STAT 652 Sec 1 - Project

Yen Tran

02/08/2019

Abstract

This project is to develop the best classification model to classify loan default for Lending Club. Applying different Machine Learning algorithms, the best model is selected based on the highest AUC value between the predicted probability and the observed label. The winning model is Logistic Regression model.

Introduction

Business Understanding

Lending Club is the world's largest peer-to-peer lending platform. It operates an online lending platforms that enables borrowers to obtain a loan, and investors to purchases notes backed by payments made on loans. Lending Club enables borrowers to create unsecured personal loans between 1,000 and 40,000 dollars.

Problem Statement

How to classify or detect which loan will more likely to default based on the borrower's credit history?

Data

I downloaded the data from Lending Club database for data from 2012-2015. As the goal is to build classification model to classify whether a borrower, who already borrowed the loan, will repay a loan or not based on borrower's credit history.

The target variable in the dataset is '**loan_status**', which has 3 different categorical values ('Charged Off', 'Fully Paid', and 'Default').
- Fully Paid: Loan has been fully repaid.
- Default: Loan has not been current for 121 days or more.
- Charged Off: Loan for which the is no longer a reasonable expectation of further payments.

The objective is to predict whether a borrower will default the loan, loan_status will be either Fully Paid or Charged Off, with binary of 0 and 1 respectively.

Model applied

1. kNN
2. C5.0
3. Naive Bayes
4. Logistic Regression with LASSO
5. Random Forest
6. Bagged Tree
7. GBM - Boosted Tree

Evaluation Approach

The are under the ROC curve (AUC) is used to compare model performances, where the ROC plots True Positive Rate versus False Positive Rate. Below is the summary table of evaluation methods:

Model	AUC	Accuracy	Kappa
kNN	0.213	0.9111	0.6253
C5.0 Tree	0.998	0.9972	0.9905
Naive Bayes	0.903	0.7378	0.4081
Logistic Regression	1.000	0.9997	0.9991
Random Forest	0.999	0.9919	0.9721
Bagged Trees	0.997	0.9964	0.9877
GBM	0.999	0.9876	0.9566

Conclusions

The logistic regression model performs better than other algorithms with highest area under the ROC curve between the predicted probability and the observed label.

Code

Step 1a: Download data

```
#Unzip data files
unzip("./data/2012-2013.zip",exdir="./data")
```

```
## Warning in unzip("./data/2012-2013.zip", exdir = "./data"): error 1 in
## extracting from zip file
```

```
unzip("./data/2014.zip",exdir="./data")
```

```
## Warning in unzip("./data/2014.zip", exdir = "./data"): error 1 in
## extracting from zip file
```

```
#Clean up data directory
fn <- "./data/2012-2013.zip"
if (file.exists(fn)) file.remove(fn)
fn <- "./data/2014.zip"
if (file.exists(fn)) file.remove(fn)
```

```
#Read the.csv files
Loan_1213 <- read.csv(file="./data/LoanStats3b.csv", skip = 1, na.strings=c(""))
Loan_14 <- read.csv(file="./data/LoanStats3c.csv", skip = 1, na.strings=c(""))
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
#Removing the last two rows, empty rows due to importing error
Loan_1213 <- Loan_1213[1:(nrow(Loan_1213)-2),]
Loan_14 <- Loan_14[1:(nrow(Loan_1213)-2),]

Loan_1213 %>% group_by(issue_d) %>% summarise(count = n())
```

```
## # A tibble: 24 x 2
##    issue_d  count
##    <fct>    <int>
##  1 Apr-2012  3230
##  2 Apr-2013  9419
##  3 Aug-2012  5419
##  4 Aug-2013 12674
##  5 Dec-2012  6066
##  6 Dec-2013 15020
##  7 Feb-2012  2560
##  8 Feb-2013  7561
##  9 Jan-2012  2602
## 10 Jan-2013  6872
## # ... with 14 more rows
```

```r
Loan_14 %>% group_by(issue_d) %>% summarise(count = n())
```

```
## # A tibble: 10 x 2
##    issue_d  count
##    <fct>    <int>
##  1 Apr-2014 18413
##  2 Aug-2014 18814
##  3 Dec-2014 10307
##  4 Jul-2014 29306
##  5 Jun-2014 17179
##  6 Mar-2014   618
##  7 May-2014 19099
##  8 Nov-2014 25054
##  9 Oct-2014 38783
## 10 Sep-2014 10606
```

```r
Loan_14 %>% group_by(member_id) %>% summarise(count = n()) %>% arrange(count)
```

```
## # A tibble: 1 x 2
##   member_id  count
##   <lgl>      <int>
## 1 NA        188179
```

```r
#Merging data from 2012 - 2014, converting same data type for both files.
class(Loan_1213$hardship_dpd)
```

```
## [1] "integer"
```

```r
class(Loan_14$hardship_dpd)
```

```
## [1] "integer"
```

```r
Loan_1213$hardship_dpd <- as.numeric(Loan_1213$hardship_dpd)
class(Loan_1213$hardship_dpd)
```

```
## [1] "numeric"
```

```r
Loan_merged <- bind_rows(Loan_1213, Loan_14)
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector

## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character

## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
## Warning in bind_rows_(x, .id): binding character and factor vector,
## coercing into character vector
```

```
saveRDS(Loan_merged, file = "Loan_merged")
```

Step 1b: Load data

```
library(dplyr)
Loan <- readRDS(file="Loan_merged")
```

Step 2: Explore and prepare data for analysis

```
#Explore data table structure
dim(Loan)
```

```
## [1] 376360    145
```

```
head(Loan)
```

```
##     id member_id loan_amnt funded_amnt funded_amnt_inv       term int_rate
## 1 <NA>        NA      7550        7550            7550  36 months   16.24%
## 2 <NA>        NA      3000        3000            3000  36 months   12.85%
## 3 <NA>        NA     20800       20800           20800  36 months   13.53%
## 4 <NA>        NA      4800        4800            4800  36 months   10.99%
## 5 <NA>        NA     14000       14000           14000  36 months   12.85%
## 6 <NA>        NA     15000       15000           15000  36 months   14.47%
##   installment grade sub_grade                            emp_title
## 1      266.34     C        C5      Special Order Fulfillment Clerk
## 2      100.87     B        B4                              Auditor
## 3      706.16     B        B5                   Operations Manager
## 4      157.13     B        B2                  Surgical Technician
## 5      470.71     B        B4 Assistant Director - Human Resources
## 6      516.10     C        C2                 building maint. mgr.
##   emp_length home_ownership annual_inc verification_status  issue_d
## 1    3 years           RENT      28000        Not Verified Dec-2013
## 2  10+ years           RENT      25000            Verified Dec-2013
## 3  10+ years           RENT      81500            Verified Dec-2013
## 4    2 years       MORTGAGE      39600     Source Verified Dec-2013
## 5    4 years           RENT      88000        Not Verified Dec-2013
## 6  10+ years           RENT      98000        Not Verified Dec-2013
##   loan_status pymnt_plan url
## 1  Fully Paid          n  NA
## 2  Fully Paid          n  NA
## 3  Fully Paid          n  NA
## 4  Fully Paid          n  NA
## 5  Fully Paid          n  NA
## 6  Fully Paid          n  NA
##
## 1
## 2
## 3                     Borrower added on 12/31/13 > My goal is to purchase a home. I am consolidating
## 4   Borrower added on 12/31/13 > Just bought a house, and would like a little extra funds to improve
## 5
```

```
## 6
##             purpose                         title zip_code addr_state
## 1 debt_consolidation            Debt consolidation    951xx         CA
## 2 debt_consolidation                          debt    322xx         FL
## 3 debt_consolidation Reducing Debt to Purchase Home    100xx         NY
## 4   home_improvement                For The House    782xx         TX
## 5 debt_consolidation            Debt consolidation    282xx         NC
## 6 debt_consolidation                       pay off    117xx         NY
##     dti delinq_2yrs earliest_cr_line inq_last_6mths mths_since_last_delinq
## 1  8.40          0         Oct-2010              0                     NA
## 2 24.68          0         May-1991              0                     58
## 3 16.73          0         Jun-1998              2                     64
## 4  2.49          0         Aug-1995              2                     NA
## 5 10.02          1         Jun-1988              0                     16
## 6  6.15          0         Jul-1992              2                     NA
##   mths_since_last_record open_acc pub_rec revol_bal revol_util total_acc
## 1                     NA        4       0      5759        72%         5
## 2                     53        5       2      2875      54.2%        26
## 3                     NA       29       0     23473      54.5%        41
## 4                     NA        3       0      4136      16.1%         8
## 5                    115        6       1      3686      81.9%        14
## 6                     NA       16       0      5749      22.3%        16
##   initial_list_status out_prncp out_prncp_inv total_pymnt total_pymnt_inv
## 1                   w         0             0    9600.455         9600.45
## 2                   f         0             0    3181.549         3181.55
## 3                   f         0             0   23926.640        23926.64
## 4                   w         0             0    5157.519         5157.52
## 5                   f         0             0   16945.319        16945.32
## 6                   f         0             0   15699.052        15699.05
##   total_rec_prncp total_rec_int total_rec_late_fee recoveries
## 1            7550       2050.45                 0          0
## 2            3000        181.55                 0          0
## 3           20800       3126.64                 0          0
## 4            4800        357.52                 0          0
## 5           14000       2945.32                 0          0
## 6           15000        699.05                 0          0
##   collection_recovery_fee last_pymnt_d last_pymnt_amnt next_pymnt_d
## 1                       0     Dec-2016          529.67        <NA>
## 2                       0     Jul-2014         2677.23        <NA>
## 3                       0     May-2015        13334.93        <NA>
## 4                       0     Sep-2014         3900.48        <NA>
## 5                       0     Jan-2017          470.47        <NA>
## 6                       0     May-2014        14150.76        <NA>
##   last_credit_pull_d collections_12_mths_ex_med
## 1          Oct-2018                          0
## 2          Oct-2016                          0
## 3          Jan-2019                          0
## 4          Jan-2017                          0
## 5          Jan-2019                          0
## 6          Oct-2018                          0
##   mths_since_last_major_derog policy_code application_type
## 1                          NA           1       Individual
## 2                          69           1       Individual
## 3                          71           1       Individual
```

```
## 4                          NA           1       Individual
## 5                          NA           1       Individual
## 6                          NA           1       Individual
##   annual_inc_joint dti_joint verification_status_joint acc_now_delinq
## 1               NA        NA                        NA             0
## 2               NA        NA                        NA             0
## 3               NA        NA                        NA             0
## 4               NA        NA                        NA             0
## 5               NA        NA                        NA             0
## 6               NA        NA                        NA             0
##   tot_coll_amt tot_cur_bal open_acc_6m open_act_il open_il_12m open_il_24m
## 1            0        5759          NA          NA          NA          NA
## 2          154       19530          NA          NA          NA          NA
## 3            0       23473          NA          NA          NA          NA
## 4            0        4136          NA          NA          NA          NA
## 5            0       17672          NA          NA          NA          NA
## 6            0       13038          NA          NA          NA          NA
##   mths_since_rcnt_il total_bal_il il_util open_rv_12m open_rv_24m
## 1                 NA           NA      NA          NA          NA
## 2                 NA           NA      NA          NA          NA
## 3                 NA           NA      NA          NA          NA
## 4                 NA           NA      NA          NA          NA
## 5                 NA           NA      NA          NA          NA
## 6                 NA           NA      NA          NA          NA
##   max_bal_bc all_util total_rev_hi_lim inq_fi total_cu_tl inq_last_12m
## 1         NA       NA             8000     NA          NA           NA
## 2         NA       NA             5300     NA          NA           NA
## 3         NA       NA            43100     NA          NA           NA
## 4         NA       NA            25700     NA          NA           NA
## 5         NA       NA             4500     NA          NA           NA
## 6         NA       NA            25800     NA          NA           NA
##   acc_open_past_24mths avg_cur_bal bc_open_to_buy bc_util
## 1                    1        1440            160    96.0
## 2                    3        3906           2050    52.3
## 3                    9         869           6811    54.6
## 4                    0        1379          21564    16.1
## 5                    3        2945            480    87.7
## 6                    6         815          15051    27.6
##   chargeoff_within_12_mths delinq_amnt mo_sin_old_il_acct
## 1                        0          0                 NA
## 2                        0          0                164
## 3                        0          0                115
## 4                        0          0                104
## 5                        0          0                111
## 6                        0          0                  2
##   mo_sin_old_rev_tl_op mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl mort_acc
## 1                   38                    17             17        0
## 2                  271                     7              7        6
## 3                  186                     0              0        0
## 4                  220                    25             25        0
## 5                  103                    24             13        0
## 6                  257                     7              2        0
##   mths_since_recent_bc mths_since_recent_bc_dlq mths_since_recent_inq
## 1                   17                      NA                   17
```

```
## 2                  14                     69                     8
## 3                   0                     70                     0
## 4                  25                     NA                     3
## 5                  38                     16                    NA
## 6                   7                     NA                     2
##   mths_since_recent_revol_delinq num_accts_ever_120_pd num_actv_bc_tl
## 1                             NA                     0              2
## 2                             69                     1              2
## 3                             70                     1              8
## 4                             NA                     0              2
## 5                             16                     0              3
## 6                             NA                     0              8
##   num_actv_rev_tl num_bc_sats num_bc_tl num_il_tl num_op_rev_tl
## 1               4           2         2         0             4
## 2               3           3         6        11             4
## 3              24          11        17         1            29
## 4               2           3         4         1             3
## 5               4           3         9         3             4
## 6               8          13        13         1            15
##   num_rev_accts num_rev_tl_bal_gt_0 num_sats num_tl_120dpd_2m num_tl_30dpd
## 1             5                   4        4                0            0
## 2             9                   3        5                0            0
## 3            40                  24       29                0            0
## 4             7                   2        3                0            0
## 5            10                   4        6                0            0
## 6            15                   8       16                0            0
##   num_tl_90g_dpd_24m num_tl_op_past_12m pct_tl_nvr_dlq percent_bc_gt_75
## 1                  0                  0          100.0            100.0
## 2                  0                  1           91.3             66.7
## 3                  0                  3           90.2             50.0
## 4                  0                  0          100.0              0.0
## 5                  0                  0           78.6            100.0
## 6                  0                  2          100.0              7.7
##   pub_rec_bankruptcies tax_liens tot_hi_cred_lim total_bal_ex_mort
## 1                    0         0            8000              5759
## 2                    2         0           32082             19530
## 3                    0         0           43100             23473
## 4                    0         0           25700              4136
## 5                    1         0           31840             17672
## 6                    0         0           33300             13038
##   total_bc_limit total_il_high_credit_limit revol_bal_joint
## 1           4000                         0             NA
## 2           4300                     26782             NA
## 3          15000                         0             NA
## 4          25700                         0             NA
## 5           3900                     27340             NA
## 6          20800                      7500             NA
##   sec_app_earliest_cr_line sec_app_inq_last_6mths sec_app_mort_acc
## 1                       NA                     NA               NA
## 2                       NA                     NA               NA
## 3                       NA                     NA               NA
## 4                       NA                     NA               NA
## 5                       NA                     NA               NA
## 6                       NA                     NA               NA
```

```
##   sec_app_open_acc sec_app_revol_util sec_app_open_act_il
## 1               NA                 NA                  NA
## 2               NA                 NA                  NA
## 3               NA                 NA                  NA
## 4               NA                 NA                  NA
## 5               NA                 NA                  NA
## 6               NA                 NA                  NA
##   sec_app_num_rev_accts sec_app_chargeoff_within_12_mths
## 1                    NA                              NA
## 2                    NA                              NA
## 3                    NA                              NA
## 4                    NA                              NA
## 5                    NA                              NA
## 6                    NA                              NA
##   sec_app_collections_12_mths_ex_med sec_app_mths_since_last_major_derog
## 1                                 NA                                   NA
## 2                                 NA                                   NA
## 3                                 NA                                   NA
## 4                                 NA                                   NA
## 5                                 NA                                   NA
## 6                                 NA                                   NA
##   hardship_flag hardship_type hardship_reason hardship_status
## 1             N          <NA>            <NA>            <NA>
## 2             N          <NA>            <NA>            <NA>
## 3             N          <NA>            <NA>            <NA>
## 4             N          <NA>            <NA>            <NA>
## 5             N          <NA>            <NA>            <NA>
## 6             N          <NA>            <NA>            <NA>
##   deferral_term hardship_amount hardship_start_date hardship_end_date
## 1            NA              NA                <NA>              <NA>
## 2            NA              NA                <NA>              <NA>
## 3            NA              NA                <NA>              <NA>
## 4            NA              NA                <NA>              <NA>
## 5            NA              NA                <NA>              <NA>
## 6            NA              NA                <NA>              <NA>
##   payment_plan_start_date hardship_length hardship_dpd
## 1                    <NA>              NA           NA
## 2                    <NA>              NA           NA
## 3                    <NA>              NA           NA
## 4                    <NA>              NA           NA
## 5                    <NA>              NA           NA
## 6                    <NA>              NA           NA
##   hardship_loan_status orig_projected_additional_accrued_interest
## 1                <NA>                                        NA
## 2                <NA>                                        NA
## 3                <NA>                                        NA
## 4                <NA>                                        NA
## 5                <NA>                                        NA
## 6                <NA>                                        NA
##   hardship_payoff_balance_amount hardship_last_payment_amount
## 1                             NA                           NA
## 2                             NA                           NA
## 3                             NA                           NA
## 4                             NA                           NA
```

```
## 5                            NA                           NA
## 6                            NA                           NA
##   disbursement_method debt_settlement_flag debt_settlement_flag_date
## 1                Cash                    N                      <NA>
## 2                Cash                    N                      <NA>
## 3                Cash                    N                      <NA>
## 4                Cash                    N                      <NA>
## 5                Cash                    N                      <NA>
## 6                Cash                    N                      <NA>
##   settlement_status settlement_date settlement_amount
## 1              <NA>            <NA>                NA
## 2              <NA>            <NA>                NA
## 3              <NA>            <NA>                NA
## 4              <NA>            <NA>                NA
## 5              <NA>            <NA>                NA
## 6              <NA>            <NA>                NA
##   settlement_percentage settlement_term
## 1                    NA              NA
## 2                    NA              NA
## 3                    NA              NA
## 4                    NA              NA
## 5                    NA              NA
## 6                    NA              NA
```

```
tail(Loan)
```

```
##          id member_id loan_amnt funded_amnt funded_amnt_inv       term
## 376355 <NA>        NA     14000       14000           14000  36 months
## 376356 <NA>        NA     12000       12000           12000  36 months
## 376357 <NA>        NA     10000       10000           10000  36 months
## 376358 <NA>        NA     30000       30000           30000  36 months
## 376359 <NA>        NA     20000       20000           20000  60 months
## 376360 <NA>        NA     13300       13300           13300  36 months
##        int_rate installment grade sub_grade                   emp_title
## 376355   14.64%      482.86     C        C3               General Manager
## 376356    9.67%      385.35     B        B1                       foreman
## 376357   11.99%      332.10     B        B3 Business Development Analyst
## 376358   13.65%     1020.24     C        C1                     Principal
## 376359   14.64%      472.03     C        C3                     Carpenter
## 376360   10.99%      435.37     B        B2       Regional Sales Manager
##        emp_length home_ownership annual_inc verification_status  issue_d
## 376355    6 years           RENT      90000     Source Verified Mar-2014
## 376356    6 years       MORTGAGE      53000        Not Verified Apr-2014
## 376357    2 years           RENT      60000            Verified Mar-2014
## 376358    6 years       MORTGAGE      78000            Verified Mar-2014
## 376359  10+ years       MORTGAGE      70000            Verified Mar-2014
## 376360   < 1 year       MORTGAGE      72000     Source Verified Mar-2014
##        loan_status pymnt_plan url desc            purpose
## 376355 Charged Off          n  NA <NA>        credit_card
## 376356 Charged Off          n  NA <NA>        credit_card
## 376357  Fully Paid          n  NA <NA>        credit_card
## 376358  Fully Paid          n  NA <NA>        credit_card
## 376359  Fully Paid          n  NA <NA> debt_consolidation
## 376360  Fully Paid          n  NA <NA>   home_improvement
##                         title zip_code addr_state   dti delinq_2yrs
```

```
## 376355 Credit card refinancing    891xx         NV  7.64           0
## 376356 Credit card refinancing    338xx         FL 23.80           2
## 376357 Credit card refinancing    606xx         IL  6.13           0
## 376358 Credit card refinancing    310xx         GA 31.94           0
## 376359      Debt consolidation    481xx         MI 24.58           0
## 376360        Home improvement    633xx         MO 11.75           0
##        earliest_cr_line inq_last_6mths mths_since_last_delinq
## 376355         Jul-2003              1                     61
## 376356         Apr-2003              0                      9
## 376357         Oct-2007              0                     NA
## 376358         Aug-1989              0                     46
## 376359         Dec-1994              2                     NA
## 376360         Sep-1993              1                     NA
##        mths_since_last_record open_acc pub_rec revol_bal revol_util
## 376355                     90        5       1      8236        67%
## 376356                     NA       10       0     11549      56.1%
## 376357                     NA        6       0      9424      77.9%
## 376358                     NA       30       0     38228      52.2%
## 376359                     78        9       1      6056      32.9%
## 376360                     NA        9       0      5199      43.7%
##        total_acc initial_list_status out_prncp out_prncp_inv total_pymnt
## 376355        16                   w         0             0     2897.16
## 376356        36                   w         0             0     5109.88
## 376357        11                   w         0             0    10764.65
## 376358        62                   f         0             0    36728.49
## 376359        26                   w         0             0    27470.73
## 376360        17                   w         0             0    13656.78
##        total_pymnt_inv total_rec_prncp total_rec_int total_rec_late_fee
## 376355         2897.16         1930.40        966.76                 0
## 376356         5109.88         2683.18        784.97                 0
## 376357        10764.65        10000.00        764.65                 0
## 376358        36728.49        30000.00       6728.49                 0
## 376359        27470.73        20000.00       7470.73                 0
## 376360        13656.78        13300.00        356.78                 0
##        recoveries collection_recovery_fee last_pymnt_d last_pymnt_amnt
## 376355       0.00                  0.0000     Oct-2014          482.86
## 376356    1641.73                295.5114     Jan-2015          385.35
## 376357       0.00                  0.0000     May-2015         1779.45
## 376358       0.00                  0.0000     Apr-2017         1020.09
## 376359       0.00                  0.0000     Oct-2017         8117.50
## 376360       0.00                  0.0000     Jul-2014        12786.04
##        next_pymnt_d last_credit_pull_d collections_12_mths_ex_med
## 376355         <NA>           Jan-2019                         0
## 376356         <NA>           Oct-2016                         0
## 376357         <NA>           Feb-2017                         0
## 376358         <NA>           Jan-2019                         0
## 376359         <NA>           Jan-2019                         0
## 376360         <NA>           Sep-2018                         0
##        mths_since_last_major_derog policy_code application_type
## 376355                          61           1       Individual
## 376356                           9           1       Individual
## 376357                          NA           1       Individual
## 376358                          NA           1       Individual
## 376359                          NA           1       Individual
```

```
## 376360                          NA           1       Individual
##        annual_inc_joint dti_joint verification_status_joint acc_now_delinq
## 376355               NA        NA                        NA              0
## 376356               NA        NA                        NA              0
## 376357               NA        NA                        NA              0
## 376358               NA        NA                        NA              0
## 376359               NA        NA                        NA              0
## 376360               NA        NA                        NA              0
##        tot_coll_amt tot_cur_bal open_acc_6m open_act_il open_il_12m
## 376355           76       26821          NA          NA          NA
## 376356          266      149948          NA          NA          NA
## 376357            0        9424          NA          NA          NA
## 376358            0      343959          NA          NA          NA
## 376359            0       70240          NA          NA          NA
## 376360            0      166895          NA          NA          NA
##        open_il_24m mths_since_rcnt_il total_bal_il il_util open_rv_12m
## 376355          NA                NA           NA      NA          NA
## 376356          NA                NA           NA      NA          NA
## 376357          NA                NA           NA      NA          NA
## 376358          NA                NA           NA      NA          NA
## 376359          NA                NA           NA      NA          NA
## 376360          NA                NA           NA      NA          NA
##        open_rv_24m max_bal_bc all_util total_rev_hi_lim inq_fi total_cu_tl
## 376355          NA         NA       NA            12300     NA          NA
## 376356          NA         NA       NA            20600     NA          NA
## 376357          NA         NA       NA            12100     NA          NA
## 376358          NA         NA       NA            73300     NA          NA
## 376359          NA         NA       NA            18400     NA          NA
## 376360          NA         NA       NA            11900     NA          NA
##        inq_last_12m acc_open_past_24mths avg_cur_bal bc_open_to_buy
## 376355           NA                    3        5364             64
## 376356           NA                    4       16661           5568
## 376357           NA                    3        1571            320
## 376358           NA                    4       13229          12215
## 376359           NA                    5        7804          10344
## 376360           NA                    6       20862           6701
##        bc_util chargeoff_within_12_mths delinq_amnt mo_sin_old_il_acct
## 376355    99.2                        0           0                128
## 376356    34.5                        0           0                131
## 376357    95.3                        0           0                 NA
## 376358    72.6                        0           0                145
## 376359    36.9                        0           0                160
## 376360    43.7                        0           0                127
##        mo_sin_old_rev_tl_op mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl mort_acc
## 376355                  110                    13              8        3
## 376356                  130                    20              1        1
## 376357                   77                     3              3        0
## 376358                  295                     2              2        2
## 376359                  231                     3              3        0
## 376360                  246                     3              3        1
##        mths_since_recent_bc mths_since_recent_bc_dlq mths_since_recent_inq
## 376355                   18                       63                    2
## 376356                   27                       NA                   21
## 376357                   23                       NA                   18
```

```
## 376358                  16                      NA                     7
## 376359                   3                      NA                     3
## 376360                   3                      NA                     3
##        mths_since_recent_revol_delinq num_accts_ever_120_pd num_actv_bc_tl
## 376355                             61                    1             3
## 376356                             NA                    1             2
## 376357                             NA                    0             3
## 376358                             46                    0             5
## 376359                             NA                    0             3
## 376360                             NA                    0             5
##        num_actv_rev_tl num_bc_sats num_bc_tl num_il_tl num_op_rev_tl
## 376355               3           3         4         4             4
## 376356               4           3        15        11             7
## 376357               5           3         5         0             6
## 376358              15           5        13        28            23
## 376359               3           5        12        11             6
## 376360               5           5         7         6             6
##        num_rev_accts num_rev_tl_bal_gt_0 num_sats num_tl_120dpd_2m
## 376355             9                   3        5                0
## 376356            23                   4       10                0
## 376357            10                   5        6                0
## 376358            32                  15       30                0
## 376359            15                   3        9                0
## 376360            10                   5        9                0
##        num_tl_30dpd num_tl_90g_dpd_24m num_tl_op_past_12m pct_tl_nvr_dlq
## 376355            0                  0                  1           86.7
## 376356            0                  1                  1           94.4
## 376357            0                  0                  1          100.0
## 376358            0                  0                  2           98.3
## 376359            0                  0                  3          100.0
## 376360            0                  0                  3          100.0
##        percent_bc_gt_75 pub_rec_bankruptcies tax_liens tot_hi_cred_lim
## 376355            100.0                    1         0           31908
## 376356             33.3                    0         0          183419
## 376357            100.0                    0         0           12100
## 376358             20.0                    0         0          390184
## 376359             40.0                    0         0          109532
## 376360             20.0                    0         0          206107
##        total_bal_ex_mort total_bc_limit total_il_high_credit_limit
## 376355             26821           8300                      19608
## 376356             47313           8500                      43760
## 376357              9424           6800                          0
## 376358            165718          44500                     127023
## 376359             70240          16400                      91132
## 376360             26322          11900                      25207
##        revol_bal_joint sec_app_earliest_cr_line sec_app_inq_last_6mths
## 376355              NA                      NA                     NA
## 376356              NA                      NA                     NA
## 376357              NA                      NA                     NA
## 376358              NA                      NA                     NA
## 376359              NA                      NA                     NA
## 376360              NA                      NA                     NA
##        sec_app_mort_acc sec_app_open_acc sec_app_revol_util
## 376355               NA               NA                 NA
```

```
## 376356              NA              NA                 NA
## 376357              NA              NA                 NA
## 376358              NA              NA                 NA
## 376359              NA              NA                 NA
## 376360              NA              NA                 NA
##        sec_app_open_act_il sec_app_num_rev_accts
## 376355                  NA                   NA
## 376356                  NA                   NA
## 376357                  NA                   NA
## 376358                  NA                   NA
## 376359                  NA                   NA
## 376360                  NA                   NA
##        sec_app_chargeoff_within_12_mths sec_app_collections_12_mths_ex_med
## 376355                               NA                                 NA
## 376356                               NA                                 NA
## 376357                               NA                                 NA
## 376358                               NA                                 NA
## 376359                               NA                                 NA
## 376360                               NA                                 NA
##        sec_app_mths_since_last_major_derog hardship_flag hardship_type
## 376355                                  NA             N          <NA>
## 376356                                  NA             N          <NA>
## 376357                                  NA             N          <NA>
## 376358                                  NA             N          <NA>
## 376359                                  NA             N          <NA>
## 376360                                  NA             N          <NA>
##        hardship_reason hardship_status deferral_term hardship_amount
## 376355            <NA>            <NA>            NA              NA
## 376356            <NA>            <NA>            NA              NA
## 376357            <NA>            <NA>            NA              NA
## 376358            <NA>            <NA>            NA              NA
## 376359            <NA>            <NA>            NA              NA
## 376360            <NA>            <NA>            NA              NA
##        hardship_start_date hardship_end_date payment_plan_start_date
## 376355                <NA>              <NA>                    <NA>
## 376356                <NA>              <NA>                    <NA>
## 376357                <NA>              <NA>                    <NA>
## 376358                <NA>              <NA>                    <NA>
## 376359                <NA>              <NA>                    <NA>
## 376360                <NA>              <NA>                    <NA>
##        hardship_length hardship_dpd hardship_loan_status
## 376355              NA           NA                 <NA>
## 376356              NA           NA                 <NA>
## 376357              NA           NA                 <NA>
## 376358              NA           NA                 <NA>
## 376359              NA           NA                 <NA>
## 376360              NA           NA                 <NA>
##        orig_projected_additional_accrued_interest
## 376355                                         NA
## 376356                                         NA
## 376357                                         NA
## 376358                                         NA
## 376359                                         NA
## 376360                                         NA
```

```
##        hardship_payoff_balance_amount hardship_last_payment_amount
## 376355                             NA                           NA
## 376356                             NA                           NA
## 376357                             NA                           NA
## 376358                             NA                           NA
## 376359                             NA                           NA
## 376360                             NA                           NA
##        disbursement_method debt_settlement_flag debt_settlement_flag_date
## 376355                Cash                    N                      <NA>
## 376356                Cash                    N                      <NA>
## 376357                Cash                    N                      <NA>
## 376358                Cash                    N                      <NA>
## 376359                Cash                    N                      <NA>
## 376360                Cash                    N                      <NA>
##        settlement_status settlement_date settlement_amount
## 376355              <NA>            <NA>                NA
## 376356              <NA>            <NA>                NA
## 376357              <NA>            <NA>                NA
## 376358              <NA>            <NA>                NA
## 376359              <NA>            <NA>                NA
## 376360              <NA>            <NA>                NA
##        settlement_percentage settlement_term
## 376355                    NA              NA
## 376356                    NA              NA
## 376357                    NA              NA
## 376358                    NA              NA
## 376359                    NA              NA
## 376360                    NA              NA
```

```
str(Loan)
```

```
## 'data.frame':    376360 obs. of  145 variables:
##  $ id                                        : chr  NA NA NA NA ...
##  $ member_id                                 : logi  NA NA NA NA NA NA ...
##  $ loan_amnt                                 : int  7550 3000 20800 4800 14000 15000 11100 12000 975
##  $ funded_amnt                               : int  7550 3000 20800 4800 14000 15000 11100 12000 975
##  $ funded_amnt_inv                           : num  7550 3000 20800 4800 14000 15000 11100 12000 975
##  $ term                                      : Factor w/ 2 levels " 36 months"," 60 months": 1 1 1 1
##  $ int_rate                                  : chr  " 16.24%" " 12.85%" " 13.53%" " 10.99%" ...
##  $ installment                               : num  266 101 706 157 471 ...
##  $ grade                                     : Factor w/ 7 levels "A","B","C","D",..: 3 2 2 2 2 3 3 1
##  $ sub_grade                                 : Factor w/ 35 levels "A1","A2","A3",..: 15 9 10 7 9 12
##  $ emp_title                                 : chr  "Special Order Fulfillment Clerk" "Auditor" "Oper
##  $ emp_length                                : Factor w/ 12 levels "< 1 year","1 year",..: 5 3 3 4 6
##  $ home_ownership                            : chr  "RENT" "RENT" "RENT" "MORTGAGE" ...
##  $ annual_inc                                : num  28000 25000 81500 39600 88000 98000 90000 40000 1
##  $ verification_status                       : Factor w/ 3 levels "Not Verified",..: 1 3 3 2 1 1 1 2
##  $ issue_d                                   : chr  "Dec-2013" "Dec-2013" "Dec-2013" "Dec-2013" ...
##  $ loan_status                               : chr  "Fully Paid" "Fully Paid" "Fully Paid" "Fully Pai
##  $ pymnt_plan                                : chr  "n" "n" "n" "n" ...
##  $ url                                       : logi  NA NA NA NA NA NA ...
##  $ desc                                      : chr  NA NA "  Borrower added on 12/31/13 > My goal is
##  $ purpose                                   : Factor w/ 13 levels "car","credit_card",..: 3 3 3 4 3
##  $ title                                     : chr  "Debt consolidation" "debt" "Reducing Debt to Pur
##  $ zip_code                                  : chr  "951xx" "322xx" "100xx" "782xx" ...
```

```
##  $ addr_state                    : chr  "CA" "FL" "NY" "TX" ...
##  $ dti                           : num  8.4 24.68 16.73 2.49 10.02 ...
##  $ delinq_2yrs                   : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ earliest_cr_line              : chr  "Oct-2010" "May-1991" "Jun-1998" "Aug-1995" ...
##  $ inq_last_6mths                : int  0 0 2 2 0 2 0 0 0 0 ...
##  $ mths_since_last_delinq        : int  NA 58 64 NA 16 NA 16 53 NA 34 ...
##  $ mths_since_last_record        : int  NA 53 NA NA 115 NA NA 33 NA NA ...
##  $ open_acc                      : int  4 5 29 3 6 16 9 7 12 8 ...
##  $ pub_rec                       : int  0 2 0 0 1 0 0 2 0 0 ...
##  $ revol_bal                     : int  5759 2875 23473 4136 3686 5749 6619 5572 7967 114
##  $ revol_util                    : chr  "72%" "54.2%" "54.5%" "16.1%" ...
##  $ total_acc                     : int  5 26 41 8 14 16 12 32 28 29 ...
##  $ initial_list_status           : Factor w/ 2 levels "f","w": 2 1 1 2 1 1 1 2 1 2 ...
##  $ out_prncp                     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ out_prncp_inv                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ total_pymnt                   : num  9600 3182 23927 5158 16945 ...
##  $ total_pymnt_inv               : num  9600 3182 23927 5158 16945 ...
##  $ total_rec_prncp               : num  7550 3000 20800 4800 14000 15000 11100 12000 9750
##  $ total_rec_int                 : num  2050 182 3127 358 2945 ...
##  $ 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                  : chr  "Dec-2016" "Jul-2014" "May-2015" "Sep-2014" ...
##  $ last_pymnt_amnt               : num  530 2677 13335 3900 470 ...
##  $ next_pymnt_d                  : Factor w/ 1 level "Feb-2019": NA NA NA NA NA NA NA NA
##  $ last_credit_pull_d            : chr  "Oct-2018" "Oct-2016" "Jan-2019" "Jan-2017" ...
##  $ collections_12_mths_ex_med    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog   : int  NA 69 71 NA NA NA 16 53 NA 34 ...
##  $ 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  0 154 0 0 0 0 0 15386 0 1514 ...
##  $ tot_cur_bal                   : int  5759 19530 23473 4136 17672 13038 353402 13605 14
##  $ open_acc_6m                   : logi  NA NA NA NA NA NA ...
##  $ open_act_il                   : 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  8000 5300 43100 25700 4500 25800 10000 8100 15100
##  $ 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  1 3 9 0 3 6 2 4 2 3 ...
##  $ avg_cur_bal                   : int  1440 3906 869 1379 2945 815 39267 2268 1177 3892
##  $ bc_open_to_buy                : int  160 2050 6811 21564 480 15051 1016 1428 1752 2969
```

```
##  $ bc_util                                : num  96 52.3 54.6 16.1 87.7 27.6 74.6 79.6 75.7 79.1
##  $ chargeoff_within_12_mths               : int  0 0 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 164 115 104 111 2 NA 124 67 147 ...
##  $ mo_sin_old_rev_tl_op                   : int  38 271 186 220 103 257 150 182 83 189 ...
##  $ mo_sin_rcnt_rev_tl_op                  : int  17 7 0 25 24 7 11 1 12 24 ...
##  $ mo_sin_rcnt_tl                         : int  17 7 0 25 13 2 11 1 12 13 ...
##  $ mort_acc                               : int  0 6 0 0 0 0 1 0 0 4 ...
##  $ mths_since_recent_bc                   : int  17 14 0 25 38 7 11 11 12 24 ...
##  $ mths_since_recent_bc_dlq               : int  NA 69 70 NA 16 NA 35 53 NA 75 ...
##  $ mths_since_recent_inq                  : int  17 8 0 3 NA 2 11 17 20 12 ...
##  $ mths_since_recent_revol_delinq         : int  NA 69 70 NA 16 NA 35 53 NA 75 ...
##  $ num_accts_ever_120_pd                  : int  0 1 1 0 0 0 1 6 0 3 ...
##  $ num_actv_bc_tl                         : int  2 2 8 2 3 8 4 2 6 3 ...
##  $ num_actv_rev_tl                        : int  4 3 24 2 4 8 8 2 7 4 ...
##  $ num_bc_sats                            : int  2 3 11 3 3 13 4 3 6 3 ...
##  $ num_bc_tl                              : int  2 6 17 4 9 13 4 14 11 10 ...
##  $ num_il_tl                              : int  0 11 1 1 3 1 0 8 8 8 ...
##  $ num_op_rev_tl                          : int  4 4 29 3 4 15 8 6 9 6 ...
##  $ num_rev_accts                          : int  5 9 40 7 10 15 11 24 20 17 ...
##  $ num_rev_tl_bal_gt_0                    : int  4 3 24 2 4 8 8 2 7 4 ...
##  $ num_sats                               : int  4 5 29 3 6 16 9 7 12 8 ...
##   [list output truncated]
```

```
#Response variable
table(Loan$loan_status)
```

```
##
##        Charged Off            Current            Default
##              62855              12860                  1
##         Fully Paid    In Grace Period  Late (16-30 days)
##             299845                316                108
## Late (31-120 days)
##                375
```

```
#Get the indices of those loans categorized as Charged Off, Fully Paid, or Default
indx <- which(Loan[["loan_status"]] == c("Charged Off", "Fully Paid"))
length(indx)
```

```
## [1] 181344
```

```
Loan <- Loan[indx,]
```

```
#Converting response variable as a factor of binary outcomes
Loan$loan_status <- ifelse(Loan$loan_status == "Fully Paid", 0, 1)
Loan$loan_status <- as.factor(Loan$loan_status)
table(Loan$loan_status)
```

```
##
##      0      1
## 149937  31407
```

```
## Drop columns that have more than half number of NA values
# specify columns that have no information (including ID columns)
idx <- which(sapply(Loan,function(x) sum(is.na(x))) > (nrow(Loan)*0.5))
Loan_dropped <- Loan[,-idx]
```

```r
# Remove rows that contains NAs
Loan_dropped <- na.omit(Loan_dropped)

dim(Loan_dropped)
```

```
## [1] 134289     87
```

```r
head(Loan_dropped)
```

```
##    loan_amnt funded_amnt funded_amnt_inv       term int_rate installment
## 2       3000        3000            3000  36 months   12.85%      100.87
## 4       4800        4800            4800  36 months   10.99%      157.13
## 6      15000       15000           15000  36 months   14.47%      516.10
## 8      12000       12000           12000  36 months   13.53%      407.40
## 10     15000       15000           15000  36 months    8.90%      476.30
## 14     24000       24000           24000  36 months   13.53%      814.80
##    grade sub_grade                     emp_title emp_length home_ownership
## 2      B        B4                       Auditor  10+ years           RENT
## 4      B        B2           Surgical Technician    2 years       MORTGAGE
## 6      C        C2          building maint. mgr.  10+ years           RENT
## 8      B        B5               On road manager  10+ years           RENT
## 10     A        A5 aircraft maintenance engineer    2 years       MORTGAGE
## 14     B        B5                        driver  10+ years       MORTGAGE
##    annual_inc verification_status  issue_d loan_status pymnt_plan
## 2       25000            Verified Dec-2013           0          n
## 4       39600     Source Verified Dec-2013           0          n
## 6       98000        Not Verified Dec-2013           0          n
## 8       40000     Source Verified Dec-2013           0          n
## 10      63000        Not Verified Dec-2013           0          n
## 14     100000            Verified Dec-2013           0          n
##                purpose              title zip_code addr_state   dti
## 2   debt_consolidation               debt    322xx         FL 24.68
## 4     home_improvement      For The House    782xx         TX  2.49
## 6   debt_consolidation            pay off    117xx         NY  6.15
## 8   debt_consolidation Debt consolidation    871xx         NM 16.94
## 10  debt_consolidation            Pay off    334xx         FL 16.51
## 14         credit_card        credit card    493xx         MI 22.18
##    delinq_2yrs earliest_cr_line inq_last_6mths open_acc pub_rec revol_bal
## 2            0         May-1991              0        5       2      2875
## 4            0         Aug-1995              2        3       0      4136
## 6            0         Jul-1992              2       16       0      5749
## 8            0         Oct-1998              0        7       2      5572
## 10           0         Mar-1998              0        8       0     11431
## 14           0         Jan-1989              0       14       0     21617
##    revol_util total_acc initial_list_status out_prncp out_prncp_inv
## 2       54.2%        26                   f         0             0
## 4       16.1%         8                   w         0             0
## 6       22.3%        16                   f         0             0
## 8       68.8%        32                   w         0             0
## 10      74.2%        29                   w         0             0
## 14      76.7%        39                   w         0             0
##    total_pymnt total_pymnt_inv total_rec_prncp total_rec_int
## 2     3181.549         3181.55            3000        181.55
## 4     5157.519         5157.52            4800        357.52
```

```
## 6    15699.052        15699.05           15000        699.05
## 8    13359.777        13359.78           12000       1359.78
## 10   17146.725        17146.73           15000       2146.73
## 14   28652.210        28652.21           24000       4652.21
##    total_rec_late_fee recoveries collection_recovery_fee last_pymnt_d
## 2                   0          0                      0     Jul-2014
## 4                   0          0                      0     Sep-2014
## 6                   0          0                      0     May-2014
## 8                   0          0                      0     Sep-2015
## 10                  0          0                      0     Jan-2017
## 14                  0          0                      0     Dec-2015
##    last_pymnt_amnt last_credit_pull_d collections_12_mths_ex_med
## 2          2677.23           Oct-2016                         0
## 4          3900.48           Jan-2017                         0
## 6         14150.76           Oct-2018                         0
## 8           119.17           Jan-2019                         0
## 10          476.23           Dec-2016                         0
## 14        10726.61           Jan-2019                         0
##    policy_code application_type acc_now_delinq tot_coll_amt tot_cur_bal
## 2            1       Individual              0          154       19530
## 4            1       Individual              0            0        4136
## 6            1       Individual              0            0       13038
## 8            1       Individual              0        15386       13605
## 10           1       Individual              0         1514      272492
## 14           1       Individual              0          539      199834
##    total_rev_hi_lim acc_open_past_24mths avg_cur_bal bc_open_to_buy
## 2              5300                    3        3906           2050
## 4             25700                    0        1379          21564
## 6             25800                    6         815          15051
## 8              8100                    4        2268           1428
## 10            15400                    3       38927           2969
## 14            28200                    7       15372           4822
##    bc_util chargeoff_within_12_mths delinq_amnt mo_sin_old_il_acct
## 2     52.3                        0           0                164
## 4     16.1                        0           0                104
## 6     27.6                        0           0                  2
## 8     79.6                        0           0                124
## 10    79.1                        0           0                147
## 14    77.6                        0           0                179
##    mo_sin_old_rev_tl_op mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl mort_acc
## 2                   271                     7              7        6
## 4                   220                    25             25        0
## 6                   257                     7              2        0
## 8                   182                     1              1        0
## 10                  189                    24             13        4
## 14                  299                    18              7        3
##    mths_since_recent_bc mths_since_recent_inq num_accts_ever_120_pd
## 2                    14                     8                     1
## 4                    25                     3                     0
## 6                     7                     2                     0
## 8                    11                    17                     6
## 10                   24                    12                     3
## 14                   18                     7                     0
##    num_actv_bc_tl num_actv_rev_tl num_bc_sats num_bc_tl num_il_tl
```

```
## 2               2               3           3         6        11
## 4               2               2           3         4         1
## 6               8               8          13        13         1
## 8               2               2           3        14         8
## 10              3               4           3        10         8
## 14              3               5           5        10        17
##    num_op_rev_tl num_rev_accts num_rev_tl_bal_gt_0 num_sats
## 2              4             9                   3        5
## 4              3             7                   2        3
## 6             15            15                   8       16
## 8              6            24                   2        7
## 10             6            17                   4        8
## 14             8            19                   5       14
##    num_tl_120dpd_2m num_tl_30dpd num_tl_90g_dpd_24m num_tl_op_past_12m
## 2                 0            0                  0                  1
## 4                 0            0                  0                  0
## 6                 0            0                  0                  2
## 8                 0            0                  0                  2
## 10                0            0                  0                  0
## 14                0            0                  0                  2
##    pct_tl_nvr_dlq percent_bc_gt_75 pub_rec_bankruptcies tax_liens
## 2            91.3             66.7                   2         0
## 4           100.0              0.0                   0         0
## 6           100.0              7.7                   0         0
## 8            81.2             33.3                   0         0
## 10           89.3             66.7                   0         0
## 14          100.0             75.0                   0         0
##    tot_hi_cred_lim total_bal_ex_mort total_bc_limit
## 2            32082             19530          4300
## 4            25700              4136         25700
## 6            33300             13038         20800
## 8            18130             13605          7000
## 10          288195             39448         14200
## 14          229072             61397         21500
##    total_il_high_credit_limit hardship_flag disbursement_method
## 2                       26782            N                Cash
## 4                           0            N                Cash
## 6                        7500            N                Cash
## 8                       10030            N                Cash
## 10                      33895            N                Cash
## 14                      58847            N                Cash
##    debt_settlement_flag
## 2                     N
## 4                     N
## 6                     N
## 8                     N
## 10                    N
## 14                    N
```

```r
#Converting interest rate variable to numeric
Loan_dropped$int_rate <- as.numeric(sub("%","",Loan_dropped$int_rate))/100
Loan_dropped$revol_util <- as.numeric(sub("%","",Loan_dropped$revol_util))/100

Loan_dropped$emp_length <- as.factor(Loan_dropped$emp_length)
```

```r
Loan_dropped$home_ownership <- as.factor(Loan_dropped$home_ownership)
```

```r
#Drop character variables
Loan_dropped <- Loan_dropped[, !sapply(Loan_dropped, is.character)]

#Drop variables that are caused by default status
Loan_dropped <- Loan_dropped[ , !names(Loan_dropped) %in% c('recoveries', 'collection_recovery_fee', 'd

#Remove all zeros columns
Loan_dropped <- Loan_dropped[ , !names(Loan_dropped) %in% c("out_prncp","out_prncp_inv","policy_code","

#Remove categorical variable sub_grade to avoid new categories in test data
Loan_dropped <- Loan_dropped[, !names(Loan_dropped)== 'sub_grade']
```

```r
dim(Loan_dropped)
```

```
## [1] 134289     68
```

```r
summary(Loan_dropped)
```

```
##    loan_amnt      funded_amnt    funded_amnt_inv         term
##  Min.   : 1000   Min.   : 1000   Min.   : 1000    36 months:98161
##  1st Qu.: 8400   1st Qu.: 8400   1st Qu.: 8400    60 months:36128
##  Median :13000   Median :13000   Median :13000
##  Mean   :14791   Mean   :14791   Mean   :14784
##  3rd Qu.:20000   3rd Qu.:20000   3rd Qu.:20000
##  Max.   :35000   Max.   :35000   Max.   :35000
##
##     int_rate       installment      grade         emp_length
##  Min.   :0.0600   Min.   :  27.85   A:20103   10+ years:46688
##  1st Qu.:0.1099   1st Qu.: 273.11   B:39551   2 years  :12297
##  Median :0.1398   Median : 398.38   C:37017   3 years  :10726
##  Mean   :0.1401   Mean   : 449.46   D:22270   < 1 year :10268
##  3rd Qu.:0.1699   3rd Qu.: 587.34   E:10355   5 years  : 8936
##  Max.   :0.2606   Max.   :1408.13   F: 4054   1 year   : 8609
##                                     G:  939   (Other)  :36765
##   home_ownership    annual_inc           verification_status loan_status
##  MORTGAGE:71789   Min.   :   4000   Not Verified   :42807    0:110817
##  NONE    :   16   1st Qu.:  48000   Source Verified:45109    1: 23472
##  OTHER   :   15   Median :  65000   Verified       :46373
##  OWN     :11483   Mean   :  75954
##  RENT    :50986   3rd Qu.:  90000
##                   Max.   :7500000
##
##               purpose           dti         delinq_2yrs
##  debt_consolidation:82302   Min.   : 0.00   Min.   : 0.0000
##  credit_card       :30598   1st Qu.:12.20   1st Qu.: 0.0000
##  home_improvement  : 7430   Median :17.66   Median : 0.0000
##  other             : 5823   Mean   :18.03   Mean   : 0.3059
##  major_purchase    : 2310   3rd Qu.:23.61   3rd Qu.: 0.0000
##  small_business    : 1274   Max.   :39.99   Max.   :21.0000
##  (Other)           : 4552
##  inq_last_6mths      open_acc        pub_rec          revol_bal
##  Min.   :0.0000   Min.   : 1.00   Min.   : 0.0000   Min.   :      0
##  1st Qu.:0.0000   1st Qu.: 8.00   1st Qu.: 0.0000   1st Qu.:   6858
```

```
##  Median :1.0000   Median :11.00   Median : 0.0000   Median :  12241
##  Mean   :0.8493   Mean   :11.87   Mean   : 0.1682   Mean   :  16796
##  3rd Qu.:1.0000   3rd Qu.:14.00   3rd Qu.: 0.0000   3rd Qu.:  20916
##  Max.   :8.0000   Max.   :84.00   Max.   :54.0000   Max.   :2568995
##
##    revol_util       total_acc      initial_list_status  total_pymnt
##  Min.   :0.0000   Min.   :  3.00   f:79504             Min.   :   35.79
##  1st Qu.:0.4010   1st Qu.: 18.00   w:54785             1st Qu.: 8395.40
##  Median :0.5750   Median : 25.00                       Median :13836.67
##  Mean   :0.5648   Mean   : 26.41                       Mean   :16380.10
##  3rd Qu.:0.7410   3rd Qu.: 33.00                       3rd Qu.:22280.60
##  Max.   :1.5070   Max.   :156.00                       Max.   :62862.51
##
##  total_pymnt_inv    total_rec_prncp total_rec_int      total_rec_late_fee
##  Min.   :   35.79   Min.   :    0   Min.   :    1.34   Min.   :  0.000
##  1st Qu.: 8392.49   1st Qu.: 6250   1st Qu.: 1124.58   1st Qu.:  0.000
##  Median :13829.10   Median :11125   Median : 2096.78   Median :  0.000
##  Mean   :16373.08   Mean   :12965   Mean   : 3173.85   Mean   :  1.175
##  3rd Qu.:22271.03   3rd Qu.:18000   3rd Qu.: 3955.54   3rd Qu.:  0.000
##  Max.   :62862.51   Max.   :35000   Max.   :27862.51   Max.   :455.760
##
##  last_pymnt_amnt   collections_12_mths_ex_med acc_now_delinq
##  Min.   :    0.0   Min.   : 0.00000           Min.   :0.000000
##  1st Qu.:  395.4   1st Qu.: 0.00000           1st Qu.:0.000000
##  Median : 1486.4   Median : 0.00000           Median :0.000000
##  Mean   : 4777.3   Mean   : 0.01133           Mean   :0.004706
##  3rd Qu.: 7355.2   3rd Qu.: 0.00000           3rd Qu.:0.000000
##  Max.   :36234.4   Max.   :20.00000           Max.   :5.000000
##
##   tot_coll_amt      tot_cur_bal      total_rev_hi_lim
##  Min.   :      0   Min.   :      0   Min.   :    300
##  1st Qu.:      0   1st Qu.:  32447   1st Qu.:  14000
##  Median :      0   Median :  94550   Median :  23300
##  Mean   :    230   Mean   : 147201   Mean   :  30938
##  3rd Qu.:      0   3rd Qu.: 219742   3rd Qu.:  38500
##  Max.   :9152545   Max.   :4772549   Max.   :9999999
##
##  acc_open_past_24mths  avg_cur_bal     bc_open_to_buy      bc_util
##  Min.   : 0.000       Min.   :     0   Min.   :     0   Min.   :  0.0
##  1st Qu.: 3.000       1st Qu.:  3330   1st Qu.:  1091   1st Qu.: 47.2
##  Median : 4.000       Median :  8361   Median :  3683   Median : 69.9
##  Mean   : 4.514       Mean   : 13956   Mean   :  8477   Mean   : 65.4
##  3rd Qu.: 6.000       3rd Qu.: 19731   3rd Qu.: 10032   3rd Qu.: 88.0
##  Max.   :53.000       Max.   :502002   Max.   :278899   Max.   :197.0
##
##  chargeoff_within_12_mths  delinq_amnt       mo_sin_old_il_acct
##  Min.   :0.000000         Min.   :    0.00   Min.   :  0.0
##  1st Qu.:0.000000         1st Qu.:    0.00   1st Qu.: 97.0
##  Median :0.000000         Median :    0.00   Median :129.0
##  Mean   :0.009085         Mean   :    8.21   Mean   :126.1
##  3rd Qu.:0.000000         3rd Qu.:    0.00   3rd Qu.:152.0
##  Max.   :4.000000         Max.   :65000.00   Max.   :649.0
##
##  mo_sin_old_rev_tl_op mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl
```

```
##  Min.   :  4.0        Min.   :  0.00        Min.   :  0.000
##  1st Qu.:117.0        1st Qu.:  4.00        1st Qu.:  3.000
##  Median :162.0        Median :  8.00        Median :  6.000
##  Mean   :177.8        Mean   : 12.33        Mean   :  7.429
##  3rd Qu.:223.0        3rd Qu.: 15.00        3rd Qu.: 10.000
##  Max.   :818.0        Max.   :372.00        Max.   :121.000
##
##     mort_acc      mths_since_recent_bc mths_since_recent_inq
##  Min.   : 0.000   Min.   :  0.00       Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:  6.00       1st Qu.: 2.000
##  Median : 1.000   Median : 14.00       Median : 5.000
##  Mean   : 1.896   Mean   : 23.47       Mean   : 6.936
##  3rd Qu.: 3.000   3rd Qu.: 28.00       3rd Qu.:10.000
##  Max.   :34.000   Max.   :538.00       Max.   :24.000
##
##  num_accts_ever_120_pd num_actv_bc_tl   num_actv_rev_tl   num_bc_sats
##  Min.   : 0.0000       Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.0000       1st Qu.: 2.000   1st Qu.: 4.000   1st Qu.: 3.000
##  Median : 0.0000       Median : 3.000   Median : 5.000   Median : 4.000
##  Mean   : 0.4479       Mean   : 3.756   Mean   : 5.806   Mean   : 4.686
##  3rd Qu.: 0.0000       3rd Qu.: 5.000   3rd Qu.: 7.000   3rd Qu.: 6.000
##  Max.   :33.0000       Max.   :26.000   Max.   :38.000   Max.   :35.000
##
##    num_bc_tl        num_il_tl       num_op_rev_tl    num_rev_accts
##  Min.   : 1.000   Min.   :  1.000   Min.   : 1.000   Min.   :  1.00
##  1st Qu.: 5.000   1st Qu.:  4.000   1st Qu.: 5.000   1st Qu.: 10.00
##  Median : 8.000   Median :  7.000   Median : 8.000   Median : 14.00
##  Mean   : 8.886   Mean   :  8.762   Mean   : 8.405   Mean   : 15.45
##  3rd Qu.:11.000   3rd Qu.: 12.000   3rd Qu.:11.000   3rd Qu.: 20.00
##  Max.   :65.000   Max.   :150.000   Max.   :58.000   Max.   :105.00
##
##  num_rev_tl_bal_gt_0    num_sats     num_tl_120dpd_2m
##  Min.   : 0.000      Min.   : 1.00   Min.   :0.0000000
##  1st Qu.: 4.000      1st Qu.: 8.00   1st Qu.:0.0000000
##  Median : 5.000      Median :11.00   Median :0.0000000
##  Mean   : 5.823      Mean   :11.84   Mean   :0.0007298
##  3rd Qu.: 7.000      3rd Qu.:14.00   3rd Qu.:0.0000000
##  Max.   :38.000      Max.   :84.00   Max.   :2.0000000
##
##   num_tl_30dpd      num_tl_90g_dpd_24m num_tl_op_past_12m pct_tl_nvr_dlq
##  Min.   :0.000000   Min.   : 0.00000   Min.   : 0.000     Min.   : 16.00
##  1st Qu.:0.000000   1st Qu.: 0.00000   1st Qu.: 1.000     1st Qu.: 92.30
##  Median :0.000000   Median : 0.00000   Median : 2.000     Median :100.00
##  Mean   :0.003061   Mean   : 0.07833   Mean   : 2.053     Mean   : 94.91
##  3rd Qu.:0.000000   3rd Qu.: 0.00000   3rd Qu.: 3.000     3rd Qu.:100.00
##  Max.   :3.000000   Max.   :20.00000   Max.   :23.000     Max.   :100.00
##
##  percent_bc_gt_75 pub_rec_bankruptcies   tax_liens
##  Min.   :  0.0    Min.   :0.000        Min.   : 0.00000
##  1st Qu.: 25.0    1st Qu.:0.000        1st Qu.: 0.00000
##  Median : 50.0    Median :0.000        Median : 0.00000
##  Mean   : 51.7    Mean   :0.112        Mean   : 0.03613
##  3rd Qu.: 80.0    3rd Qu.:0.000        3rd Qu.: 0.00000
##  Max.   :100.0    Max.   :7.000        Max.   :53.00000
```

```
##
##  tot_hi_cred_lim   total_bal_ex_mort total_bc_limit
##  Min.   :    500   Min.   :      0   Min.   :   100
##  1st Qu.:  51614   1st Qu.:  22411   1st Qu.:  7559
##  Median : 123621   Median :  37733   Median : 14500
##  Mean   : 177609   Mean   :  49108   Mean   : 20462
##  3rd Qu.: 257112   3rd Qu.:  61072   3rd Qu.: 26700
##  Max.   :9999999   Max.   :2644442   Max.   :760000
##
##  total_il_high_credit_limit
##  Min.   :      0
##  1st Qu.:  15239
##  Median :  30772
##  Mean   :  40546
##  3rd Qu.:  53885
##  Max.   :1241783
##
```

```
# Split data to training and testing
set.seed(12345)
n <- nrow(Loan_dropped)
train_ind <- sample.int(n, size = round(n*0.75))

Loan_train <- Loan_dropped[train_ind, ]
Loan_test <- Loan_dropped[-train_ind, ]
```

Step 3: Fit model

```
library(caret)
library(Metrics)
library(ROCR)
```

```
## Warning: package 'gplots' was built under R version 3.5.2
```

kNN

```
# create normalization function
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

# drop other categorical variables
train_num <- Loan_train[, !sapply(Loan_train, is.factor)]
test_num <- Loan_test[, !sapply(Loan_test, is.factor)]

# normalize the train and test data
train_norm <- as.data.frame(lapply(train_num, normalize))
test_norm <- as.data.frame(lapply(test_num, normalize))

# create labels for training and test data
train_label <- Loan_train$loan_status
test_label <- Loan_test$loan_status
```

```r
summary(train_norm)
```

```
##    loan_amnt       funded_amnt     funded_amnt_inv     int_rate
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2176   1st Qu.:0.2176   1st Qu.:0.2176   1st Qu.:0.2488
##  Median :0.3529   Median :0.3529   Median :0.3529   Median :0.3978
##  Mean   :0.4056   Mean   :0.4056   Mean   :0.4054   Mean   :0.3988
##  3rd Qu.:0.5588   3rd Qu.:0.5588   3rd Qu.:0.5588   3rd Qu.:0.5479
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##   installment       annual_inc            dti          delinq_2yrs
##  Min.   :0.0000   Min.   :0.000000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.1779   1st Qu.:0.005870   1st Qu.:0.3047   1st Qu.:0.00000
##  Median :0.2679   Median :0.008138   Median :0.4412   Median :0.00000
##  Mean   :0.3055   Mean   :0.009633   Mean   :0.4506   Mean   :0.01461
##  3rd Qu.:0.4055   3rd Qu.:0.011473   3rd Qu.:0.5905   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.000000   Max.   :1.0000   Max.   :1.00000
##  inq_last_6mths      open_acc          pub_rec           revol_bal
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.08434   1st Qu.:0.000000   1st Qu.:0.002672
##  Median :0.1250   Median :0.12048   Median :0.000000   Median :0.004775
##  Mean   :0.1064   Mean   :0.13110   Mean   :0.003127   Mean   :0.006548
##  3rd Qu.:0.1250   3rd Qu.:0.15663   3rd Qu.:0.000000   3rd Qu.:0.008152
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.000000   Max.   :1.000000
##    revol_util       total_acc        total_pymnt     total_pymnt_inv
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.2661   1st Qu.:0.09804   1st Qu.:0.1326   1st Qu.:0.1326
##  Median :0.3809   Median :0.14379   Median :0.2196   Median :0.2195
##  Mean   :0.3746   Mean   :0.15306   Mean   :0.2600   Mean   :0.2599
##  3rd Qu.:0.4917   3rd Qu.:0.19608   3rd Qu.:0.3541   3rd Qu.:0.3539
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##  total_rec_prncp  total_rec_int     total_rec_late_fee last_pymnt_amnt
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.1794   1st Qu.:0.04026   1st Qu.:0.000000   1st Qu.:0.01098
##  Median :0.3200   Median :0.07526   Median :0.000000   Median :0.04153
##  Mean   :0.3710   Mean   :0.11387   Mean   :0.002543   Mean   :0.13275
##  3rd Qu.:0.5143   3rd Qu.:0.14194   3rd Qu.:0.000000   3rd Qu.:0.20449
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000
##  collections_12_mths_ex_med acc_now_delinq       tot_coll_amt
##  Min.   :0.0000000          Min.   :0.0000000   Min.   :0.00e+00
##  1st Qu.:0.0000000          1st Qu.:0.0000000   1st Qu.:0.00e+00
##  Median :0.0000000          Median :0.0000000   Median :0.00e+00
##  Mean   :0.0005704          Mean   :0.0009552   Mean   :2.77e-05
##  3rd Qu.:0.0000000          3rd Qu.:0.0000000   3rd Qu.:0.00e+00
##  Max.   :1.0000000          Max.   :1.0000000   Max.   :1.00e+00
##   tot_cur_bal       total_rev_hi_lim   acc_open_past_24mths
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.006814   1st Qu.:0.001370   1st Qu.:0.05660
##  Median :0.019905   Median :0.002310   Median :0.07547
##  Mean   :0.030909   Mean   :0.003072   Mean   :0.08523
##  3rd Qu.:0.046102   3rd Qu.:0.003830   3rd Qu.:0.11321
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.00000
##   avg_cur_bal       bc_open_to_buy        bc_util
##  Min.   :0.000000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:0.006708   1st Qu.:0.003923   1st Qu.:0.2396
```

```
##  Median :0.016831   Median :0.013227   Median :0.3548
##  Mean   :0.028059   Mean   :0.030478   Mean   :0.3319
##  3rd Qu.:0.039613   3rd Qu.:0.036038   3rd Qu.:0.4467
##  Max.   :1.000000   Max.   :1.000000   Max.   :1.0000
##  chargeoff_within_12_mths  delinq_amnt        mo_sin_old_il_acct
##  Min.   :0.000000         Min.   :0.0000000   Min.   :0.0000
##  1st Qu.:0.000000         1st Qu.:0.0000000   1st Qu.:0.1481
##  Median :0.000000         Median :0.0000000   Median :0.1975
##  Mean   :0.002289         Mean   :0.0001321   Mean   :0.1933
##  3rd Qu.:0.000000         3rd Qu.:0.0000000   3rd Qu.:0.2330
##  Max.   :1.000000         Max.   :1.0000000   Max.   :1.0000
##  mo_sin_old_rev_tl_op mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl
##  Min.   :0.0000       Min.   :0.00000       Min.   :0.00000
##  1st Qu.:0.1485       1st Qu.:0.01075       1st Qu.:0.02479
##  Median :0.2082       Median :0.02151       Median :0.04959
##  Mean   :0.2295       Mean   :0.03317       Mean   :0.06134
##  3rd Qu.:0.2891       3rd Qu.:0.04032       3rd Qu.:0.08264
##  Max.   :1.0000       Max.   :1.00000       Max.   :1.00000
##     mort_acc       mths_since_recent_bc mths_since_recent_inq
##  Min.   :0.00000   Min.   :0.00000      Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.01115      1st Qu.:0.08333
##  Median :0.02941   Median :0.02602      Median :0.20833
##  Mean   :0.05589   Mean   :0.04376      Mean   :0.28881
##  3rd Qu.:0.08824   3rd Qu.:0.05390      3rd Qu.:0.41667
##  Max.   :1.00000   Max.   :1.00000      Max.   :1.00000
##  num_accts_ever_120_pd num_actv_bc_tl    num_actv_rev_tl
##  Min.   :0.00000       Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000       1st Qu.:0.08696   1st Qu.:0.1053
##  Median :0.00000       Median :0.13043   Median :0.1316
##  Mean   :0.01874       Mean   :0.16345   Mean   :0.1529
##  3rd Qu.:0.00000       3rd Qu.:0.21739   3rd Qu.:0.1842
##  Max.   :1.00000       Max.   :1.00000   Max.   :1.0000
##   num_bc_sats        num_bc_tl        num_il_tl       num_op_rev_tl
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.08571   1st Qu.:0.0625   1st Qu.:0.02013   1st Qu.:0.07018
##  Median :0.11429   Median :0.1094   Median :0.04027   Median :0.12281
##  Mean   :0.13392   Mean   :0.1232   Mean   :0.05210   Mean   :0.13002
##  3rd Qu.:0.17143   3rd Qu.:0.1562   3rd Qu.:0.07383   3rd Qu.:0.17544
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##  num_rev_accts     num_rev_tl_bal_gt_0    num_sats
##  Min.   :0.00000   Min.   :0.0000      Min.   :0.00000
##  1st Qu.:0.08654   1st Qu.:0.1053      1st Qu.:0.08434
##  Median :0.12500   Median :0.1316      Median :0.12048
##  Mean   :0.13898   Mean   :0.1534      Mean   :0.13073
##  3rd Qu.:0.18269   3rd Qu.:0.1842      3rd Qu.:0.15663
##  Max.   :1.00000   Max.   :1.0000      Max.   :1.00000
##  num_tl_120dpd_2m     num_tl_30dpd      num_tl_90g_dpd_24m
##  Min.   :0.0000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.0000000   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.0000000   Median :0.000000   Median :0.000000
##  Mean   :0.0003823   Mean   :0.001039   Mean   :0.003956
##  3rd Qu.:0.0000000   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.0000000   Max.   :1.000000   Max.   :1.000000
##  num_tl_op_past_12m pct_tl_nvr_dlq   percent_bc_gt_75 pub_rec_bankruptcies
```

```
##  Min.   :0.00000    Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.04348    1st Qu.:0.9083   1st Qu.:0.2500   1st Qu.:0.00000
##  Median :0.08696    Median :1.0000   Median :0.5000   Median :0.00000
##  Mean   :0.08933    Mean   :0.9394   Mean   :0.5172   Mean   :0.01602
##  3rd Qu.:0.13043    3rd Qu.:1.0000   3rd Qu.:0.8000   3rd Qu.:0.00000
##  Max.   :1.00000    Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##    tax_liens         tot_hi_cred_lim    total_bal_ex_mort
##  Min.   :0.0000000   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.0000000   1st Qu.:0.005112   1st Qu.:0.008478
##  Median :0.0000000   Median :0.012327   Median :0.014303
##  Mean   :0.0006858   Mean   :0.017741   Mean   :0.018599
##  3rd Qu.:0.0000000   3rd Qu.:0.025681   3rd Qu.:0.023146
##  Max.   :1.0000000   Max.   :1.000000   Max.   :1.000000
##  total_bc_limit     total_il_high_credit_limit
##  Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.009738   1st Qu.:0.01229
##  Median :0.018950   Median :0.02482
##  Mean   :0.026863   Mean   :0.03268
##  3rd Qu.:0.035136   3rd Qu.:0.04340
##  Max.   :1.000000   Max.   :1.00000
```

```
#Train model
library(class)
```

```
## Warning: package 'class' was built under R version 3.5.2
```

```
loan_pred <- knn(train = train_norm, test = test_norm,
                      cl = train_label, k = 200, prob = TRUE)

#prediction result - class
head(loan_pred)
```

```
## [1] 0 0 0 0 0 0
## Levels: 0 1
```

```
# prediction result - probability
pred.knn <- attributes(loan_pred)$prob



#Evaluate model performance
#Area under the ROC curve
auc.knn <- auc(actual = Loan_test$loan_status, predicted = pred.knn)
sprintf("kNN Test AUC: %.3f", auc.knn)
```

```
## [1] "kNN Test AUC: 0.234"
```

```
#Confusion matrix
confusionMatrix(data = loan_pred, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 27550  2901
##          1     0  3121
##
##                Accuracy : 0.9136
```

```
##                  95% CI : (0.9105, 0.9166)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6384
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.5183
##          Pos Pred Value : 0.9047
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8206
##          Detection Rate : 0.8206
##    Detection Prevalence : 0.9070
##       Balanced Accuracy : 0.7591
##
##        'Positive' Class : 0
##
```

C5.0

```
#Fit model
library(C50)
model.c50 <- C5.0(loan_status~., data = Loan_train, rules = FALSE)

#Prediction
pred.c50 <- predict(model.c50, Loan_test, type='prob')
head(pred.c50)
```

```
##            0           1
## 6  0.9995770 0.0004229613
## 10 0.9990296 0.0009704365
## 28 0.9995770 0.0004229613
## 32 0.9995770 0.0004229613
## 36 0.9935564 0.0064436318
## 38 0.9995770 0.0004229613
```

```
pr <- pred.c50[,'1']

#Evaluate model performance
#Area under the ROC curve
auc.c50 <- auc(actual = Loan_test$loan_status, predicted = pr)
sprintf("C5.0 Tree Test AUC: %.3f", auc.c50)
```

```
## [1] "C5.0 Tree Test AUC: 0.998"
```

```
#Converting prediction result to binary class
class.c50 <- ifelse(pr > 0.5, 1, 0)
class.c50 <- as.factor(class.c50)

#Confusion matrix
confusionMatrix(data = class.c50, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction     0     1
##          0 27546    89
##          1     4  5933
##
##                Accuracy : 0.9972
##                  95% CI : (0.9966, 0.9978)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9905
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9999
##             Specificity : 0.9852
##          Pos Pred Value : 0.9968
##          Neg Pred Value : 0.9993
##              Prevalence : 0.8206
##          Detection Rate : 0.8205
##    Detection Prevalence : 0.8232
##       Balanced Accuracy : 0.9925
##
##        'Positive' Class : 0
##
```

Naive Bayes

```
#Fit model
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.5.2
```

```
model.nb <- naiveBayes(loan_status~., data = Loan_train, laplace = 1, prob=TRUE)

#Prediction
pred.nb <- predict(model.nb, Loan_test, type = 'raw')
head(pred.nb)
```

```
##              0             1
## [1,] 1.0000000 2.429523e-210
## [2,] 0.9133796  8.662037e-02
## [3,] 0.6851178  3.148822e-01
## [4,] 1.0000000 9.900778e-106
## [5,] 0.2084170  7.915830e-01
## [6,] 1.0000000 1.177332e-130
```

```
#Evaluate model performance
#Area under the ROC curve
auc.nb <- auc(actual = Loan_test$loan_status, predicted = pred.nb[,'1'])
sprintf("Naive Bayes Test AUC: %.3f", auc.nb)
```

```
## [1] "Naive Bayes Test AUC: 0.904"
```

```
pr <- pred.nb[,'1']

#Converting prediction result to binary class
class.nb <- ifelse(pr > 0.5, 1, 0)
```

```r
class.nb <- as.factor(class.nb)

#Confusion matrix
confusionMatrix(data = class.nb, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 19264   509
##          1  8286  5513
##
##                Accuracy : 0.738
##                  95% CI : (0.7333, 0.7427)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4086
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.6992
##             Specificity : 0.9155
##          Pos Pred Value : 0.9743
##          Neg Pred Value : 0.3995
##              Prevalence : 0.8206
##          Detection Rate : 0.5738
##    Detection Prevalence : 0.5890
##       Balanced Accuracy : 0.8074
##
##        'Positive' Class : 0
##
```

Logistic Regression

```r
library(dplyr)
library(stringr)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
##
## Attaching package: 'glmnet'
```

```
## The following object is masked from 'package:Metrics':
##
##     auc
```

```r
library(Matrix)

trainX <- model.matrix(loan_status~., data=Loan_train)[,-1]
trainY <- Loan_train$loan_status
```

```r
#update new levels
levels(Loan_test$purpose) <- levels(Loan_train$purpose)
#Turn test data to sparse matrix for same dimension as train data
test <- model.matrix(loan_status~., data=Loan_test)[,-1]
testX <- as.matrix(test)

#Fit model - using LASSO for regularization
mod.logit <-  glmnet(x=trainX, y=trainY,family="binomial",alpha=1, lambda = 0.001, standardize = FALSE )

#Prediction
pred.logit <- predict(mod.logit, testX, type = 'response')
head(pred.logit)
```

```
##              s0
## 6  1.302526e-06
## 10 4.475019e-05
## 28 7.903351e-08
## 32 2.213548e-05
## 36 4.946062e-05
## 38 1.678097e-06
```

```r
detach("package:glmnet", unload=TRUE)


#Evaluate model performance
#Area under the ROC curve22
auc.logit <- auc(actual = Loan_test$loan_status, predicted = pred.logit)
sprintf("Logistic Regression Test AUC: %.3f", auc.logit)
```

```
## [1] "Logistic Regression Test AUC: 1.000"
```

```r
#Converting prediction result to binary class
class.logit <- ifelse(pred.logit > 0.5, 1, 0)
class.logit <- as.factor(class.logit)

confusionMatrix(data = class.logit, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 27549     8
##          1     1  6014
##
##                Accuracy : 0.9997
##                  95% CI : (0.9995, 0.9999)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9991
##  Mcnemar's Test P-Value : 0.0455
##
##             Sensitivity : 1.0000
##             Specificity : 0.9987
##          Pos Pred Value : 0.9997
```

```
##          Neg Pred Value : 0.9998
##              Prevalence : 0.8206
##          Detection Rate : 0.8206
##    Detection Prevalence : 0.8208
##       Balanced Accuracy : 0.9993
##
##        'Positive' Class : 0
##
```

Cross validation for glmnet (LASSO)

```
#Fit the model - using LASSO
# Using caret to perform CV
myControl <- trainControl(
  method = "cv", number = 5,
  summaryFunction = twoClassSummary,
  classProbs = TRUE, # IMPORTANT!
  verboseIter = TRUE
)

model.glmnet <- train(
  x = trainX, y=make.names(trainY),
  metric = "ROC",
  tuneGrid = expand.grid(alpha = 1, lambda = seq(0.0001, 1, length = 20)),
  method = "glmnet",
  trControl = myControl)
```

```
## + Fold1: alpha=1, lambda=1
## - Fold1: alpha=1, lambda=1
## + Fold2: alpha=1, lambda=1
## - Fold2: alpha=1, lambda=1
## + Fold3: alpha=1, lambda=1
## - Fold3: alpha=1, lambda=1
## + Fold4: alpha=1, lambda=1
## - Fold4: alpha=1, lambda=1
## + Fold5: alpha=1, lambda=1
## - Fold5: alpha=1, lambda=1
## Aggregating results
## Selecting tuning parameters
## Fitting alpha = 1, lambda = 1e-04 on full training set
```

```
model.glmnet
```

```
## glmnet
##
## 100717 samples
##     97 predictor
##      2 classes: 'X0', 'X1'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 80574, 80574, 80573, 80573, 80574
## Resampling results across tuning parameters:
##
##   lambda      ROC        Sens      Spec
##   0.00010000  0.9998202  0.999964  0.9841834
```

```
##   0.05272632  0.9703967  1.000000  0.2841834
##   0.10535263  0.8497963  1.000000  0.0000000
##   0.15797895  0.5000000  1.000000  0.0000000
##   0.21060526  0.5000000  1.000000  0.0000000
##   0.26323158  0.5000000  1.000000  0.0000000
##   0.31585789  0.5000000  1.000000  0.0000000
##   0.36848421  0.5000000  1.000000  0.0000000
##   0.42111053  0.5000000  1.000000  0.0000000
##   0.47373684  0.5000000  1.000000  0.0000000
##   0.52636316  0.5000000  1.000000  0.0000000
##   0.57898947  0.5000000  1.000000  0.0000000
##   0.63161579  0.5000000  1.000000  0.0000000
##   0.68424211  0.5000000  1.000000  0.0000000
##   0.73686842  0.5000000  1.000000  0.0000000
##   0.78949474  0.5000000  1.000000  0.0000000
##   0.84212105  0.5000000  1.000000  0.0000000
##   0.89474737  0.5000000  1.000000  0.0000000
##   0.94737368  0.5000000  1.000000  0.0000000
##   1.00000000  0.5000000  1.000000  0.0000000
##
## Tuning parameter 'alpha' was held constant at a value of 1
## ROC was used to select the optimal model using the largest value.
## The final values used for the model were alpha = 1 and lambda = 1e-04.
```

```
# best parameter
model.glmnet$bestTune
```

```
##   alpha lambda
## 1     1  1e-04
```

```
# best coefficient
coef(model.glmnet$finalModel, model.glmnet$bestTune$lambda)
```

```
## 98 x 1 sparse Matrix of class "dgCMatrix"
##                                           1
## (Intercept)                    -4.206796e+00
## loan_amnt                       5.700434e-03
## funded_amnt                     7.172785e-08
## funded_amnt_inv                 .
## term 60 months                  .
## int_rate                        3.908717e+00
## installment                     3.569054e-07
## gradeB                          .
## gradeC                          3.566373e-02
## gradeD                          .
## gradeE                          .
## gradeF                          .
## gradeG                          .
## emp_length1 year               -7.313710e-03
## emp_length10+ years             4.128378e-02
## emp_length2 years               .
## emp_length3 years              -1.284324e-01
## emp_length4 years               .
## emp_length5 years              -2.848723e-02
## emp_length6 years              -1.446423e-01
## emp_length7 years               .
```

```
## emp_length8 years                    5.987114e-02
## emp_length9 years                    .
## emp_lengthn/a                        1.452174e+00
## home_ownershipNONE                   .
## home_ownershipOTHER                  .
## home_ownershipOWN                    1.862761e-01
## home_ownershipRENT                   .
## annual_inc                           .
## verification_statusSource Verified   .
## verification_statusVerified          .
## purposecredit_card                  -2.828167e-02
## purposedebt_consolidation            2.042271e-02
## purposehome_improvement              .
## purposehouse                         .
## purposemajor_purchase                .
## purposemedical                      -8.989260e-02
## purposemoving                        2.185532e-01
## purposeother                         7.713804e-03
## purposerenewable_energy              1.563803e-01
## purposesmall_business                2.471433e-02
## purposevacation                      .
## purposewedding                       .
## dti                                  6.170537e-03
## delinq_2yrs                          .
## inq_last_6mths                       .
## open_acc                             .
## pub_rec                              6.751792e-02
## revol_bal                            .
## revol_util                           .
## total_acc                            .
## initial_list_statusw                 .
## total_pymnt                          .
## total_pymnt_inv                      .
## total_rec_prncp                     -5.813997e-03
## total_rec_int                        .
## total_rec_late_fee                   2.286334e-02
## last_pymnt_amnt                     -5.649901e-04
## collections_12_mths_ex_med           1.181940e-01
## acc_now_delinq                       .
## tot_coll_amt                         .
## tot_cur_bal                         -3.770332e-07
## total_rev_hi_lim                     .
## acc_open_past_24mths                 2.782219e-02
## avg_cur_bal                          .
## bc_open_to_buy                       .
## bc_util                             -1.292632e-03
## 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                -1.763935e-03
```

```
## mths_since_recent_inq               .
## num_accts_ever_120_pd               2.448968e-02
## num_actv_bc_tl                      .
## num_actv_rev_tl                     3.902840e-06
## num_bc_sats                         .
## num_bc_tl                           .
## num_il_tl                           .
## num_op_rev_tl                       .
## num_rev_accts                      -7.699913e-03
## num_rev_tl_bal_gt_0                 2.533854e-02
## num_sats                            .
## num_tl_120dpd_2m                    .
## num_tl_30dpd                        9.773363e-02
## num_tl_90g_dpd_24m                  2.438748e-02
## num_tl_op_past_12m                  .
## pct_tl_nvr_dlq                      .
## percent_bc_gt_75                    .
## pub_rec_bankruptcies                .
## tax_liens                          -3.269176e-02
## tot_hi_cred_lim                    -3.159391e-07
## total_bal_ex_mort                   .
## total_bc_limit                      .
## total_il_high_credit_limit          .
```

Random Forest

```
library(randomForest)
#Fit model
model.rf <- randomForest(loan_status~., data=Loan_train, ntree=200, prob=TRUE)

#Prediction
pred.rf <- predict(model.rf, Loan_test, type = "prob")
head(pred.rf)
```

```
##        0     1
## 6  0.955 0.045
## 10 1.000 0.000
## 28 1.000 0.000
## 32 0.975 0.025
## 36 0.970 0.030
## 38 1.000 0.000
```

```
#Evaluate model performance
#Area under the ROC curve
auc.rf <- auc(actual = Loan_test$loan_status, predicted = pred.rf[,'1'])
sprintf("Random Forest Test AUC: %.3f", auc.rf)
```

```
## [1] "Random Forest Test AUC: 0.999"
```

```
pr <- pred.rf[,'1']

#Converting prediction result to binary class
class.rf <- ifelse(pr > 0.5, 1, 0)
class.rf <- as.factor(class.rf)
```

```r
#Confusion matrix
confusionMatrix(data = class.rf, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 27550   273
##          1     0  5749
##
##                Accuracy : 0.9919
##                  95% CI : (0.9908, 0.9928)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9719
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.9547
##          Pos Pred Value : 0.9902
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8206
##          Detection Rate : 0.8206
##    Detection Prevalence : 0.8288
##       Balanced Accuracy : 0.9773
##
##        'Positive' Class : 0
##
```

Bagged Tree

```r
library(ipred)
# Train a bagged model
model.bagging <- bagging(formula = loan_status ~ .,
                         data = Loan_train,
                         coob = TRUE)

# Print the model
print(model.bagging)
```

```
##
## Bagging classification trees with 25 bootstrap replications
##
## Call: bagging.data.frame(formula = loan_status ~ ., data = Loan_train,
##     coob = TRUE)
##
## Out-of-bag estimate of misclassification error:  0.0044
```

```r
# Generate predicted classes using the model object
pred.bagging <- predict(object = model.bagging,
                            newdata = Loan_test,
                            type = "prob")
head(pred.bagging)
```

```
##         0    1
## [1,] 1.00 0.00
## [2,] 0.96 0.04
## [3,] 1.00 0.00
## [4,] 1.00 0.00
## [5,] 0.96 0.04
## [6,] 1.00 0.00
```

```
#Evaluate model performance
#Area under the ROC curve
auc.bagging <- auc(actual = Loan_test$loan_status, predicted = pred.bagging[,'1'])
sprintf("Bagged Trees Test AUC: %.3f", auc.bagging)
```

```
## [1] "Bagged Trees Test AUC: 0.998"
```

```
pr <- pred.bagging[,'1']

#Converting prediction result to binary class
class.bagging <- ifelse(pr > 0.5, 1, 0)
class.bagging <- as.factor(class.bagging)

#Confusion matrix
confusionMatrix(data = class.bagging, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 27539   118
##          1    11  5904
##
##                Accuracy : 0.9962
##                  95% CI : (0.9954, 0.9968)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9869
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9996
##             Specificity : 0.9804
##          Pos Pred Value : 0.9957
##          Neg Pred Value : 0.9981
##              Prevalence : 0.8206
##          Detection Rate : 0.8203
##    Detection Prevalence : 0.8238
##       Balanced Accuracy : 0.9900
##
##        'Positive' Class : 0
##
```

GBM

```
library(gbm)
```

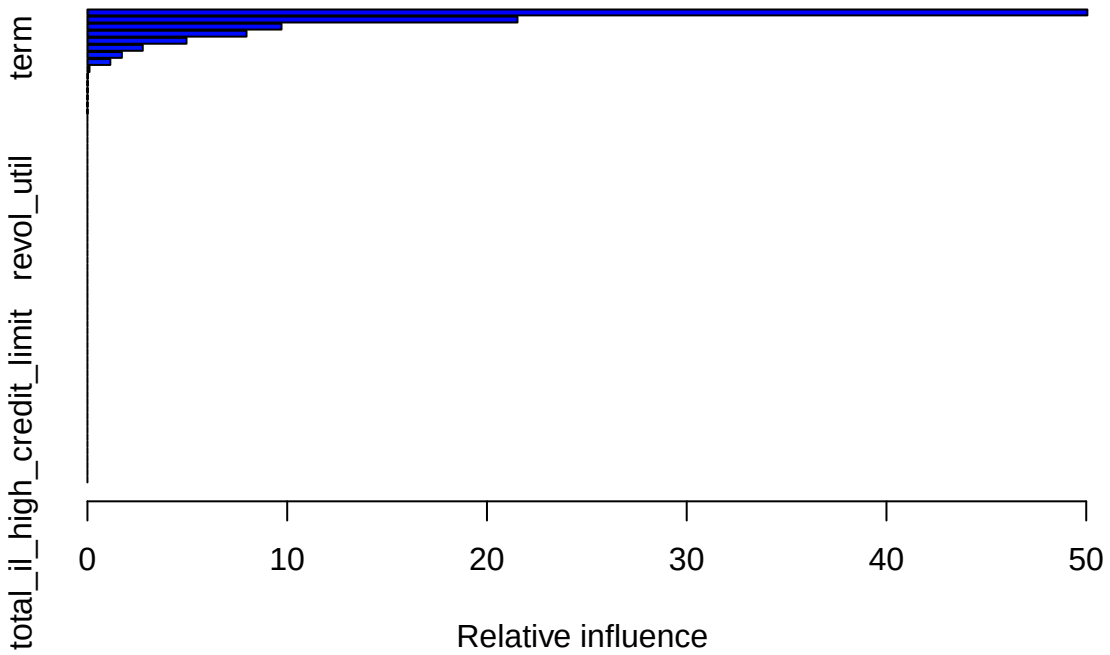
```
## Warning: package 'gbm' was built under R version 3.5.2
```

```
## Loaded gbm 2.1.5
```

```
model.gbm <- gbm(formula = as.character(loan_status) ~ .,
                 distribution = "bernoulli",
                 data = Loan_train,
                 n.trees = 1000)
print(model.gbm)
```

```
## gbm(formula = as.character(loan_status) ~ ., distribution = "bernoulli",
##     data = Loan_train, n.trees = 1000)
## A gradient boosted model with bernoulli loss function.
## 1000 iterations were performed.
## There were 67 predictors of which 15 had non-zero influence.
```

```
summary(model.gbm)
```




```
##                                       var      rel.inf
## total_rec_prncp                 total_rec_prncp 50.057189978
## last_pymnt_amnt                 last_pymnt_amnt 21.527419640
## funded_amnt_inv                 funded_amnt_inv  9.714474204
## installment                         installment  7.963692030
## loan_amnt                             loan_amnt  4.959255687
## term                                       term  2.768338587
## total_rec_late_fee           total_rec_late_fee  1.723066668
## total_rec_int                     total_rec_int  1.145918779
## total_pymnt                         total_pymnt  0.102735532
## tot_hi_cred_lim                 tot_hi_cred_lim  0.014912424
## acc_open_past_24mths       acc_open_past_24mths  0.009999982
## num_op_rev_tl                     num_op_rev_tl  0.006713400
## num_rev_tl_bal_gt_0         num_rev_tl_bal_gt_0  0.002627422
## num_actv_bc_tl                   num_actv_bc_tl  0.001989280
## total_pymnt_inv                 total_pymnt_inv  0.001666387
## funded_amnt                         funded_amnt  0.000000000
## int_rate                               int_rate  0.000000000
```

```
## grade                                         grade  0.000000000
## emp_length                                emp_length  0.000000000
## home_ownership                          home_ownership  0.000000000
## annual_inc                                 annual_inc  0.000000000
## verification_status               verification_status  0.000000000
## purpose                                       purpose  0.000000000
## dti                                               dti  0.000000000
## delinq_2yrs                               delinq_2yrs  0.000000000
## inq_last_6mths                         inq_last_6mths  0.000000000
## open_acc                                     open_acc  0.000000000
## pub_rec                                       pub_rec  0.000000000
## revol_bal                                  revol_bal  0.000000000
## revol_util                                 revol_util  0.000000000
## total_acc                                   total_acc  0.000000000
## initial_list_status               initial_list_status  0.000000000
## collections_12_mths_ex_med collections_12_mths_ex_med  0.000000000
## acc_now_delinq                         acc_now_delinq  0.000000000
## tot_coll_amt                             tot_coll_amt  0.000000000
## tot_cur_bal                               tot_cur_bal  0.000000000
## total_rev_hi_lim                     total_rev_hi_lim  0.000000000
## avg_cur_bal                               avg_cur_bal  0.000000000
## bc_open_to_buy                         bc_open_to_buy  0.000000000
## bc_util                                       bc_util  0.000000000
## chargeoff_within_12_mths     chargeoff_within_12_mths  0.000000000
## delinq_amnt                               delinq_amnt  0.000000000
## mo_sin_old_il_acct                 mo_sin_old_il_acct  0.000000000
## mo_sin_old_rev_tl_op             mo_sin_old_rev_tl_op  0.000000000
## mo_sin_rcnt_rev_tl_op           mo_sin_rcnt_rev_tl_op  0.000000000
## mo_sin_rcnt_tl                         mo_sin_rcnt_tl  0.000000000
## mort_acc                                     mort_acc  0.000000000
## mths_since_recent_bc             mths_since_recent_bc  0.000000000
## mths_since_recent_inq           mths_since_recent_inq  0.000000000
## num_accts_ever_120_pd           num_accts_ever_120_pd  0.000000000
## num_actv_rev_tl                       num_actv_rev_tl  0.000000000
## num_bc_sats                               num_bc_sats  0.000000000
## num_bc_tl                                   num_bc_tl  0.000000000
## num_il_tl                                   num_il_tl  0.000000000
## num_rev_accts                           num_rev_accts  0.000000000
## num_sats                                     num_sats  0.000000000
## num_tl_120dpd_2m                     num_tl_120dpd_2m  0.000000000
## num_tl_30dpd                             num_tl_30dpd  0.000000000
## num_tl_90g_dpd_24m                 num_tl_90g_dpd_24m  0.000000000
## num_tl_op_past_12m                 num_tl_op_past_12m  0.000000000
## pct_tl_nvr_dlq                         pct_tl_nvr_dlq  0.000000000
## percent_bc_gt_75                     percent_bc_gt_75  0.000000000
## pub_rec_bankruptcies             pub_rec_bankruptcies  0.000000000
## tax_liens                                   tax_liens  0.000000000
## total_bal_ex_mort                   total_bal_ex_mort  0.000000000
## total_bc_limit                         total_bc_limit  0.000000000
## total_il_high_credit_limit total_il_high_credit_limit  0.000000000
```

```
#Prediction
pred.gbm <- predict(object = model.gbm,
                    newdata = Loan_test,
```

```
                  n.trees = 1000,
                  type = "response")
head(pred.gbm)
```

```
## [1] 0.001648028 0.025335306 0.001849512 0.002103967 0.046754502 0.001552185
```

```
#Evaluate model performance
#Area under the ROC curve
auc.gbm <- auc(actual = Loan_test$loan_status, predicted = pred.gbm)
sprintf("GBM Test AUC: %.3f", auc.gbm)
```

```
## [1] "GBM Test AUC: 0.999"
```

```
#Converting prediction result to binary class
class.gbm <- ifelse(pred.gbm > 0.5, 1, 0)
class.gbm <- as.factor(class.gbm)

#Confusion matrix
confusionMatrix(data = class.gbm, reference = Loan_test$loan_status)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 27550   421
##          1     0  5601
##
##                Accuracy : 0.9875
##                  95% CI : (0.9862, 0.9886)
##     No Information Rate : 0.8206
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9562
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.9301
##          Pos Pred Value : 0.9849
##          Neg Pred Value : 1.0000
##              Prevalence : 0.8206
##          Detection Rate : 0.8206
##    Detection Prevalence : 0.8332
##       Balanced Accuracy : 0.9650
##
##        'Positive' Class : 0
##
```

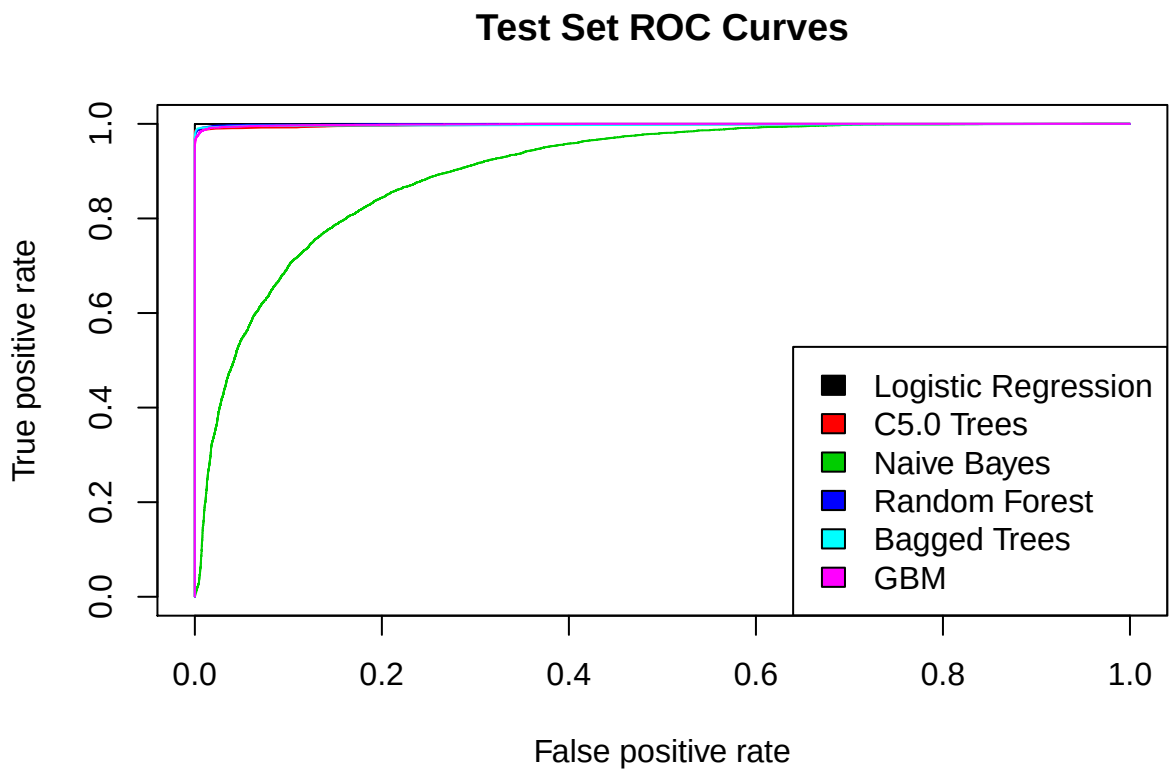
Step 4: Comparing model performances

Compute the accuracy and, if appropriate, the area under the ROC curve (AUC) to rank the classification accuracy of each model.

```
# List of predictions
preds_list <- list(pred.logit, pred.c50[,2], pred.nb[,2], pred.rf[,2], pred.bagging[,2], pred.gbm)
```

```r
# List of actual values (same for all)
m <- length(preds_list)
actuals_list <- rep(list(Loan_test$loan_status), m)

# Plot the ROC curves
pred <- prediction(preds_list, actuals_list)
rocs <- performance(pred, "tpr", "fpr")
plot(rocs, col = as.list(1:m), main = "Test Set ROC Curves")
abline=c(0,1)
legend(x = "bottomright",
       legend = c("Logistic Regression","C5.0 Trees", "Naive Bayes","Random Forest","Bagged Trees", "GBM
       fill = 1:m)
```




```r
sprintf("Logistic Regression Test AUC: %.3f", auc.logit)
```

```
## [1] "Logistic Regression Test AUC: 1.000"
```

```r
sprintf("kNN Test AUC: %.3f", auc.knn)
```

```
## [1] "kNN Test AUC: 0.234"
```

```r
sprintf("C5.0 Tree Test AUC: %.3f", auc.c50)
```

```
## [1] "C5.0 Tree Test AUC: 0.998"
```

```r
sprintf("Naive Bayes Test AUC: %.3f", auc.nb)
```

```
## [1] "Naive Bayes Test AUC: 0.904"
```

```r
sprintf("Random Forest Test AUC: %.3f", auc.rf)
```

```
## [1] "Random Forest Test AUC: 0.999"
```

```r
sprintf("Bagged Trees Test AUC: %.3f", auc.bagging)
```

```
## [1] "Bagged Trees Test AUC: 0.998"
```

```r
sprintf("GBM Test AUC: %.3f", auc.gbm)
```

```
## [1] "GBM Test AUC: 0.999"
```