.432687  1        1.196949
## delinq_2yrs          1.523308  1        1.234224
## inq_last_6mths       1.650985  1        1.284907
## open_acc             5.693245  1        2.386052
## revol_util           2.595731  1        1.611127
## total_acc            2.763890  1        1.662495
## acc_open_past_24mths 2.013864  1        1.419107
## bc_open_to_buy       1.954873  1        1.398168
## bc_util              4.973452  1        2.230124
```

```
## mo_sin_old_rev_tl_op  1.190200  1      1.090963
## mths_since_recent_bc  1.268852  1      1.126433
## mths_since_recent_inq 1.680028  1      1.296159
## num_actv_bc_tl         5.431688  1      2.330598
## num_bc_sats            4.567212  1      2.137104
## num_il_tl              1.801095  1      1.342049
## num_rev_tl_bal_gt_0    3.824969  1      1.955753
## num_sats               6.409657  1      2.531730
## num_tl_30dpd           1.020416  1      1.010157
## num_tl_90g_dpd_24m     1.439781  1      1.199909
## num_tl_op_past_12m     1.642053  1      1.281426
## percent_bc_gt_75       3.322387  1      1.822742
## pub_rec_bankruptcies   1.041556  1      1.020566
```

```r
# Drop num_sats in Group 1:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + de

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

```r
vif(log.model.vif)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## emp_length           1.110902  1      1.053993
## home_ownership       1.200134  2      1.046664
## annual_inc           1.449314  1      1.203875
## verification_status  1.203761  2      1.047454
## purpose              1.163171 12      1.006318
## dti                  1.431754  1      1.196559
## delinq_2yrs          1.516485  1      1.231456
## inq_last_6mths       1.647416  1      1.283517
## open_acc             3.010169  1      1.734984
## revol_util           2.576956  1      1.605290
## total_acc            2.692395  1      1.640852
## acc_open_past_24mths 1.996728  1      1.413056
## bc_open_to_buy       1.952781  1      1.397419
## bc_util              4.963292  1      2.227845
## mo_sin_old_rev_tl_op 1.186187  1      1.089122
## mths_since_recent_bc 1.263015  1      1.123839
## mths_since_recent_inq 1.676280  1      1.294712
## num_actv_bc_tl       5.076814  1      2.253179
## num_bc_sats          3.916812  1      1.979094
## num_il_tl            1.552356  1      1.245936
## num_rev_tl_bal_gt_0  3.310269  1      1.819415
## num_tl_30dpd         1.019926  1      1.009914
## num_tl_90g_dpd_24m   1.434735  1      1.197804
## num_tl_op_past_12m   1.636930  1      1.279426
## percent_bc_gt_75     3.323403  1      1.823020
## pub_rec_bankruptcies 1.040486  1      1.020042
```

```r
# Drop num_actv_bc_tl in Group1:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + d

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

```r
vif(log.model.vif)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## emp_length             1.110833  1        1.053960
## home_ownership         1.199709  2        1.046572
## annual_inc             1.445487  1        1.202284
## verification_status    1.203573  2        1.047413
## purpose                1.162356 12        1.006288
## dti                    1.429543  1        1.195635
## delinq_2yrs            1.519416  1        1.232646
## inq_last_6mths         1.646899  1        1.283316
## open_acc               2.882667  1        1.697842
## revol_util             2.582441  1        1.606997
## total_acc              2.686921  1        1.639183
## acc_open_past_24mths   1.994125  1        1.412135
## bc_open_to_buy         1.950024  1        1.396433
## bc_util                4.933429  1        2.221132
## mo_sin_old_rev_tl_op   1.185327  1        1.088727
## mths_since_recent_bc   1.258106  1        1.121653
## mths_since_recent_inq  1.675579  1        1.294441
## num_bc_sats            2.289327  1        1.513052
## num_il_tl              1.544361  1        1.242723
## num_rev_tl_bal_gt_0    1.920154  1        1.385696
## num_tl_30dpd           1.019976  1        1.009939
## num_tl_90g_dpd_24m     1.437513  1        1.198963
## num_tl_op_past_12m     1.630305  1        1.276834
## percent_bc_gt_75       3.317558  1        1.821417
## pub_rec_bankruptcies   1.040449  1        1.020024
```

```r
# Drop bc_util in Group 8:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + d

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

```r
vif(log.model.vif)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## emp_length             1.110575  1        1.053838
## home_ownership         1.199864  2        1.046605
## annual_inc             1.443994  1        1.201663
## verification_status    1.203440  2        1.047384
## purpose                1.159063 12        1.006169
## dti                    1.428643  1        1.195258
## delinq_2yrs            1.512283  1        1.229749
## inq_last_6mths         1.646269  1        1.283070
## open_acc               2.870340  1        1.694208
## revol_util             2.088391  1        1.445127
## total_acc              2.687913  1        1.639486
## acc_open_past_24mths   1.992730  1        1.411641
## bc_open_to_buy         1.763926  1        1.328129
## mo_sin_old_rev_tl_op   1.186201  1        1.089129
## mths_since_recent_bc   1.250811  1        1.118397
```

```
## mths_since_recent_inq 1.675011  1        1.294222
## num_bc_sats            2.289722  1        1.513183
## num_il_tl              1.541837  1        1.241707
## num_rev_tl_bal_gt_0    1.916005  1        1.384198
## num_tl_30dpd           1.020208  1        1.010054
## num_tl_90g_dpd_24m     1.432430  1        1.196842
## num_tl_op_past_12m     1.629528  1        1.276530
## percent_bc_gt_75       1.849143  1        1.359832
## pub_rec_bankruptcies   1.040323  1        1.019962
```

 # Drop total_acc in Group 1:
```
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + de

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

vif(log.model.vif)

```
##                            GVIF Df GVIF^(1/(2*Df))
## emp_length            1.109981  1        1.053557
## home_ownership        1.177384  2        1.041668
## annual_inc            1.406179  1        1.185824
## verification_status   1.200819  2        1.046814
## purpose               1.155271 12        1.006032
## dti                   1.423235  1        1.192994
## delinq_2yrs           1.502093  1        1.225599
## inq_last_6mths        1.643772  1        1.282097
## open_acc              2.307475  1        1.519038
## revol_util            2.091230  1        1.446109
## acc_open_past_24mths  1.972590  1        1.404489
## bc_open_to_buy        1.763267  1        1.327881
## mo_sin_old_rev_tl_op  1.131644  1        1.063788
## mths_since_recent_bc  1.250803  1        1.118393
## mths_since_recent_inq 1.675015  1        1.294224
## num_bc_sats           2.292700  1        1.514167
## num_il_tl             1.269462  1        1.126704
## num_rev_tl_bal_gt_0   1.916315  1        1.384310
## num_tl_30dpd          1.020180  1        1.010040
## num_tl_90g_dpd_24m    1.439637  1        1.199849
## num_tl_op_past_12m    1.628972  1        1.276312
## percent_bc_gt_75      1.849929  1        1.360121
## pub_rec_bankruptcies  1.039942  1        1.019775
```

 # Drop open_acc in Group 1:
```
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + d

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

vif(log.model.vif)

```
##                           GVIF Df GVIF^(1/(2*Df))
## emp_length           1.108200  1        1.052711
## home_ownership       1.170812  2        1.040212
## annual_inc           1.356839  1        1.164834
```

```
## verification_status    1.200365  2       1.046715
## purpose                1.149534 12       1.005823
## dti                    1.322867  1       1.150160
## delinq_2yrs            1.494761  1       1.222604
## inq_last_6mths         1.644117  1       1.282231
## revol_util             2.016246  1       1.419946
## acc_open_past_24mths   1.834462  1       1.354423
## bc_open_to_buy         1.765784  1       1.328828
## mo_sin_old_rev_tl_op   1.132016  1       1.063963
## mths_since_recent_bc   1.249750  1       1.117922
## mths_since_recent_inq  1.675687  1       1.294483
## num_bc_sats            2.062532  1       1.436152
## num_il_tl              1.189518  1       1.090650
## num_rev_tl_bal_gt_0    1.776301  1       1.332779
## num_tl_30dpd           1.020153  1       1.010026
## num_tl_90g_dpd_24m     1.442266  1       1.200944
## num_tl_op_past_12m     1.624035  1       1.274376
## percent_bc_gt_75       1.844789  1       1.358230
## pub_rec_bankruptcies   1.040076  1       1.019841
# Drop revol_util in Group 8:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + d

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
vif(log.model.vif)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## emp_length             1.108116  1       1.052671
## home_ownership         1.169223  2       1.039859
## annual_inc             1.317362  1       1.147764
## verification_status    1.199419  2       1.046508
## purpose                1.140726 12       1.005501
## dti                    1.302323  1       1.141194
## delinq_2yrs            1.493304  1       1.222008
## inq_last_6mths         1.638701  1       1.280117
## acc_open_past_24mths   1.792914  1       1.338998
## bc_open_to_buy         1.597359  1       1.263867
## mo_sin_old_rev_tl_op   1.131335  1       1.063643
## mths_since_recent_bc   1.249941  1       1.118008
## mths_since_recent_inq  1.673682  1       1.293709
## num_bc_sats            2.032190  1       1.425549
## num_il_tl              1.187612  1       1.089776
## num_rev_tl_bal_gt_0    1.776391  1       1.332813
## num_tl_30dpd           1.020274  1       1.010086
## num_tl_90g_dpd_24m     1.444371  1       1.201820
## num_tl_op_past_12m     1.620116  1       1.272838
## percent_bc_gt_75       1.433249  1       1.197184
## pub_rec_bankruptcies   1.032913  1       1.016323
# Drop num_bc_sats in Group 1:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + d

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

```
vif(log.model.vif)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## emp_length            1.108154  1        1.052689
## home_ownership        1.167677  2        1.039515
## annual_inc            1.311930  1        1.145395
## verification_status   1.198619  2        1.046334
## purpose               1.135142 12        1.005296
## dti                   1.301508  1        1.140836
## delinq_2yrs           1.492164  1        1.221542
## inq_last_6mths        1.638616  1        1.280084
## acc_open_past_24mths  1.790618  1        1.338140
## bc_open_to_buy        1.410508  1        1.187648
## mo_sin_old_rev_tl_op  1.131199  1        1.063579
## mths_since_recent_bc  1.194145  1        1.092769
## mths_since_recent_inq 1.673397  1        1.293599
## num_il_tl             1.186043  1        1.089056
## num_rev_tl_bal_gt_0   1.173025  1        1.083063
## num_tl_30dpd          1.020058  1        1.009979
## num_tl_90g_dpd_24m    1.443877  1        1.201614
## num_tl_op_past_12m    1.616157  1        1.271282
## percent_bc_gt_75      1.425736  1        1.194042
## pub_rec_bankruptcies  1.032829  1        1.016282
```

```
# Drop acc_open_past_24mths in Group 7:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + de

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

```
vif(log.model.vif)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## emp_length            1.108145  1        1.052685
## home_ownership        1.165421  2        1.039012
## annual_inc            1.308675  1        1.143973
## verification_status   1.198477  2        1.046303
## purpose               1.133664 12        1.005241
## dti                   1.288657  1        1.135190
## delinq_2yrs           1.488581  1        1.220074
## inq_last_6mths        1.631258  1        1.277207
## bc_open_to_buy        1.410577  1        1.187677
## mo_sin_old_rev_tl_op  1.125557  1        1.060923
## mths_since_recent_bc  1.137379  1        1.066480
## mths_since_recent_inq 1.662891  1        1.289531
## num_il_tl             1.153614  1        1.074064
## num_rev_tl_bal_gt_0   1.159154  1        1.076640
## num_tl_30dpd          1.019991  1        1.009946
## num_tl_90g_dpd_24m    1.443869  1        1.201611
## num_tl_op_past_12m    1.249558  1        1.117836
## percent_bc_gt_75      1.411768  1        1.188178
## pub_rec_bankruptcies  1.032028  1        1.015888
```

```r
# Drop mths_since_recent_inq in Group 5:
formula<-defaulted ~ emp_length + home_ownership + annual_inc + verification_status + purpose + dti + d

log.model.vif<-glm(formula,data=train,family=binomial(link=logit))
```

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

```r
vif(log.model.vif)
```

```
##                          GVIF Df GVIF^(1/(2*Df))
## emp_length           1.108238  1        1.052729
## home_ownership       1.164452  2        1.038796
## annual_inc           1.309625  1        1.144389
## verification_status  1.197087  2        1.045999
## purpose              1.133136 12        1.005221
## dti                  1.288821  1        1.135263
## delinq_2yrs          1.487648  1        1.219692
## inq_last_6mths       1.086618  1        1.042410
## bc_open_to_buy       1.410190  1        1.187514
## mo_sin_old_rev_tl_op 1.125666  1        1.060974
## mths_since_recent_bc 1.133845  1        1.064822
## num_il_tl            1.154753  1        1.074594
## num_rev_tl_bal_gt_0  1.159669  1        1.076879
## num_tl_30dpd         1.019967  1        1.009934
## num_tl_90g_dpd_24m   1.443168  1        1.201319
## num_tl_op_past_12m   1.227440  1        1.107899
## percent_bc_gt_75     1.412033  1        1.188290
## pub_rec_bankruptcies 1.031933  1        1.015841
```

```r
final.model<-log.model.vif
```

Boostrapping:

Using boostrapping for the coefficients.

```r
bootstrapping <-data.frame(matrix(rep(0,22000),nrow=1000,ncol=32))
```

```
## Warning in matrix(rep(0, 22000), nrow = 1000, ncol = 32): data length
## [22000] is not a sub-multiple or multiple of the number of columns [32]
```

```r
colnames(bootstrapping)<-rownames(summary(final.model)$coefficients)

suppressWarnings(
for (i in 1:1000){
  set.seed(12345*i)
  rnumber<-sample(seq_len(nrow(model.data)),size=n*0.632)
  boots.train<-model.data[rnumber,]
  log.model<-glm(formula,data=boots.train,family=binomial(link=logit))
  bootstrapping [i,]<-summary(log.model)$coefficients[,1]
}
)

boots.coeff<-apply(bootstrapping,2,mean)
summary(final.model)
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = logit), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4503  -0.5708  -0.4783  -0.3741   3.6954
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -2.333e+00  1.283e-01 -18.183  < 2e-16
## emp_lengthn/a                     3.278e-01  5.157e-02   6.356 2.07e-10
## home_ownershipOWN                 1.269e-01  4.465e-02   2.842 0.004486
## home_ownershipRENT                2.729e-01  2.647e-02  10.309  < 2e-16
## annual_inc                       -5.020e-06  4.099e-07 -12.246  < 2e-16
## verification_statusSource Verified 1.304e-01  3.259e-02   4.003 6.26e-05
## verification_statusVerified       8.127e-02  2.856e-02   2.846 0.004433
## purposecredit_card               -1.504e-01  1.198e-01  -1.255 0.209331
## purposedebt_consolidation         5.479e-02  1.180e-01   0.464 0.642374
## purposehome_improvement           1.119e-01  1.284e-01   0.871 0.383487
## purposehouse                      2.006e-01  1.864e-01   1.076 0.281849
## purposemajor_purchase            -1.149e-02  1.447e-01  -0.079 0.936734
## purposemedical                    2.441e-01  1.641e-01   1.487 0.136983
## purposemoving                     3.779e-02  1.786e-01   0.212 0.832433
## purposeother                      3.172e-01  1.259e-01   2.519 0.011755
## purposerenewable_energy           8.723e-01  3.915e-01   2.228 0.025871
## purposesmall_business             8.171e-01  1.403e-01   5.823 5.79e-09
## purposevacation                   1.225e-01  1.793e-01   0.683 0.494305
## purposewedding                   -4.779e-03  1.684e-01  -0.028 0.977365
## dti                               1.645e-02  1.752e-03   9.386  < 2e-16
## delinq_2yrs                       6.879e-02  1.952e-02   3.524 0.000425
## inq_last_6mths                    1.685e-01  1.124e-02  14.985  < 2e-16
## bc_open_to_buy                   -1.596e-05  1.532e-06 -10.415  < 2e-16
## mo_sin_old_rev_tl_op             -6.849e-04  1.453e-04  -4.714 2.43e-06
## mths_since_recent_bc             -3.685e-03  4.764e-04  -7.736 1.03e-14
## num_il_tl                        -1.330e-02  2.006e-03  -6.630 3.35e-11
## num_rev_tl_bal_gt_0               2.479e-02  4.229e-03   5.861 4.59e-09
## num_tl_30dpd                      6.857e-01  1.950e-01   3.516 0.000438
## num_tl_90g_dpd_24m                5.639e-02  3.401e-02   1.658 0.097328
## num_tl_op_past_12m                8.407e-02  8.204e-03  10.248  < 2e-16
## percent_bc_gt_75                  3.169e-03  4.141e-04   7.653 1.96e-14
## pub_rec_bankruptcies             -1.053e-01  4.091e-02  -2.573 0.010089
##
## (Intercept)                      ***
## emp_lengthn/a                    ***
## home_ownershipOWN                **
## home_ownershipRENT               ***
## annual_inc                       ***
## verification_statusSource Verified ***
## verification_statusVerified      **
## purposecredit_card
## purposedebt_consolidation
## purposehome_improvement
## purposehouse
```

```
## purposemajor_purchase
## purposemedical
## purposemoving
## purposeother                      *
## purposerenewable_energy           *
## purposesmall_business             ***
## purposevacation
## purposewedding
## dti                               ***
## delinq_2yrs                       ***
## inq_last_6mths                    ***
## bc_open_to_buy                    ***
## mo_sin_old_rev_tl_op              ***
## mths_since_recent_bc              ***
## num_il_tl                         ***
## num_rev_tl_bal_gt_0               ***
## num_tl_30dpd                      ***
## num_tl_90g_dpd_24m                .
## num_tl_op_past_12m                ***
## percent_bc_gt_75                  ***
## pub_rec_bankruptcies              *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 50963  on 66157  degrees of freedom
## Residual deviance: 48941  on 66126  degrees of freedom
## AIC: 49005
##
## Number of Fisher Scoring iterations: 5
final.model$coefficients <- boots.coeff
```

Test

Obviously, the model performs poorly. However, the objective of this project is not to build a predictive model to predict for default. Rather this work aims to show how to decide a cutoff point given a set of predictions of credit scores for borrowers. Based on the two senarios shown on the top, we will show two approaches in deciding this cutoff point.

```
pred.test<-predict.glm(final.model,test,type="response")
prediction.test<-ifelse(pred.test>0.5,1,0)

confusionMatrix(prediction.test,test$defaulted)

## Confusion Matrix and Statistics
##
##          Reference
## Prediction     0     1
##          0 33592  4929
##          1     2     0
##
##              Accuracy : 0.872
```

```
##                   95% CI : (0.8686, 0.8753)
##     No Information Rate : 0.8721
##     P-Value [Acc > NIR] : 0.516
##
##                   Kappa : -1e-04
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.9999
##             Specificity : 0.0000
##          Pos Pred Value : 0.8720
##          Neg Pred Value : 0.0000
##              Prevalence : 0.8721
##          Detection Rate : 0.8720
##    Detection Prevalence : 0.9999
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

Assign scores:

```r
final.pred<-predict(final.model,model.data,type="response")
scores<-round(850-(850-300)*final.pred)
```

Decide a cutoff score based on percentage of number of people:

```r
deciles<-quantile(scores, probs = seq(0,1, 0.10))
deciles
```

```
##    0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##   495  736  755  766  776  784  791  799  807  817  850
```

```r
Decile.analysis<-data.frame(Percentile=c(1:10),
                        N=rep(round(nrow(model.data)/10),10),
                        Probability=round(rev(tapply(final.pred,findInterval(final.pred,quantile(fir
                        Min_Score=tapply(scores,findInterval(scores,deciles,all.inside = TRUE),min)
                        Median_Score=tapply(scores,findInterval(scores,deciles,all.inside = TRUE),me
                        Max_Score=tapply(scores,findInterval(scores,deciles,all.inside = TRUE),max)
Decile.analysis$"E(Num_Default)"<-round(Decile.analysis$N*Decile.analysis$Probability)
for (i in 1:10){
  Decile.analysis$"Cum_E(Num_Default)"[i]<-sum(Decile.analysis$"E(Num_Default)"[i:10])
}

Decile.analysis$Percentage_Default<-round(Decile.analysis$`Cum_E(Num_Default)`/nrow(model.data),2)
Decile.analysis
```

```
##    Percentile     N Probability Min_Score Median_Score Max_Score
## 10          1 10468        0.25       495          718       735
## 9           2 10468        0.19       736          746       754
## 8           3 10468        0.16       755          760       765
## 7           4 10468        0.14       766          771       775
## 6           5 10468        0.13       776          780       783
```

```
## 5        6 10468        0.11        784        787        790
## 4        7 10468        0.10        791        794        798
## 3        8 10468        0.09        799        802        806
## 2        9 10468        0.07        807        811        816
## 1       10 10468        0.04        817        824        850
##    E(Num_Default) Cum_E(Num_Default) Percentage_Default
## 10          2617              13400               0.13
## 9           1989              10783               0.10
## 8           1675               8794               0.08
## 7           1466               7119               0.07
## 6           1361               5653               0.05
## 5           1151               4292               0.04
## 4           1047               3141               0.03
## 3            942               2094               0.02
## 2            733               1152               0.01
## 1            419                419               0.00
```

According to this approach, we decide the cutoff credit score to be 755, and then we anticipate 8% of the
borrowers will go to default.

Decide a cutoff score based on a fixed amount of loss:

```
ordered.loan.amnt<-original$loan_amnt[order(final.pred,decreasing = TRUE)]
k=round(length(ordered.loan.amnt)/10)

Decile.analysis$Ave_Amount<-tapply(ordered.loan.amnt,c(rep(1,k),rep(2,k),rep(3,k),rep(4,k),rep(5,k),rep
Decile.analysis$"E(default_amnt)"<-Decile.analysis$`E(Num_Default)`*Decile.analysis$Ave_Amount
for (i in 1:10){
  Decile.analysis$"Cum_E(default_amnt)"[i]<-sum(Decile.analysis$`E(default_amnt)`[i:10])
}

Decile.analysis
```

```
##    Percentile     N Probability Min_Score Median_Score Max_Score
## 10          1 10468        0.25        495          718       735
## 9           2 10468        0.19        736          746       754
## 8           3 10468        0.16        755          760       765
## 7           4 10468        0.14        766          771       775
## 6           5 10468        0.13        776          780       783
## 5           6 10468        0.11        784          787       790
## 4           7 10468        0.10        791          794       798
## 3           8 10468        0.09        799          802       806
## 2           9 10468        0.07        807          811       816
## 1          10 10468        0.04        817          824       850
##    E(Num_Default) Cum_E(Num_Default) Percentage_Default Ave_Amount
## 10          2617              13400               0.13   9267.195
## 9           1989              10783               0.10  10151.476
## 8           1675               8794               0.08  10654.521
## 7           1466               7119               0.07  11122.578
## 6           1361               5653               0.05  11725.781
## 5           1151               4292               0.04  12258.524
## 4           1047               3141               0.03  12796.850
## 3            942               2094               0.02  13549.670
```

```
## 2               733               1152                    0.01  14570.701
## 1               419               419                     0.00  16716.040
##    E(default_amnt) Cum_E(default_amnt)
## 10    24252250            152510343
## 9     20191286            128258093
## 8     17846323            108066807
## 7     16305700             90220485
## 6     15958788             73914785
## 5     14109561             57955997
## 4     13398302             43846436
## 3     12763790             30448134
## 2     10680324             17684345
## 1      7004021              7004021
```

Here, if we anticipate that we will not get back 90 million, we will set the cut off score to be 766.

KS Stats:

```
# KS statistics:
library("KScorrect")
kstest<-ks.test(scores[scores<755],scores[scores>=755])
```

```
## Warning in ks.test(scores[scores < 755], scores[scores >= 755]): p-value
## will be approximate in the presence of ties
```

```
kstest
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  scores[scores < 755] and scores[scores >= 755]
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```