Lending Club Loan History Challenge

Aaron Banlao

```
library(pacman)
p_load(ggplot2, dplyr, tidyverse, janitor, GGally, tidymodels, yardstick, kknn, ranger, klaR, discrim, i
```

Building the Model

Reading in the data

```
lend <- read.csv("lending_club_data_2012_2014_small.csv")
head(lend)
```

```
##         id member_id loan_amnt funded_amnt funded_amnt_inv      term int_rate
## 1  3290675        NA     12175       12175           12175 36 months    17.77
## 2  1690712        NA     15000       15000           14975 36 months    18.49
## 3  1339491        NA     15000       15000           15000 36 months    20.50
## 4 13518760        NA     15000       15000           15000 36 months     8.39
## 5  6578305        NA     10000       10000           10000 36 months     6.62
## 6 16441609        NA      4000        4000            4000 36 months    13.98
##   installment grade sub_grade                    emp_title emp_length
## 1      438.76     D        D1                      tlc-elc    7 years
## 2      545.99     D        D2       HP Enterprise Services    3 years
## 3      561.29     E        E2                         <NA>     1 year
## 4      472.75     A        A5             Project Manager  10+ years
## 5      307.04     A        A2 OBGYN CONSULTANTS OF MEMPHIS    4 years
## 6      136.68     C        C3                         <NA>       <NA>
##   home_ownership annual_inc verification_status  issue_d loan_status pymnt_plan
## 1           RENT    35000.0        Not Verified Feb-2013 Charged Off          n
## 2           RENT    71614.1        Not Verified Nov-2012  Fully Paid          n
## 3           RENT    35000.0            Verified Jun-2012  Fully Paid          n
## 4       MORTGAGE    75000.0     Source Verified Jul-2014  Fully Paid          n
## 5       MORTGAGE    29000.0        Not Verified Aug-2013 Charged Off          n
## 6       MORTGAGE    70000.0            Verified May-2014  Fully Paid          n
##                                                                  url
## 1  https://lendingclub.com/browse/loanDetail.action?loan_id=3290675
## 2  https://lendingclub.com/browse/loanDetail.action?loan_id=1690712
## 3  https://lendingclub.com/browse/loanDetail.action?loan_id=1339491
## 4 https://lendingclub.com/browse/loanDetail.action?loan_id=13518760
## 5  https://lendingclub.com/browse/loanDetail.action?loan_id=6578305
## 6 https://lendingclub.com/browse/loanDetail.action?loan_id=16441609
##
## 1
## 2                              Borrower added on 10/31/12 > I want to pay-off my Credit Card debts
```

```
## 3
## 4
## 5 Borrower added on 07/25/13 > THIS LOAN WILL BE USED TO FIX THINGS IN THE HOME AND INSTALL HARD WOOD
## 6
##              purpose               title zip_code addr_state   dti delinq_2yrs
## 1 debt_consolidation        Credit Cards    554xx         MN  8.43           0
## 2        credit_card Credit Card Pay-Off    406xx         KY 13.68           0
## 3                car                 Car    152xx         PA 34.56           0
## 4 debt_consolidation  Debt consolidation    957xx         CA 29.01           0
## 5   home_improvement         improvement    380xx         TN 24.00           0
## 6   home_improvement    Home improvement    063xx         CT 26.04           1
##   earliest_cr_line fico_range_low fico_range_high inq_last_6mths
## 1         May-2002            695             699              3
## 2         Sep-2007            670             674              1
## 3         Jul-2005            665             669              0
## 4         Mar-1997            685             689              0
## 5         Aug-1998            715             719              0
## 6         Feb-1992            675             679              0
##   mths_since_last_delinq mths_since_last_record open_acc pub_rec revol_bal
## 1                     NA                     NA       10       0      9706
## 2                     NA                     NA       12       0     15991
## 3                     NA                     NA        9       0     22081
## 4                     76                     NA       19       0     20594
## 5                     38                     NA       18       0      9801
## 6                      9                    110       14       1     19523
##   revol_util total_acc initial_list_status out_prncp out_prncp_inv total_pymnt
## 1       72.4        21                   w         0             0     8453.88
## 2       68.9        14                   f         0             0    19507.90
## 3       95.2         9                   f         0             0    20231.07
## 4       56.9        54                   w         0             0    15830.20
## 5       26.0        35                   w         0             0     8405.16
## 6       80.0        31                   f         0             0     4920.11
##   total_pymnt_inv total_rec_prncp total_rec_int total_rec_late_fee recoveries
## 1         8453.88         3675.46       2028.42                 0    2750.00
## 2        19475.39        15000.00       4507.90                 0       0.00
## 3        20231.07        15000.00       5231.07                 0       0.00
## 4        15830.20        15000.00        830.20                 0       0.00
## 5         8405.16         7020.87        962.17                 0     422.12
## 6         4920.11         4000.00        920.11                 0       0.00
##   collection_recovery_fee last_pymnt_d last_pymnt_amnt next_pymnt_d
## 1                467.5000     Mar-2014          438.76         <NA>
## 2                  0.0000     May-2015         3694.26         <NA>
## 3                  0.0000     Jul-2015           30.91         <NA>
## 4                  0.0000     Apr-2015           22.33         <NA>
## 5                 75.9816     Oct-2015          307.04         <NA>
## 6                  0.0000     May-2017          136.31         <NA>
##   last_credit_pull_d last_fico_range_high last_fico_range_low
## 1          Aug-2018                  564                 560
## 2          Mar-2019                  789                 785
## 3          Mar-2019                  729                 725
## 4          Jan-2019                  684                 680
## 5          Oct-2016                  539                 535
## 6          Mar-2019                  674                 670
##   collections_12_mths_ex_med mths_since_last_major_derog policy_code
```

```
## 1                         0                         NA           1
## 2                         0                         NA           1
## 3                         0                         NA           1
## 4                         0                         76           1
## 5                         0                         NA           1
## 6                         0                         NA           1
##   application_type annual_inc_joint dti_joint verification_status_joint
## 1       Individual               NA        NA                       NA
## 2       Individual               NA        NA                       NA
## 3       Individual               NA        NA                       NA
## 4       Individual               NA        NA                       NA
## 5       Individual               NA        NA                       NA
## 6       Individual               NA        NA                       NA
##   acc_now_delinq tot_coll_amt tot_cur_bal open_acc_6m open_act_il open_il_12m
## 1              0            0       44692          NA          NA          NA
## 2              0            0       19718          NA          NA          NA
## 3              0           NA          NA          NA          NA          NA
## 4              0            0      243234          NA          NA          NA
## 5              0            0       93725          NA          NA          NA
## 6              0            0       78233          NA          NA          NA
##   open_il_24m mths_since_rcnt_il total_bal_il il_util open_rv_12m open_rv_24m
## 1          NA                NA           NA      NA          NA          NA
## 2          NA                NA           NA      NA          NA          NA
## 3          NA                NA           NA      NA          NA          NA
## 4          NA                NA           NA      NA          NA          NA
## 5          NA                NA           NA      NA          NA          NA
## 6          NA                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            13400     NA          NA           NA
## 2         NA       NA            23200     NA          NA           NA
## 3         NA       NA               NA     NA          NA           NA
## 4         NA       NA            36200     NA          NA           NA
## 5         NA       NA            37700     NA          NA           NA
## 6         NA       NA            24400     NA          NA           NA
##   acc_open_past_24mths avg_cur_bal bc_open_to_buy bc_util
## 1                    3        4469            177    97.9
## 2                    2        1643           1278    87.3
## 3                    6          NA              0   100.6
## 4                    7       12802           6148    69.7
## 5                    6        5513          21965    30.0
## 6                    5        6018           1866    77.2
##   chargeoff_within_12_mths delinq_amnt mo_sin_old_il_acct mo_sin_old_rev_tl_op
## 1                        0          0                 38                  130
## 2                        0          0                 59                   62
## 3                        0          0                 NA                   NA
## 4                        0          0                181                  207
## 5                        0          0                147                  182
## 6                        0          0                151                  107
##   mo_sin_rcnt_rev_tl_op mo_sin_rcnt_tl mort_acc mths_since_recent_bc
## 1                    13             13        1                   13
## 2                    11             11        0                   11
## 3                    NA             NA        0                    3
## 4                     8              8        2                    8
## 5                    11             11        3                   11
```

```
## 6                    11             10        5                    11
##   mths_since_recent_bc_dlq mths_since_recent_inq mths_since_recent_revol_delinq
## 1                       NA                     1                              NA
## 2                       NA                     0                              NA
## 3                       NA                    NA                              NA
## 4                       NA                     8                              78
## 5                       NA                    11                              38
## 6                       NA                    17                               9
##   num_accts_ever_120_pd num_actv_bc_tl num_actv_rev_tl num_bc_sats num_bc_tl
## 1                     0              3               6           3         7
## 2                     0              6              10           6         7
## 3                    NA             NA              NA           3        NA
## 4                     1              7              12           7        16
## 5                     0              5               7           7        13
## 6                     0              6               8           6        12
##   num_il_tl num_op_rev_tl num_rev_accts num_rev_tl_bal_gt_0 num_sats
## 1         2             8            18                   6       10
## 2         2            11            12                  10       12
## 3        NA            NA            NA                  NA        8
## 4        24            14            27                  12       19
## 5        10            13            22                   7       18
## 6         7             9            19                   8       13
##   num_tl_120dpd_2m num_tl_30dpd num_tl_90g_dpd_24m num_tl_op_past_12m
## 1                0            0                  0                  0
## 2                0            0                  0                  1
## 3               NA           NA                 NA                 NA
## 4                0            0                  0                  2
## 5                0            0                  0                  3
## 6                0            0                  0                  2
##   pct_tl_nvr_dlq percent_bc_gt_75 pub_rec_bankruptcies tax_liens
## 1          100.0            100.0                    0         0
## 2          100.0             83.3                    0         0
## 3             NA            100.0                    0         0
## 4           94.4             57.1                    0         0
## 5           94.0              0.0                    0         0
## 6           96.8             33.3                    1         0
##   tot_hi_cred_lim total_bal_ex_mort total_bc_limit total_il_high_credit_limit
## 1           49589             44692          8300                       36189
## 2           40020             19718         10100                       16820
## 3              NA             60075         17200                          NA
## 4          276413             82065         20300                       76313
## 5          133394             27109         31400                       21794
## 6          105110             36490          8200                       33808
##   revol_bal_joint sec_app_fico_range_low sec_app_fico_range_high
## 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_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 sec_app_num_rev_accts
## 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
##   sec_app_chargeoff_within_12_mths sec_app_collections_12_mths_ex_med
## 1                               NA                                NA
## 2                               NA                                NA
## 3                               NA                                NA
## 4                               NA                                NA
## 5                               NA                                NA
## 6                               NA                                NA
##   sec_app_mths_since_last_major_derog hardship_flag hardship_type
## 1                                  NA             N          <NA>
## 2                                  NA             N          <NA>
## 3                                  NA             N          <NA>
## 4                                  NA             N          <NA>
## 5                                  NA             N          <NA>
## 6                                  NA             N          <NA>
##   hardship_reason hardship_status deferral_term hardship_amount
## 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
##   hardship_start_date hardship_end_date payment_plan_start_date hardship_length
## 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
##   hardship_dpd hardship_loan_status orig_projected_additional_accrued_interest
## 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_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                    Y                 Feb-2015
```

```
## 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 settlement_percentage
## 1          COMPLETE        Aug-2014              2750                 30.12
## 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
##   settlement_term year
## 1               0 2013
## 2              NA 2012
## 3              NA 2012
## 4              NA 2014
## 5              NA 2013
## 6              NA 2014
```

```
dim(lend)
```

```
## [1] 10000   152
```

```
colnames(lend)
```

```
##   [1] "id"
##   [2] "member_id"
##   [3] "loan_amnt"
##   [4] "funded_amnt"
##   [5] "funded_amnt_inv"
##   [6] "term"
##   [7] "int_rate"
##   [8] "installment"
##   [9] "grade"
##  [10] "sub_grade"
##  [11] "emp_title"
##  [12] "emp_length"
##  [13] "home_ownership"
##  [14] "annual_inc"
##  [15] "verification_status"
##  [16] "issue_d"
##  [17] "loan_status"
##  [18] "pymnt_plan"
##  [19] "url"
##  [20] "desc"
##  [21] "purpose"
##  [22] "title"
##  [23] "zip_code"
##  [24] "addr_state"
##  [25] "dti"
##  [26] "delinq_2yrs"
##  [27] "earliest_cr_line"
```

```
##  [28] "fico_range_low"
##  [29] "fico_range_high"
##  [30] "inq_last_6mths"
##  [31] "mths_since_last_delinq"
##  [32] "mths_since_last_record"
##  [33] "open_acc"
##  [34] "pub_rec"
##  [35] "revol_bal"
##  [36] "revol_util"
##  [37] "total_acc"
##  [38] "initial_list_status"
##  [39] "out_prncp"
##  [40] "out_prncp_inv"
##  [41] "total_pymnt"
##  [42] "total_pymnt_inv"
##  [43] "total_rec_prncp"
##  [44] "total_rec_int"
##  [45] "total_rec_late_fee"
##  [46] "recoveries"
##  [47] "collection_recovery_fee"
##  [48] "last_pymnt_d"
##  [49] "last_pymnt_amnt"
##  [50] "next_pymnt_d"
##  [51] "last_credit_pull_d"
##  [52] "last_fico_range_high"
##  [53] "last_fico_range_low"
##  [54] "collections_12_mths_ex_med"
##  [55] "mths_since_last_major_derog"
##  [56] "policy_code"
##  [57] "application_type"
##  [58] "annual_inc_joint"
##  [59] "dti_joint"
##  [60] "verification_status_joint"
##  [61] "acc_now_delinq"
##  [62] "tot_coll_amt"
##  [63] "tot_cur_bal"
##  [64] "open_acc_6m"
##  [65] "open_act_il"
##  [66] "open_il_12m"
##  [67] "open_il_24m"
##  [68] "mths_since_rcnt_il"
##  [69] "total_bal_il"
##  [70] "il_util"
##  [71] "open_rv_12m"
##  [72] "open_rv_24m"
##  [73] "max_bal_bc"
##  [74] "all_util"
##  [75] "total_rev_hi_lim"
##  [76] "inq_fi"
##  [77] "total_cu_tl"
##  [78] "inq_last_12m"
##  [79] "acc_open_past_24mths"
##  [80] "avg_cur_bal"
##  [81] "bc_open_to_buy"
```

```
##  [82] "bc_util"
##  [83] "chargeoff_within_12_mths"
##  [84] "delinq_amnt"
##  [85] "mo_sin_old_il_acct"
##  [86] "mo_sin_old_rev_tl_op"
##  [87] "mo_sin_rcnt_rev_tl_op"
##  [88] "mo_sin_rcnt_tl"
##  [89] "mort_acc"
##  [90] "mths_since_recent_bc"
##  [91] "mths_since_recent_bc_dlq"
##  [92] "mths_since_recent_inq"
##  [93] "mths_since_recent_revol_delinq"
##  [94] "num_accts_ever_120_pd"
##  [95] "num_actv_bc_tl"
##  [96] "num_actv_rev_tl"
##  [97] "num_bc_sats"
##  [98] "num_bc_tl"
##  [99] "num_il_tl"
## [100] "num_op_rev_tl"
## [101] "num_rev_accts"
## [102] "num_rev_tl_bal_gt_0"
## [103] "num_sats"
## [104] "num_tl_120dpd_2m"
## [105] "num_tl_30dpd"
## [106] "num_tl_90g_dpd_24m"
## [107] "num_tl_op_past_12m"
## [108] "pct_tl_nvr_dlq"
## [109] "percent_bc_gt_75"
## [110] "pub_rec_bankruptcies"
## [111] "tax_liens"
## [112] "tot_hi_cred_lim"
## [113] "total_bal_ex_mort"
## [114] "total_bc_limit"
## [115] "total_il_high_credit_limit"
## [116] "revol_bal_joint"
## [117] "sec_app_fico_range_low"
## [118] "sec_app_fico_range_high"
## [119] "sec_app_earliest_cr_line"
## [120] "sec_app_inq_last_6mths"
## [121] "sec_app_mort_acc"
## [122] "sec_app_open_acc"
## [123] "sec_app_revol_util"
## [124] "sec_app_open_act_il"
## [125] "sec_app_num_rev_accts"
## [126] "sec_app_chargeoff_within_12_mths"
## [127] "sec_app_collections_12_mths_ex_med"
## [128] "sec_app_mths_since_last_major_derog"
## [129] "hardship_flag"
## [130] "hardship_type"
## [131] "hardship_reason"
## [132] "hardship_status"
## [133] "deferral_term"
## [134] "hardship_amount"
## [135] "hardship_start_date"
```

```
## [136] "hardship_end_date"
## [137] "payment_plan_start_date"
## [138] "hardship_length"
## [139] "hardship_dpd"
## [140] "hardship_loan_status"
## [141] "orig_projected_additional_accrued_interest"
## [142] "hardship_payoff_balance_amount"
## [143] "hardship_last_payment_amount"
## [144] "disbursement_method"
## [145] "debt_settlement_flag"
## [146] "debt_settlement_flag_date"
## [147] "settlement_status"
## [148] "settlement_date"
## [149] "settlement_amount"
## [150] "settlement_percentage"
## [151] "settlement_term"
## [152] "year"
```

Retrieving duplicates in the dataset

```
get_dupes(lend)
```

```
## No variable names specified - using all columns.
```

```
## No duplicate combinations found of: id, member_id, loan_amnt, funded_amnt, funded_amnt_inv, term, in
```

```
##   [1] id
##   [2] member_id
##   [3] loan_amnt
##   [4] funded_amnt
##   [5] funded_amnt_inv
##   [6] term
##   [7] int_rate
##   [8] installment
##   [9] grade
##  [10] sub_grade
##  [11] emp_title
##  [12] emp_length
##  [13] home_ownership
##  [14] annual_inc
##  [15] verification_status
##  [16] issue_d
##  [17] loan_status
##  [18] pymnt_plan
##  [19] url
##  [20] desc
##  [21] purpose
##  [22] title
##  [23] zip_code
##  [24] addr_state
##  [25] dti
```

```
##  [26] delinq_2yrs
##  [27] earliest_cr_line
##  [28] fico_range_low
##  [29] fico_range_high
##  [30] inq_last_6mths
##  [31] mths_since_last_delinq
##  [32] mths_since_last_record
##  [33] open_acc
##  [34] pub_rec
##  [35] revol_bal
##  [36] revol_util
##  [37] total_acc
##  [38] initial_list_status
##  [39] out_prncp
##  [40] out_prncp_inv
##  [41] total_pymnt
##  [42] total_pymnt_inv
##  [43] total_rec_prncp
##  [44] total_rec_int
##  [45] total_rec_late_fee
##  [46] recoveries
##  [47] collection_recovery_fee
##  [48] last_pymnt_d
##  [49] last_pymnt_amnt
##  [50] next_pymnt_d
##  [51] last_credit_pull_d
##  [52] last_fico_range_high
##  [53] last_fico_range_low
##  [54] collections_12_mths_ex_med
##  [55] mths_since_last_major_derog
##  [56] policy_code
##  [57] application_type
##  [58] annual_inc_joint
##  [59] dti_joint
##  [60] verification_status_joint
##  [61] acc_now_delinq
##  [62] tot_coll_amt
##  [63] tot_cur_bal
##  [64] open_acc_6m
##  [65] open_act_il
##  [66] open_il_12m
##  [67] open_il_24m
##  [68] mths_since_rcnt_il
##  [69] total_bal_il
##  [70] il_util
##  [71] open_rv_12m
##  [72] open_rv_24m
##  [73] max_bal_bc
##  [74] all_util
##  [75] total_rev_hi_lim
##  [76] inq_fi
##  [77] total_cu_tl
##  [78] inq_last_12m
##  [79] acc_open_past_24mths
```

```
##  [80] avg_cur_bal
##  [81] bc_open_to_buy
##  [82] bc_util
##  [83] chargeoff_within_12_mths
##  [84] delinq_amnt
##  [85] mo_sin_old_il_acct
##  [86] mo_sin_old_rev_tl_op
##  [87] mo_sin_rcnt_rev_tl_op
##  [88] mo_sin_rcnt_tl
##  [89] mort_acc
##  [90] mths_since_recent_bc
##  [91] mths_since_recent_bc_dlq
##  [92] mths_since_recent_inq
##  [93] mths_since_recent_revol_delinq
##  [94] num_accts_ever_120_pd
##  [95] num_actv_bc_tl
##  [96] num_actv_rev_tl
##  [97] num_bc_sats
##  [98] num_bc_tl
##  [99] num_il_tl
## [100] num_op_rev_tl
## [101] num_rev_accts
## [102] num_rev_tl_bal_gt_0
## [103] num_sats
## [104] num_tl_120dpd_2m
## [105] num_tl_30dpd
## [106] num_tl_90g_dpd_24m
## [107] num_tl_op_past_12m
## [108] pct_tl_nvr_dlq
## [109] percent_bc_gt_75
## [110] pub_rec_bankruptcies
## [111] tax_liens
## [112] tot_hi_cred_lim
## [113] total_bal_ex_mort
## [114] total_bc_limit
## [115] total_il_high_credit_limit
## [116] revol_bal_joint
## [117] sec_app_fico_range_low
## [118] sec_app_fico_range_high
## [119] sec_app_earliest_cr_line
## [120] sec_app_inq_last_6mths
## [121] sec_app_mort_acc
## [122] sec_app_open_acc
## [123] sec_app_revol_util
## [124] sec_app_open_act_il
## [125] sec_app_num_rev_accts
## [126] sec_app_chargeoff_within_12_mths
## [127] sec_app_collections_12_mths_ex_med
## [128] sec_app_mths_since_last_major_derog
## [129] hardship_flag
## [130] hardship_type
## [131] hardship_reason
## [132] hardship_status
## [133] deferral_term
```

```
## [134] hardship_amount
## [135] hardship_start_date
## [136] hardship_end_date
## [137] payment_plan_start_date
## [138] hardship_length
## [139] hardship_dpd
## [140] hardship_loan_status
## [141] orig_projected_additional_accrued_interest
## [142] hardship_payoff_balance_amount
## [143] hardship_last_payment_amount
## [144] disbursement_method
## [145] debt_settlement_flag
## [146] debt_settlement_flag_date
## [147] settlement_status
## [148] settlement_date
## [149] settlement_amount
## [150] settlement_percentage
## [151] settlement_term
## [152] year
## [153] dupe_count
## <0 rows> (or 0-length row.names)
```

Finding the number and percentage of nulls in columns

```
apply(lend, 2, function(x)sum(is.na(x)))
```

```
##                                     id
##                                      0
##                                member_id
##                                  10000
##                                loan_amnt
##                                      0
##                              funded_amnt
##                                      0
##                            funded_amnt_inv
##                                      0
##                                   term
##                                      0
##                                int_rate
##                                      0
##                             installment
##                                      0
##                                  grade
##                                      0
##                               sub_grade
##                                      0
##                               emp_title
##                                    609
##                              emp_length
##                                    470
##                            home_ownership
##                                      0
```

```
##                          annual_inc
##                                   0
##                 verification_status
##                                   0
##                             issue_d
##                                   0
##                         loan_status
##                                   0
##                          pymnt_plan
##                                   0
##                                 url
##                                   0
##                                desc
##                                7724
##                             purpose
##                                   0
##                               title
##                                   0
##                            zip_code
##                                   0
##                          addr_state
##                                   0
##                                 dti
##                                   0
##                           delinq_2yrs
##                                   0
##                      earliest_cr_line
##                                   0
##                        fico_range_low
##                                   0
##                       fico_range_high
##                                   0
##                        inq_last_6mths
##                                   0
##                  mths_since_last_delinq
##                                5293
##                  mths_since_last_record
##                                8611
##                              open_acc
##                                   0
##                               pub_rec
##                                   0
##                             revol_bal
##                                   0
##                            revol_util
##                                   5
##                             total_acc
##                                   0
##                     initial_list_status
##                                   0
##                             out_prncp
##                                   0
##                          out_prncp_inv
##                                   0
```

```
##                             total_pymnt
##                                       0
##                         total_pymnt_inv
##                                       0
##                         total_rec_prncp
##                                       0
##                           total_rec_int
##                                       0
##                       total_rec_late_fee
##                                       0
##                              recoveries
##                                       0
##                     collection_recovery_fee
##                                       0
##                               last_pymnt_d
##                                       6
##                         last_pymnt_amnt
##                                       0
##                               next_pymnt_d
##                                    9716
##                        last_credit_pull_d
##                                       1
##                      last_fico_range_high
##                                       0
##                       last_fico_range_low
##                                       0
##                  collections_12_mths_ex_med
##                                       0
##                 mths_since_last_major_derog
##                                    7685
##                             policy_code
##                                       0
##                          application_type
##                                       0
##                          annual_inc_joint
##                                   10000
##                                 dti_joint
##                                   10000
##                   verification_status_joint
##                                   10000
##                            acc_now_delinq
##                                       0
##                           tot_coll_amt
##                                     638
##                             tot_cur_bal
##                                     638
##                             open_acc_6m
##                                   10000
##                             open_act_il
##                                   10000
##                             open_il_12m
##                                   10000
##                             open_il_24m
##                                   10000
```

```
##                        mths_since_rcnt_il
##                                     10000
##                              total_bal_il
##                                     10000
##                                   il_util
##                                     10000
##                                open_rv_12m
##                                     10000
##                                open_rv_24m
##                                     10000
##                                 max_bal_bc
##                                     10000
##                                   all_util
##                                     10000
##                          total_rev_hi_lim
##                                       638
##                                    inq_fi
##                                     10000
##                                total_cu_tl
##                                     10000
##                               inq_last_12m
##                                     10000
##                       acc_open_past_24mths
##                                       152
##                                avg_cur_bal
##                                       639
##                             bc_open_to_buy
##                                       245
##                                  bc_util
##                                       252
##                    chargeoff_within_12_mths
##                                         0
##                                delinq_amnt
##                                         0
##                          mo_sin_old_il_acct
##                                       954
##                         mo_sin_old_rev_tl_op
##                                       638
##                        mo_sin_rcnt_rev_tl_op
##                                       638
##                               mo_sin_rcnt_tl
##                                       638
##                                  mort_acc
##                                       152
##                         mths_since_recent_bc
##                                       236
##                    mths_since_recent_bc_dlq
##                                      7576
##                        mths_since_recent_inq
##                                      1150
##               mths_since_recent_revol_delinq
##                                      6664
##                        num_accts_ever_120_pd
##                                       638
```

```
##                          num_actv_bc_tl
##                                     638
##                        num_actv_rev_tl
##                                     638
##                           num_bc_sats
##                                     343
##                             num_bc_tl
##                                     638
##                             num_il_tl
##                                     638
##                          num_op_rev_tl
##                                     638
##                          num_rev_accts
##                                     638
##                       num_rev_tl_bal_gt_0
##                                     638
##                                num_sats
##                                     343
##                       num_tl_120dpd_2m
##                                     837
##                           num_tl_30dpd
##                                     638
##                        num_tl_90g_dpd_24m
##                                     638
##                        num_tl_op_past_12m
##                                     638
##                           pct_tl_nvr_dlq
##                                     645
##                         percent_bc_gt_75
##                                     247
##                      pub_rec_bankruptcies
##                                       0
##                                tax_liens
##                                       0
##                           tot_hi_cred_lim
##                                     638
##                         total_bal_ex_mort
##                                     152
##                            total_bc_limit
##                                     152
##                  total_il_high_credit_limit
##                                     638
##                               revol_bal_joint
##                                   10000
##                      sec_app_fico_range_low
##                                   10000
##                     sec_app_fico_range_high
##                                   10000
##                    sec_app_earliest_cr_line
##                                   10000
##                     sec_app_inq_last_6mths
##                                   10000
##                           sec_app_mort_acc
##                                   10000
```

```
##                           sec_app_open_acc
##                                      10000
##                          sec_app_revol_util
##                                      10000
##                        sec_app_open_act_il
##                                      10000
##                       sec_app_num_rev_accts
##                                      10000
##            sec_app_chargeoff_within_12_mths
##                                      10000
##          sec_app_collections_12_mths_ex_med
##                                      10000
##         sec_app_mths_since_last_major_derog
##                                      10000
##                                hardship_flag
##                                          0
##                                hardship_type
##                                       9989
##                              hardship_reason
##                                       9989
##                              hardship_status
##                                       9989
##                                deferral_term
##                                       9989
##                              hardship_amount
##                                       9989
##                           hardship_start_date
##                                       9989
##                             hardship_end_date
##                                       9989
##                      payment_plan_start_date
##                                       9989
##                                hardship_length
##                                       9989
##                                  hardship_dpd
##                                       9989
##                         hardship_loan_status
##                                       9989
## orig_projected_additional_accrued_interest
##                                       9990
##              hardship_payoff_balance_amount
##                                       9989
##                hardship_last_payment_amount
##                                       9989
##                           disbursement_method
##                                          0
##                         debt_settlement_flag
##                                          0
##                    debt_settlement_flag_date
##                                       9838
##                            settlement_status
##                                       9838
##                              settlement_date
##                                       9838
```

```
##                         settlement_amount
##                                      9838
##                       settlement_percentage
##                                      9838
##                             settlement_term
##                                      9838
##                                      year
##                                         0
```

```
apply(lend, 2, function(x)sum(is.na(x))/length(x))
```

```
##                                       id
##                                   0.0000
##                                member_id
##                                   1.0000
##                                loan_amnt
##                                   0.0000
##                              funded_amnt
##                                   0.0000
##                            funded_amnt_inv
##                                   0.0000
##                                     term
##                                   0.0000
##                                 int_rate
##                                   0.0000
##                              installment
##                                   0.0000
##                                    grade
##                                   0.0000
##                                sub_grade
##                                   0.0000
##                                emp_title
##                                   0.0609
##                               emp_length
##                                   0.0470
##                           home_ownership
##                                   0.0000
##                               annual_inc
##                                   0.0000
##                        verification_status
##                                   0.0000
##                                  issue_d
##                                   0.0000
##                              loan_status
##                                   0.0000
##                               pymnt_plan
##                                   0.0000
##                                      url
##                                   0.0000
##                                     desc
##                                   0.7724
##                                  purpose
##                                   0.0000
##                                    title
```

```
##                                 0.0000
##                               zip_code
##                                 0.0000
##                             addr_state
##                                 0.0000
##                                    dti
##                                 0.0000
##                             delinq_2yrs
##                                 0.0000
##                         earliest_cr_line
##                                 0.0000
##                           fico_range_low
##                                 0.0000
##                          fico_range_high
##                                 0.0000
##                           inq_last_6mths
##                                 0.0000
##                       mths_since_last_delinq
##                                 0.5293
##                       mths_since_last_record
##                                 0.8611
##                               open_acc
##                                 0.0000
##                                pub_rec
##                                 0.0000
##                              revol_bal
##                                 0.0000
##                             revol_util
##                                 0.0005
##                              total_acc
##                                 0.0000
##                         initial_list_status
##                                 0.0000
##                               out_prncp
##                                 0.0000
##                           out_prncp_inv
##                                 0.0000
##                             total_pymnt
##                                 0.0000
##                         total_pymnt_inv
##                                 0.0000
##                         total_rec_prncp
##                                 0.0000
##                           total_rec_int
##                                 0.0000
##                       total_rec_late_fee
##                                 0.0000
##                            recoveries
##                                 0.0000
##                     collection_recovery_fee
##                                 0.0000
##                          last_pymnt_d
##                                 0.0006
##                         last_pymnt_amnt
```

```
##                                 0.0000
##                           next_pymnt_d
##                                 0.9716
##                        last_credit_pull_d
##                                 0.0001
##                      last_fico_range_high
##                                 0.0000
##                       last_fico_range_low
##                                 0.0000
##                 collections_12_mths_ex_med
##                                 0.0000
##                mths_since_last_major_derog
##                                 0.7685
##                            policy_code
##                                 0.0000
##                       application_type
##                                 0.0000
##                       annual_inc_joint
##                                 1.0000
##                               dti_joint
##                                 1.0000
##                  verification_status_joint
##                                 1.0000
##                          acc_now_delinq
##                                 0.0000
##                            tot_coll_amt
##                                 0.0638
##                             tot_cur_bal
##                                 0.0638
##                            open_acc_6m
##                                 1.0000
##                            open_act_il
##                                 1.0000
##                            open_il_12m
##                                 1.0000
##                            open_il_24m
##                                 1.0000
##                        mths_since_rcnt_il
##                                 1.0000
##                           total_bal_il
##                                 1.0000
##                               il_util
##                                 1.0000
##                            open_rv_12m
##                                 1.0000
##                            open_rv_24m
##                                 1.0000
##                             max_bal_bc
##                                 1.0000
##                               all_util
##                                 1.0000
##                          total_rev_hi_lim
##                                 0.0638
##                                 inq_fi
```

```
##                                 1.0000
##                            total_cu_tl
##                                 1.0000
##                           inq_last_12m
##                                 1.0000
##                     acc_open_past_24mths
##                                 0.0152
##                            avg_cur_bal
##                                 0.0639
##                         bc_open_to_buy
##                                 0.0245
##                                bc_util
##                                 0.0252
##                  chargeoff_within_12_mths
##                                 0.0000
##                             delinq_amnt
##                                 0.0000
##                       mo_sin_old_il_acct
##                                 0.0954
##                     mo_sin_old_rev_tl_op
##                                 0.0638
##                    mo_sin_rcnt_rev_tl_op
##                                 0.0638
##                           mo_sin_rcnt_tl
##                                 0.0638
##                                mort_acc
##                                 0.0152
##                     mths_since_recent_bc
##                                 0.0236
##                  mths_since_recent_bc_dlq
##                                 0.7576
##                    mths_since_recent_inq
##                                 0.1150
##              mths_since_recent_revol_delinq
##                                 0.6664
##                    num_accts_ever_120_pd
##                                 0.0638
##                           num_actv_bc_tl
##                                 0.0638
##                          num_actv_rev_tl
##                                 0.0638
##                             num_bc_sats
##                                 0.0343
##                               num_bc_tl
##                                 0.0638
##                               num_il_tl
##                                 0.0638
##                            num_op_rev_tl
##                                 0.0638
##                            num_rev_accts
##                                 0.0638
##                       num_rev_tl_bal_gt_0
##                                 0.0638
##                                num_sats
```

```
##                                  0.0343
##                        num_tl_120dpd_2m
##                                  0.0837
##                           num_tl_30dpd
##                                  0.0638
##                       num_tl_90g_dpd_24m
##                                  0.0638
##                       num_tl_op_past_12m
##                                  0.0638
##                          pct_tl_nvr_dlq
##                                  0.0645
##                        percent_bc_gt_75
##                                  0.0247
##                     pub_rec_bankruptcies
##                                  0.0000
##                                tax_liens
##                                  0.0000
##                           tot_hi_cred_lim
##                                  0.0638
##                         total_bal_ex_mort
##                                  0.0152
##                            total_bc_limit
##                                  0.0152
##               total_il_high_credit_limit
##                                  0.0638
##                              revol_bal_joint
##                                  1.0000
##                     sec_app_fico_range_low
##                                  1.0000
##                    sec_app_fico_range_high
##                                  1.0000
##                   sec_app_earliest_cr_line
##                                  1.0000
##                     sec_app_inq_last_6mths
##                                  1.0000
##                           sec_app_mort_acc
##                                  1.0000
##                           sec_app_open_acc
##                                  1.0000
##                        sec_app_revol_util
##                                  1.0000
##                       sec_app_open_act_il
##                                  1.0000
##                      sec_app_num_rev_accts
##                                  1.0000
##           sec_app_chargeoff_within_12_mths
##                                  1.0000
##         sec_app_collections_12_mths_ex_med
##                                  1.0000
##        sec_app_mths_since_last_major_derog
##                                  1.0000
##                              hardship_flag
##                                  0.0000
##                              hardship_type
```

```
##                                     0.9989
##                              hardship_reason
##                                     0.9989
##                              hardship_status
##                                     0.9989
##                                deferral_term
##                                     0.9989
##                              hardship_amount
##                                     0.9989
##                           hardship_start_date
##                                     0.9989
##                             hardship_end_date
##                                     0.9989
##                       payment_plan_start_date
##                                     0.9989
##                                hardship_length
##                                     0.9989
##                                   hardship_dpd
##                                     0.9989
##                           hardship_loan_status
##                                     0.9989
## orig_projected_additional_accrued_interest
##                                     0.9990
##                hardship_payoff_balance_amount
##                                     0.9989
##                  hardship_last_payment_amount
##                                     0.9989
##                            disbursement_method
##                                     0.0000
##                          debt_settlement_flag
##                                     0.0000
##                     debt_settlement_flag_date
##                                     0.9838
##                               settlement_status
##                                     0.9838
##                                 settlement_date
##                                     0.9838
##                               settlement_amount
##                                     0.9838
##                          settlement_percentage
##                                     0.9838
##                                settlement_term
##                                     0.9838
##                                       year
##                                     0.0000
```

Selecting columns with less than 50% missing values

```
lend <- lend[,which(colMeans(!is.na(lend)) > 0.5)]
```

```
apply(lend, 2, function(x)sum(is.na(x))/length(x))
```

```
##                         id                  loan_amnt
##                     0.0000                     0.0000
##                funded_amnt            funded_amnt_inv
##                     0.0000                     0.0000
##                       term                   int_rate
##                     0.0000                     0.0000
##                installment                      grade
##                     0.0000                     0.0000
##                  sub_grade                  emp_title
##                     0.0000                     0.0609
##                 emp_length             home_ownership
##                     0.0470                     0.0000
##                 annual_inc        verification_status
##                     0.0000                     0.0000
##                    issue_d                loan_status
##                     0.0000                     0.0000
##                pymnt_plan                        url
##                     0.0000                     0.0000
##                    purpose                      title
##                     0.0000                     0.0000
##                   zip_code                 addr_state
##                     0.0000                     0.0000
##                        dti                delinq_2yrs
##                     0.0000                     0.0000
##           earliest_cr_line             fico_range_low
##                     0.0000                     0.0000
##            fico_range_high             inq_last_6mths
##                     0.0000                     0.0000
##                   open_acc                    pub_rec
##                     0.0000                     0.0000
##                  revol_bal                  revol_util
##                     0.0000                     0.0005
##                  total_acc        initial_list_status
##                     0.0000                     0.0000
##                  out_prncp              out_prncp_inv
##                     0.0000                     0.0000
##                total_pymnt            total_pymnt_inv
##                     0.0000                     0.0000
##            total_rec_prncp              total_rec_int
##                     0.0000                     0.0000
##         total_rec_late_fee                 recoveries
##                     0.0000                     0.0000
##    collection_recovery_fee               last_pymnt_d
##                     0.0000                     0.0006
##            last_pymnt_amnt         last_credit_pull_d
##                     0.0000                     0.0001
##       last_fico_range_high        last_fico_range_low
##                     0.0000                     0.0000
## collections_12_mths_ex_med                policy_code
##                     0.0000                     0.0000
##           application_type             acc_now_delinq
```

```
##                     0.0000                     0.0000
##               tot_coll_amt               tot_cur_bal
##                     0.0638                     0.0638
##           total_rev_hi_lim       acc_open_past_24mths
##                     0.0638                     0.0152
##                avg_cur_bal             bc_open_to_buy
##                     0.0639                     0.0245
##                    bc_util   chargeoff_within_12_mths
##                     0.0252                     0.0000
##                delinq_amnt         mo_sin_old_il_acct
##                     0.0000                     0.0954
##       mo_sin_old_rev_tl_op      mo_sin_rcnt_rev_tl_op
##                     0.0638                     0.0638
##             mo_sin_rcnt_tl                   mort_acc
##                     0.0638                     0.0152
##       mths_since_recent_bc      mths_since_recent_inq
##                     0.0236                     0.1150
##      num_accts_ever_120_pd             num_actv_bc_tl
##                     0.0638                     0.0638
##            num_actv_rev_tl                num_bc_sats
##                     0.0638                     0.0343
##                  num_bc_tl                  num_il_tl
##                     0.0638                     0.0638
##              num_op_rev_tl              num_rev_accts
##                     0.0638                     0.0638
##        num_rev_tl_bal_gt_0                   num_sats
##                     0.0638                     0.0343
##           num_tl_120dpd_2m               num_tl_30dpd
##                     0.0837                     0.0638
##         num_tl_90g_dpd_24m         num_tl_op_past_12m
##                     0.0638                     0.0638
##             pct_tl_nvr_dlq           percent_bc_gt_75
##                     0.0645                     0.0247
##       pub_rec_bankruptcies                  tax_liens
##                     0.0000                     0.0000
##             tot_hi_cred_lim          total_bal_ex_mort
##                     0.0638                     0.0152
##             total_bc_limit total_il_high_credit_limit
##                     0.0152                     0.0638
##              hardship_flag        disbursement_method
##                     0.0000                     0.0000
##       debt_settlement_flag                       year
##                     0.0000                     0.0000
```

Selcting relevant predictor variables

```
lend <- lend %>%
  dplyr::select(annual_inc, loan_amnt, verification_status, fico_range_high, grade, total_acc, loan_sta

head(lend)
```

```
##   annual_inc loan_amnt verification_status fico_range_high grade total_acc
```

```
## 1    35000.0     12175        Not Verified             699     D        21
## 2    71614.1     15000        Not Verified             674     D        14
## 3    35000.0     15000            Verified             669     E         9
## 4    75000.0     15000     Source Verified             689     A        54
## 5    29000.0     10000        Not Verified             719     A        35
## 6    70000.0      4000            Verified             679     C        31
##   loan_status inq_last_6mths emp_length home_ownership            purpose
## 1 Charged Off              3    7 years           RENT debt_consolidation
## 2  Fully Paid              1    3 years           RENT        credit_card
## 3  Fully Paid              0     1 year           RENT                car
## 4  Fully Paid              0  10+ years       MORTGAGE debt_consolidation
## 5 Charged Off              0    4 years       MORTGAGE   home_improvement
## 6  Fully Paid              0       <NA>       MORTGAGE   home_improvement
##   int_rate tot_cur_bal
## 1    17.77       44692
## 2    18.49       19718
## 3    20.50          NA
## 4     8.39      243234
## 5     6.62       93725
## 6    13.98       78233
```

```
lend <- lend %>%
  drop_na()
```

Exploring the levels of the categorical Variables

```
lend %>%
  distinct(verification_status)
```

```
##   verification_status
## 1        Not Verified
## 2     Source Verified
## 3            Verified
```

```
lend %>%
  distinct(grade)
```

```
##   grade
## 1     D
## 2     A
## 3     E
## 4     B
## 5     F
## 6     C
## 7     G
```

```
lend %>%
  distinct(loan_status)
```

```
##          loan_status
## 1        Charged Off
## 2         Fully Paid
## 3            Current
## 4 Late (31-120 days)
## 5    In Grace Period
## 6  Late (16-30 days)
```

```
lend %>%
  distinct(emp_length)
```

```
##    emp_length
## 1     7 years
## 2     3 years
## 3    10+ years
## 4     4 years
## 5     5 years
## 6     2 years
## 7     6 years
## 8      1 year
## 9     8 years
## 10   < 1 year
## 11    9 years
```

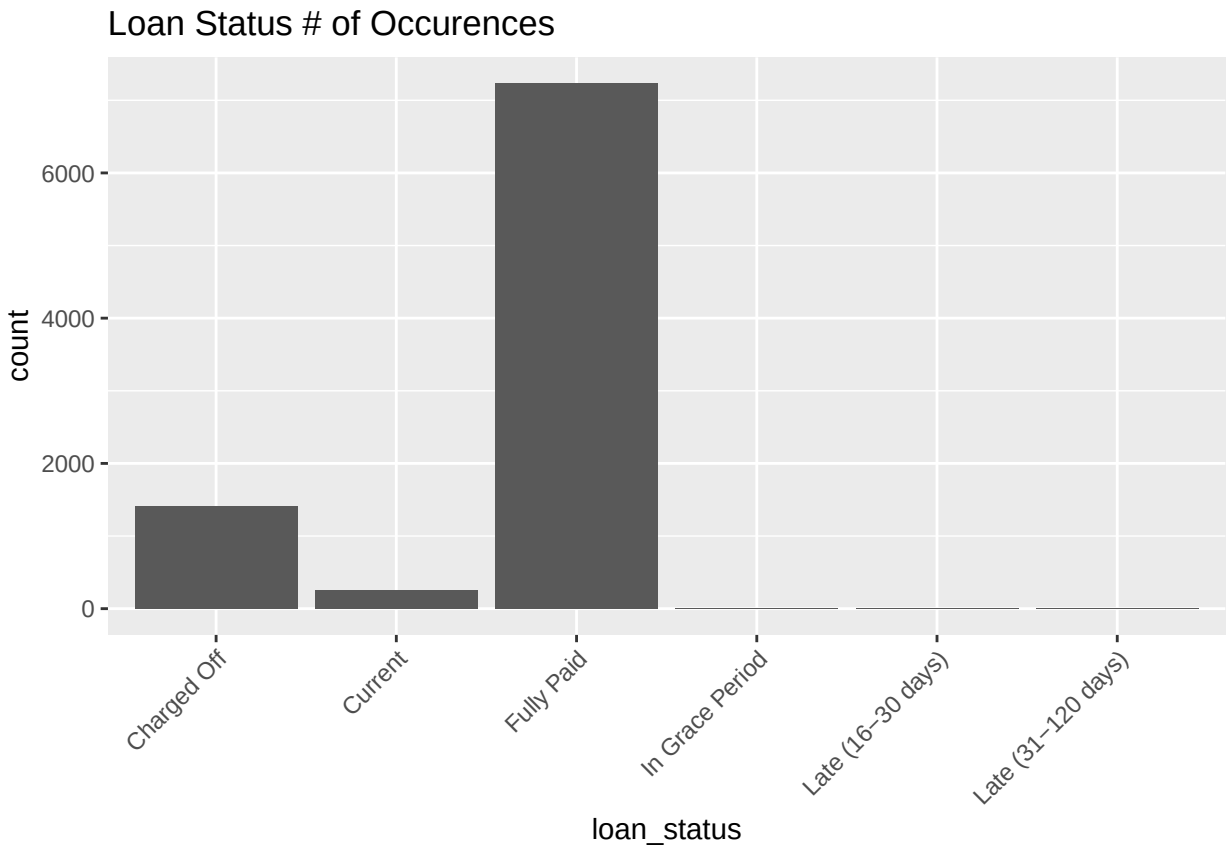
```
lend %>%
  distinct(purpose)
```

```
##               purpose
## 1  debt_consolidation
## 2         credit_card
## 3    home_improvement
## 4               other
## 5               house
## 6                 car
## 7              moving
## 8            vacation
## 9             medical
## 10     major_purchase
## 11     small_business
## 12            wedding
## 13   renewable_energy
```

```
lend %>%
  distinct(home_ownership)
```

```
##   home_ownership
## 1           RENT
## 2       MORTGAGE
## 3            OWN
## 4          OTHER
```

```
lend %>%
  ggplot(aes(x = loan_status)) +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Loan Status # of Occurences")
```




```
table(lend$loan_status)
```

```
##
##        Charged Off            Current         Fully Paid    In Grace Period
##               1415                260               7230                  5
##  Late (16-30 days) Late (31-120 days)
##                  2                  6
```

Since we are only interested about response variable being a 2 level factor, we are filtering rows that are not "Charged Off" or "Fully Paid"

```
lend <- lend %>%
  filter(loan_status == "Charged Off" | loan_status == "Fully Paid")
```

Reducing the number of categories in emp_length

```
lend %>%
  distinct(emp_length)
```

```
##    emp_length
## 1     7 years
## 2     3 years
## 3    10+ years
## 4     4 years
## 5     2 years
## 6     6 years
## 7      1 year
## 8     8 years
## 9    < 1 year
## 10    5 years
## 11    9 years
```

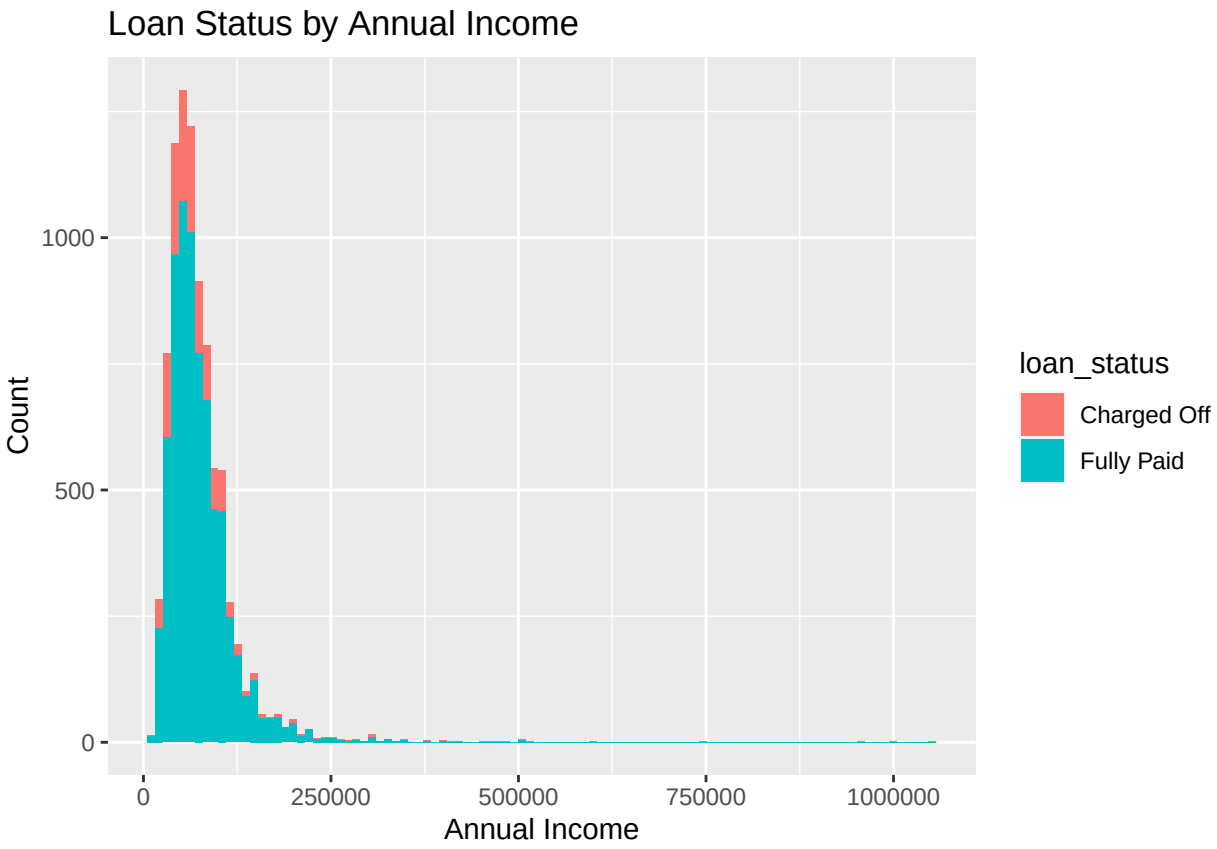
```
lend$emp_length[lend$emp_length == "< 1 year" | lend$emp_length == "1 year" | lend$emp_length == "2 yea
lend$emp_length[lend$emp_length == "4 years" | lend$emp_length == "5 years" | lend$emp_length == "6 yea
lend$emp_length[lend$emp_length == "8 years" | lend$emp_length == "9 years" | lend$emp_length == "10+ y
```

Plotting the histogram of the distribution of the numeric variables

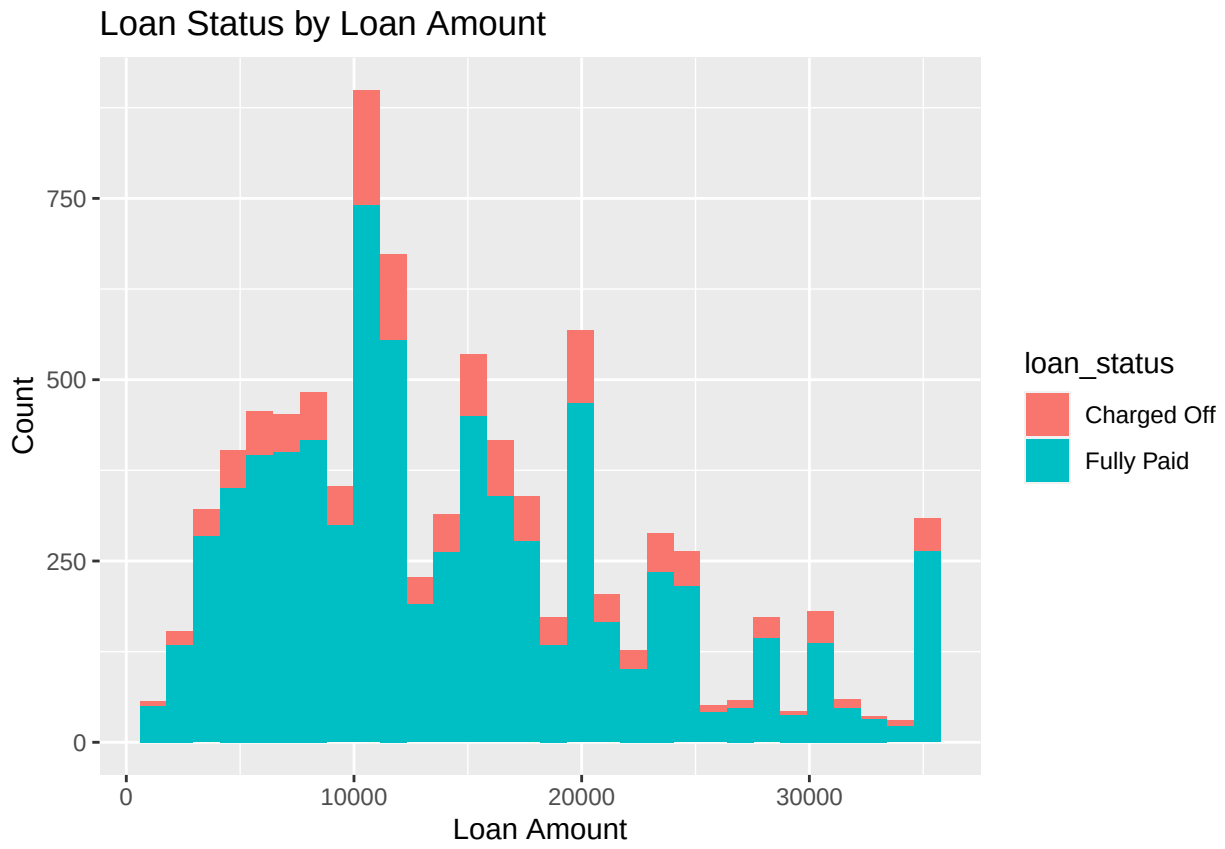
```
select_if(lend, is.numeric) %>%
  head()
```

```
##   annual_inc loan_amnt fico_range_high total_acc inq_last_6mths int_rate
## 1    35000.0     12175             699        21              3    17.77
## 2    71614.1     15000             674        14              1    18.49
## 3    75000.0     15000             689        54              0     8.39
## 4    29000.0     10000             719        35              0     6.62
## 5    75000.0      6000             694        37              0    10.16
## 6    80000.0      6000             704        19              0     9.67
##   tot_cur_bal
## 1       44692
## 2       19718
## 3      243234
## 4       93725
## 5      656431
## 6      166464
```

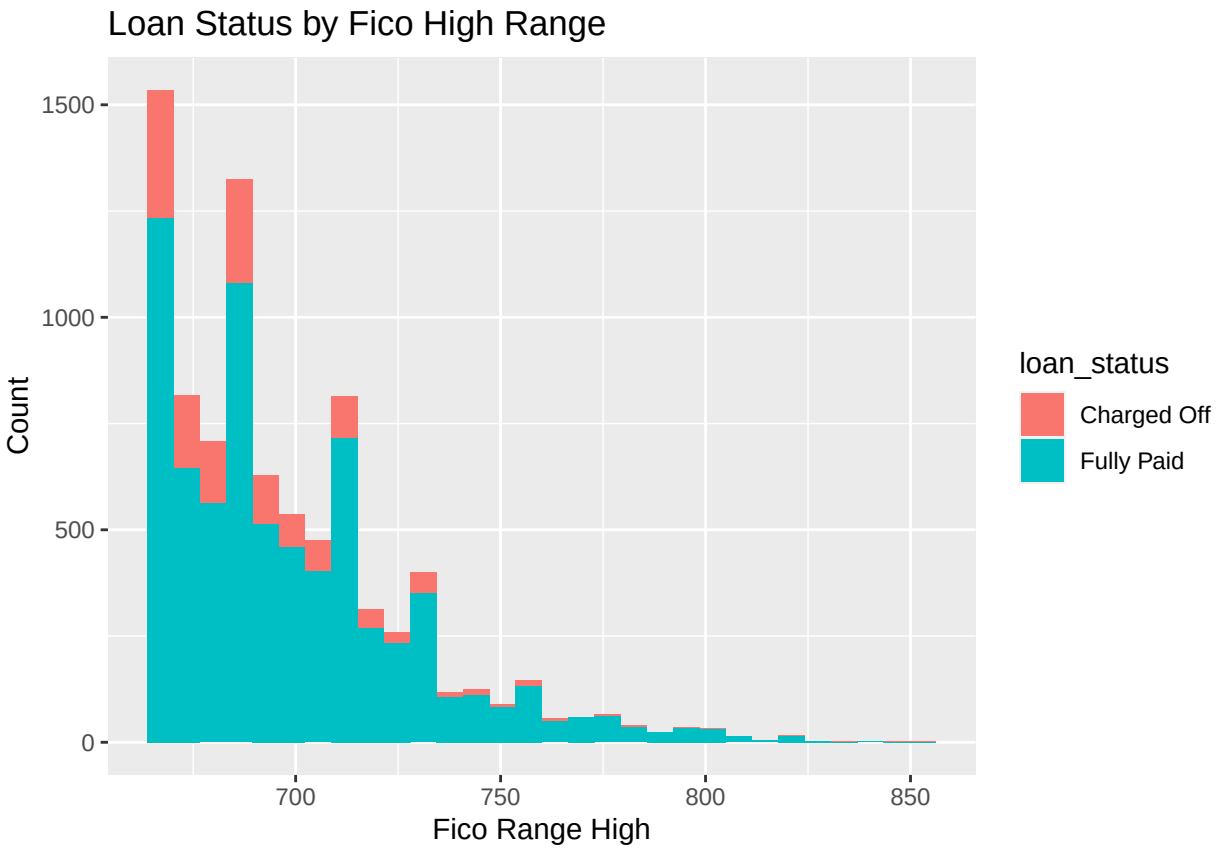
```
ggplot(lend, aes(x = annual_inc, fill = loan_status)) +
  geom_histogram(bins = 100) +
  labs(title = "Loan Status by Annual Income", x = "Annual Income", y = "Count")
```



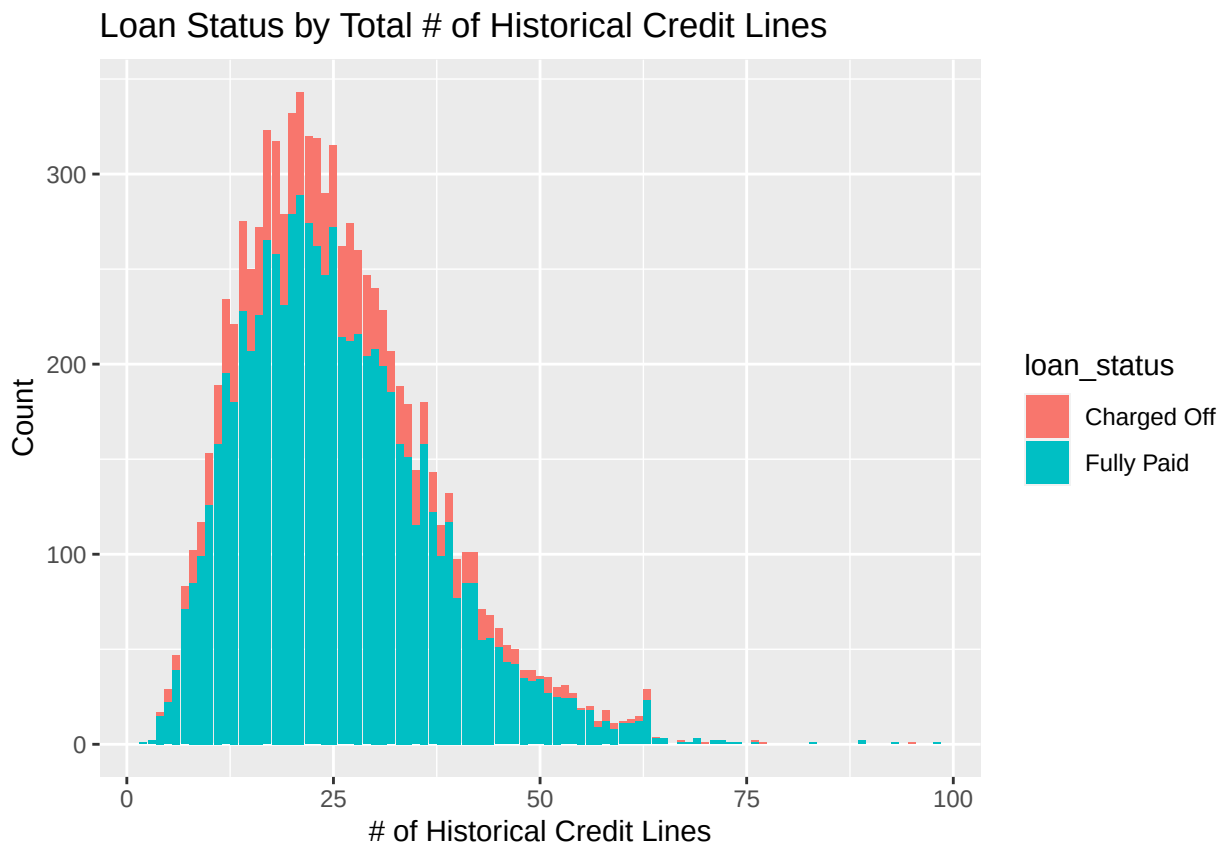

```
ggplot(lend, aes(x = loan_amnt, fill = loan_status)) +
  geom_histogram(bins = 30) +
  labs(title = "Loan Status by Loan Amount", x= "Loan Amount", y = "Count")
```



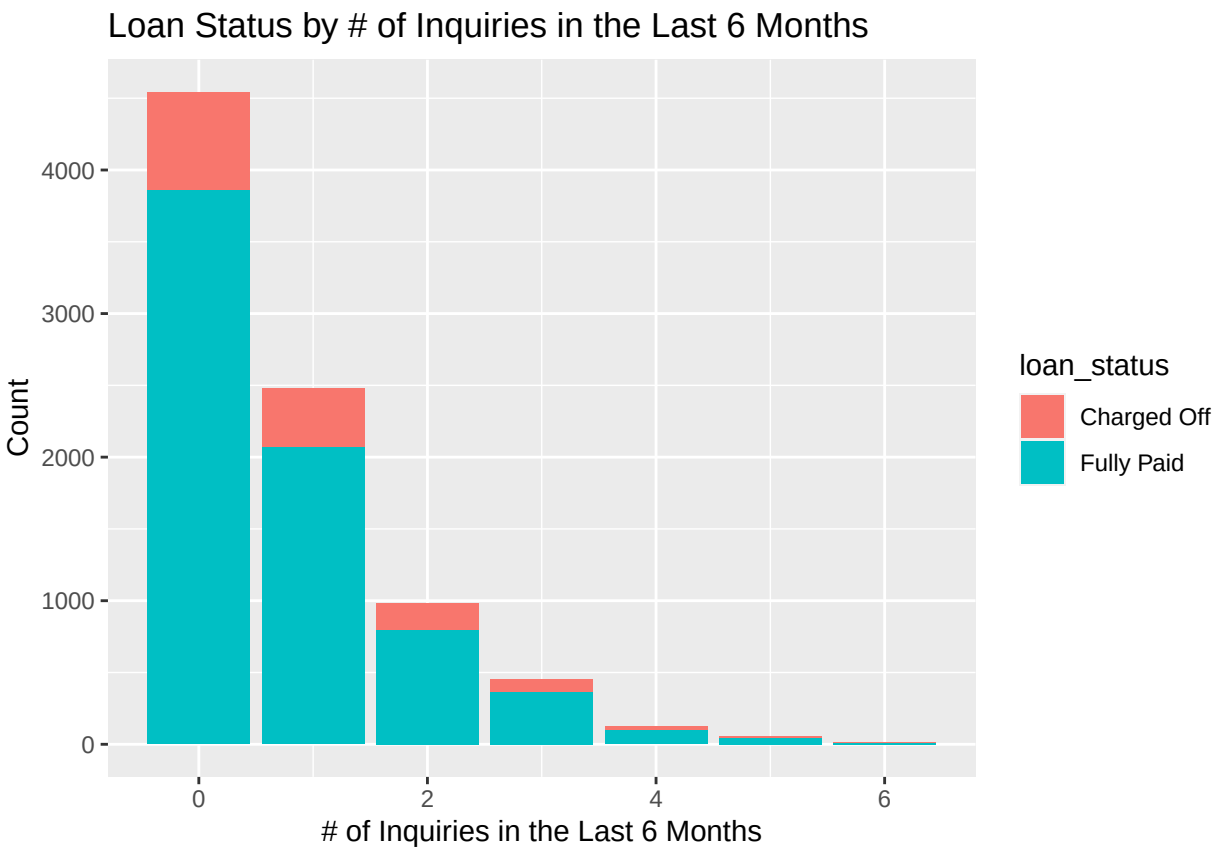

```
ggplot(lend, aes(x = fico_range_high, fill = loan_status)) +
  geom_histogram(bins = 30) +
  labs(title = "Loan Status by Fico High Range", x = "Fico Range High", y = "Count")
```



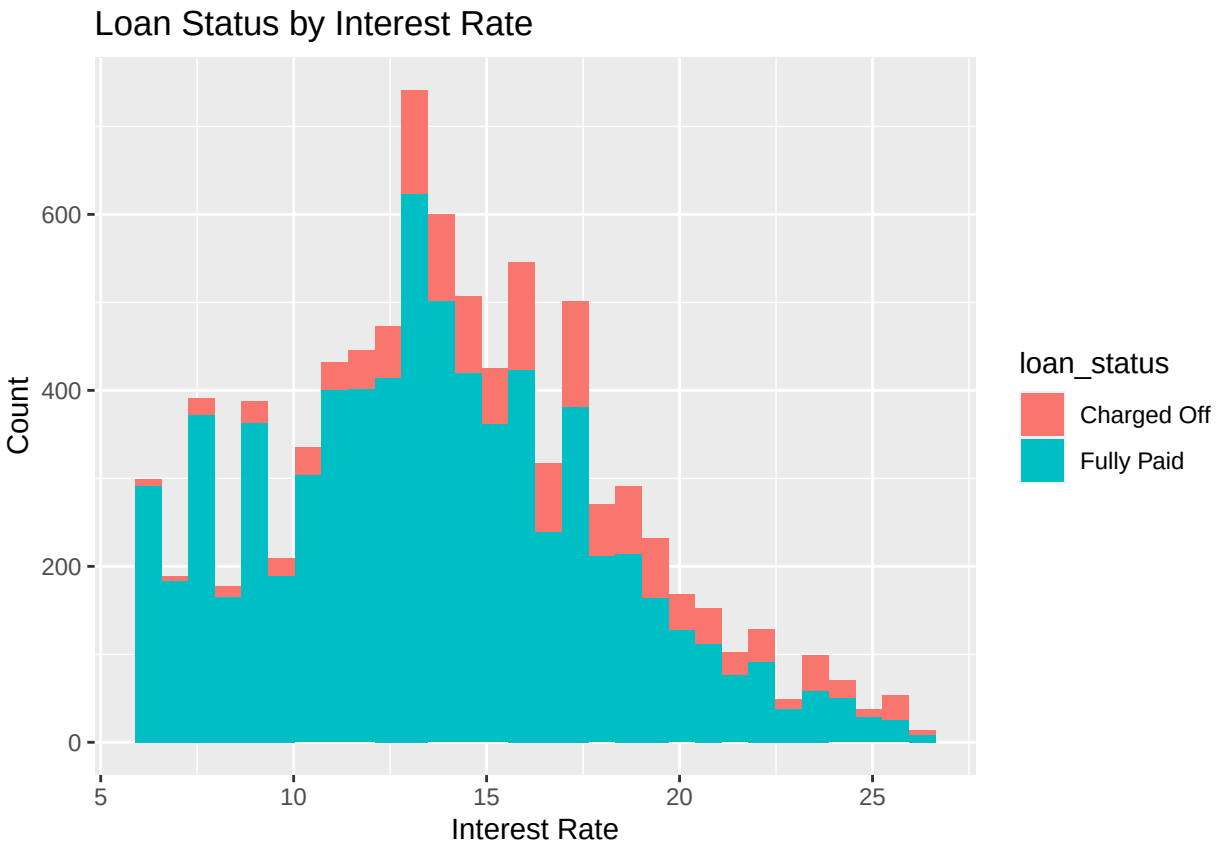

```
ggplot(lend, aes(x = total_acc, fill = loan_status)) +
  geom_bar() +
  labs(title = "Loan Status by Total # of Historical Credit Lines", x= "# of Historical Credit Lines", y
```



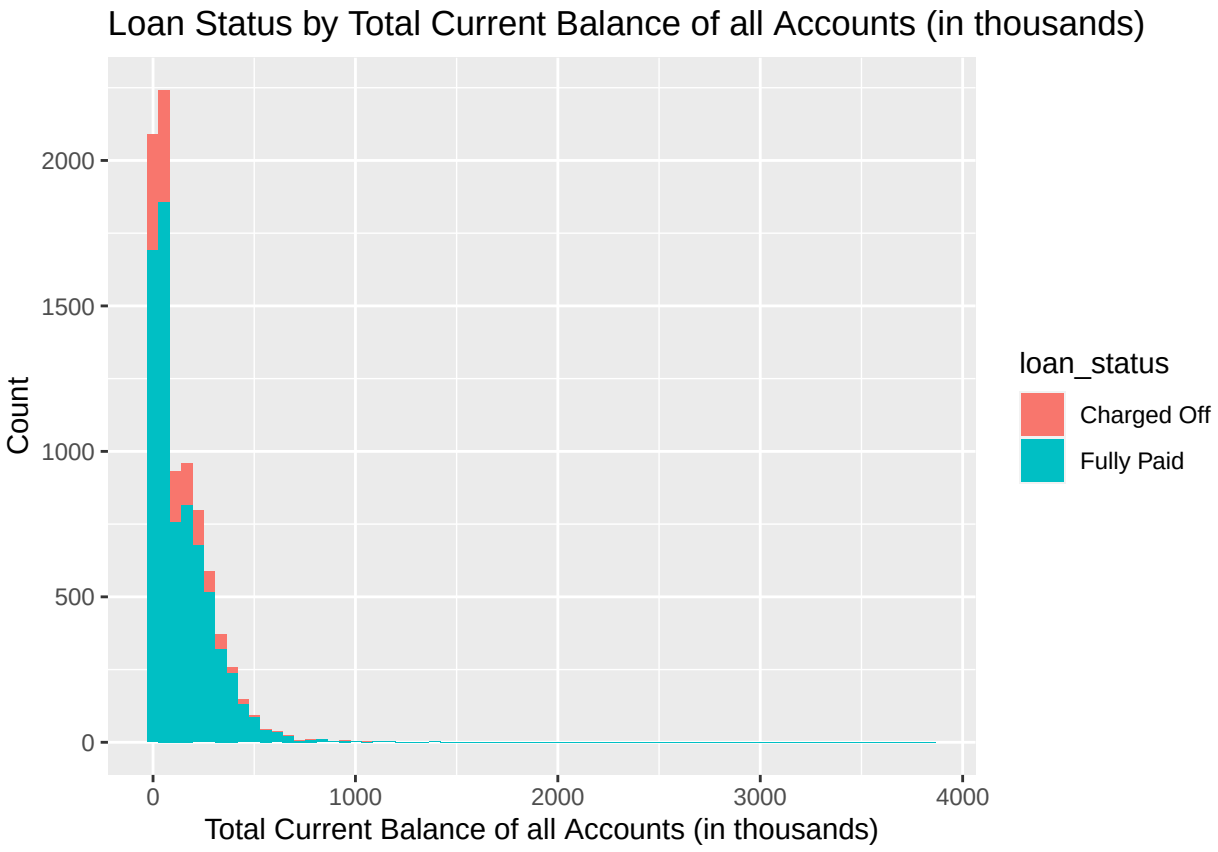

```
ggplot(lend, aes(x = inq_last_6mths, fill = loan_status)) +
  geom_bar() +
  labs(title = "Loan Status by # of Inquiries in the Last 6 Months", x = "# of Inquiries in the Last 6 M
```




```
ggplot(lend, aes(x = int_rate, fill = loan_status)) +
  geom_histogram(bins = 30) +
  labs(title = "Loan Status by Interest Rate", x = "Interest Rate", y = "Count")
```




```
ggplot(lend, aes(x = tot_cur_bal/1000, fill = loan_status)) +
  geom_histogram(bins = 70) +
  labs(title = "Loan Status by Total Current Balance of all Accounts (in thousands)", x = "Total Curren
```



Changing the Vatiables to the correct data types

```
head(lend)
```

```
##   annual_inc loan_amnt verification_status fico_range_high grade total_acc
## 1    35000.0     12175        Not Verified             699     D        21
## 2    71614.1     15000        Not Verified             674     D        14
## 3    75000.0     15000     Source Verified             689     A        54
## 4    29000.0     10000        Not Verified             719     A        35
## 5    75000.0      6000     Source Verified             694     B        37
## 6    80000.0      6000     Source Verified             704     B        19
##   loan_status inq_last_6mths emp_length home_ownership            purpose
## 1 Charged Off              3  4-7 years           RENT debt_consolidation
## 2  Fully Paid              1  0-3 years           RENT        credit_card
## 3  Fully Paid              0   8+ years       MORTGAGE debt_consolidation
## 4 Charged Off              0  4-7 years       MORTGAGE   home_improvement
## 5  Fully Paid              0  0-3 years       MORTGAGE debt_consolidation
## 6  Fully Paid              0   8+ years       MORTGAGE debt_consolidation
##   int_rate tot_cur_bal
## 1    17.77       44692
## 2    18.49       19718
## 3     8.39      243234
## 4     6.62       93725
## 5    10.16      656431
```

```
## 6     9.67      166464
```

```
lend$verification_status <- as.factor(lend$verification_status)
lend$loan_status <- as.factor(lend$loan_status)
lend$grade <- as.factor(lend$grade)
lend$emp_length <- as.factor(lend$emp_length)
lend$home_ownership <- as.factor(lend$home_ownership)
lend$purpose <- as.factor(lend$purpose)
```

Creating a training and testing set

```
lend_parts <- lend %>%
  initial_split(prop = 0.75)

train <- lend_parts %>%
  training()

test <- lend_parts %>%
  testing

list(train, test) %>%
  map_int(nrow)
```

```
## [1] 6483 2162
```

Null Model

```
lend_null <- logistic_reg(mode = "classification") %>%
  set_engine("glm") %>%
  fit(loan_status ~ ., data = train)
```

```
pred <- train %>%
  dplyr::select(annual_inc, loan_amnt, verification_status, fico_range_high, grade, total_acc, loan_sta
  bind_cols(
    predict(lend_null, new_data = train, type = "class")
  ) %>%
  rename(loan_null = .pred_class)
```

kNN

```
lend_knn <- nearest_neighbor(neighbors = 15) %>%
  set_engine("kknn", scale = TRUE) %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)
```

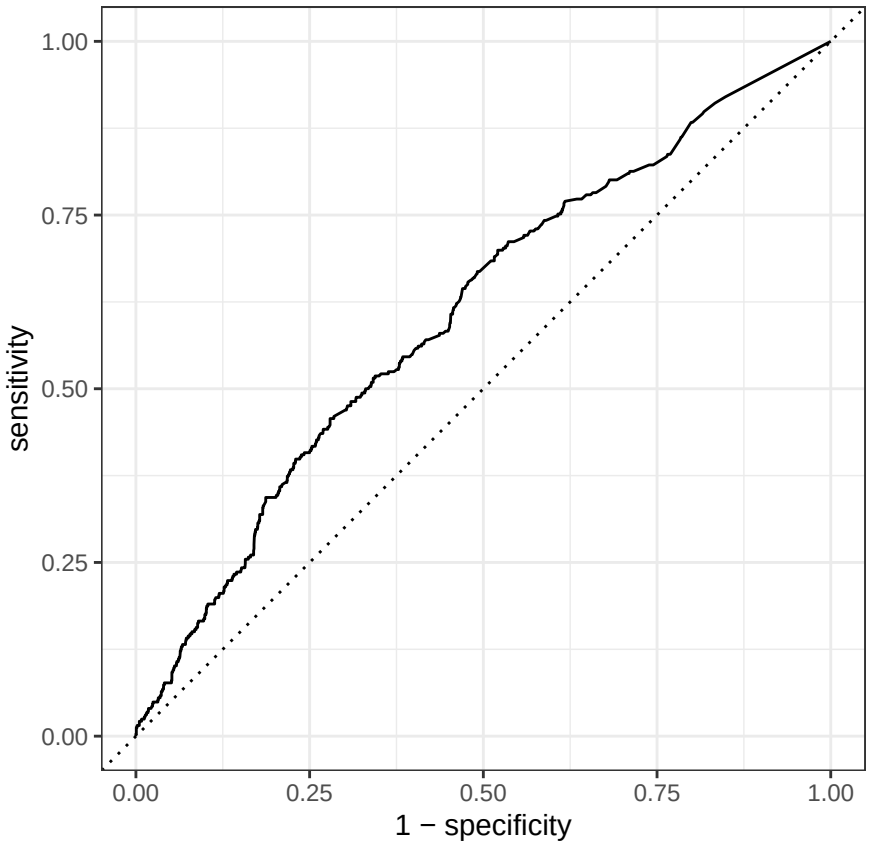
```
lend_knn %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.835
## 2 kap      binary         0.0287
```

```
lend_knn %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off          14         44
##   Fully Paid          312       1792
```

```
lend_knn %>%
  predict(test, type = "prob") %>%
  bind_cols(test) %>%
  roc_curve(loan_status, `.pred_Charged Off`) %>%
  autoplot()
```



Random Forest

```
lend_rf <- rand_forest(trees = 100) %>%
  set_engine("ranger") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)
```

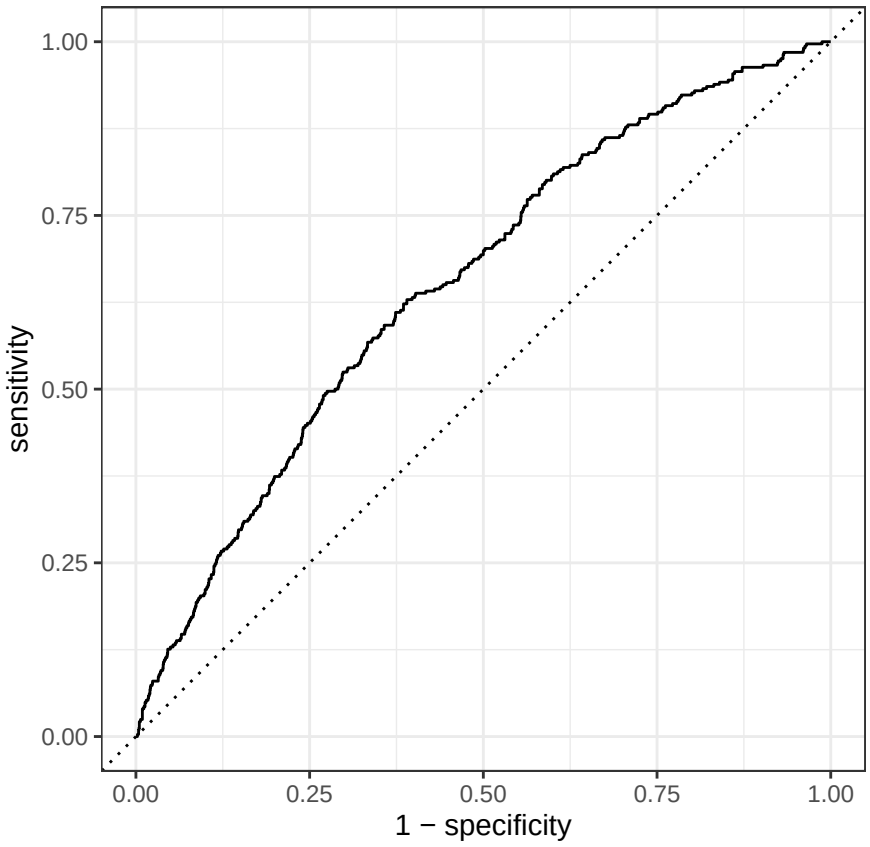
```
lend_rf %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary        0.848
## 2 kap      binary        0.0263
```

```
lend_rf %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off           7         10
##   Fully Paid          319       1826
```

```
lend_rf %>%
  predict(test, type = "prob") %>%
  bind_cols(test) %>%
  roc_curve(loan_status, `.pred_Charged Off`) %>%
  autoplot()
```



Naive Bayes

```
lend_nb <- naive_Bayes(Laplace = 1) %>%
  set_engine("klaR") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)
```

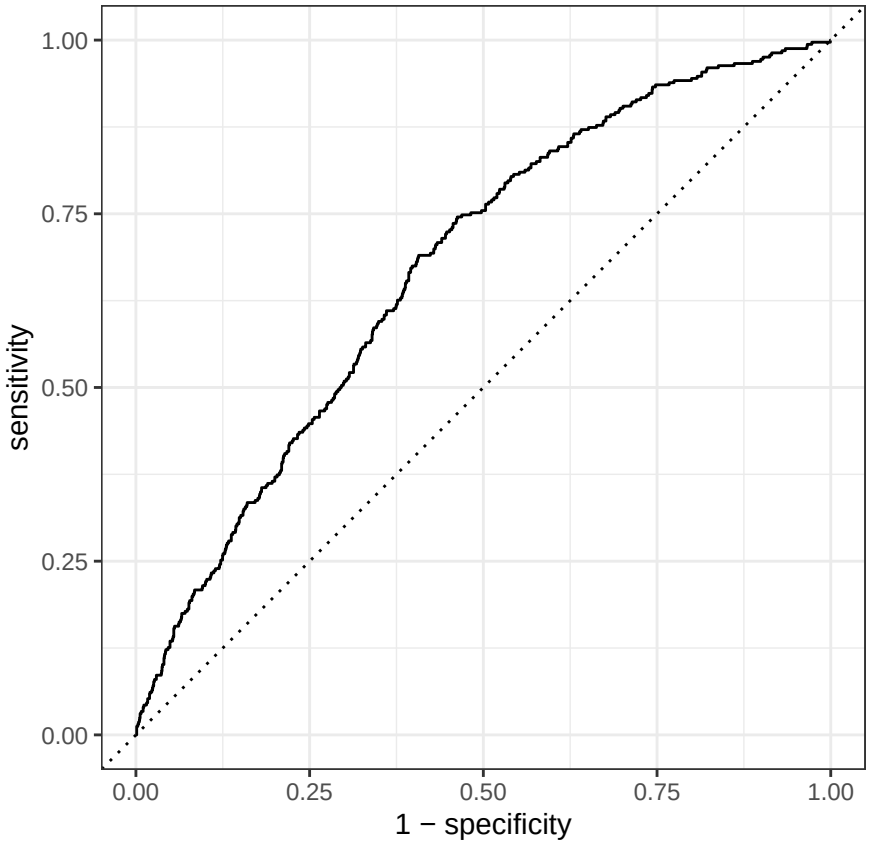
```
suppressWarnings({lend_nb %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)})
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.827
## 2 kap      binary         0.110
```

```
suppressWarnings({lend_nb %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)})
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off          43         90
##   Fully Paid          283       1746
```

```
suppressWarnings({lend_nb %>%
    predict(test, type = "prob") %>%
    bind_cols(test)} %>%
      roc_curve(loan_status, `.pred_Charged Off`)) %>%
  autoplot()
```



GLM using Regularlization

```
lend_glm <- logistic_reg(penalty = .00001, mixture = 0.1) %>%
  set_engine("glmnet") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)
```

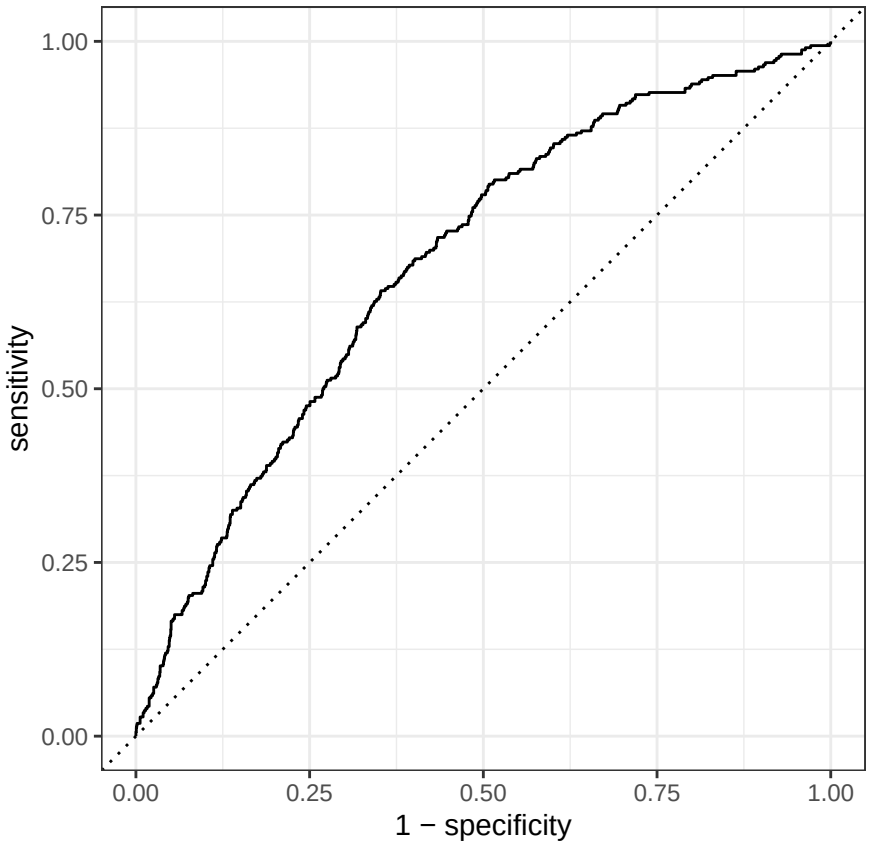
```
lend_glm %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary       0.850
## 2 kap      binary       0.00520
```

```
lend_glm %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off           1          0
##   Fully Paid          325       1836
```

```
lend_glm %>%
  predict(test, type = "prob") %>%
  bind_cols(test) %>%
  roc_curve(loan_status, `.pred_Charged Off`) %>%
  autoplot()
```



XGBoost

```
lend_xgb <- boost_tree(trees = 55) %>%
  set_engine("xgboost") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)
```

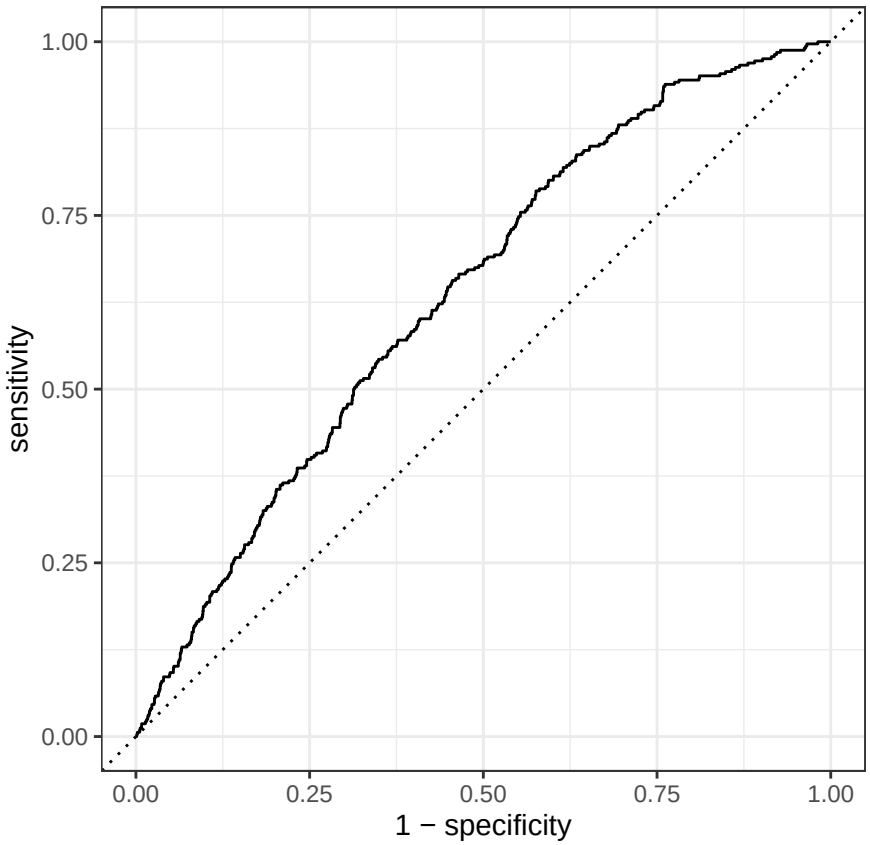
```
lend_xgb %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.833
## 2 kap      binary         0.0433
```

```
lend_xgb %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off          19         53
##   Fully Paid          307       1783
```

```
lend_xgb %>%
  predict(test, type = "prob") %>%
  bind_cols(test) %>%
  roc_curve(loan_status, `.pred_Charged Off`) %>%
  autoplot()
```



C5.0

```
lend_c50 <- boost_tree(trees = 55) %>%
  set_engine("C5.0") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)
```

```
lend_c50 %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.849
## 2 kap      binary         0
```

```
lend_c50 %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off           0          0
##   Fully Paid          326       1836
```

Decision Tree

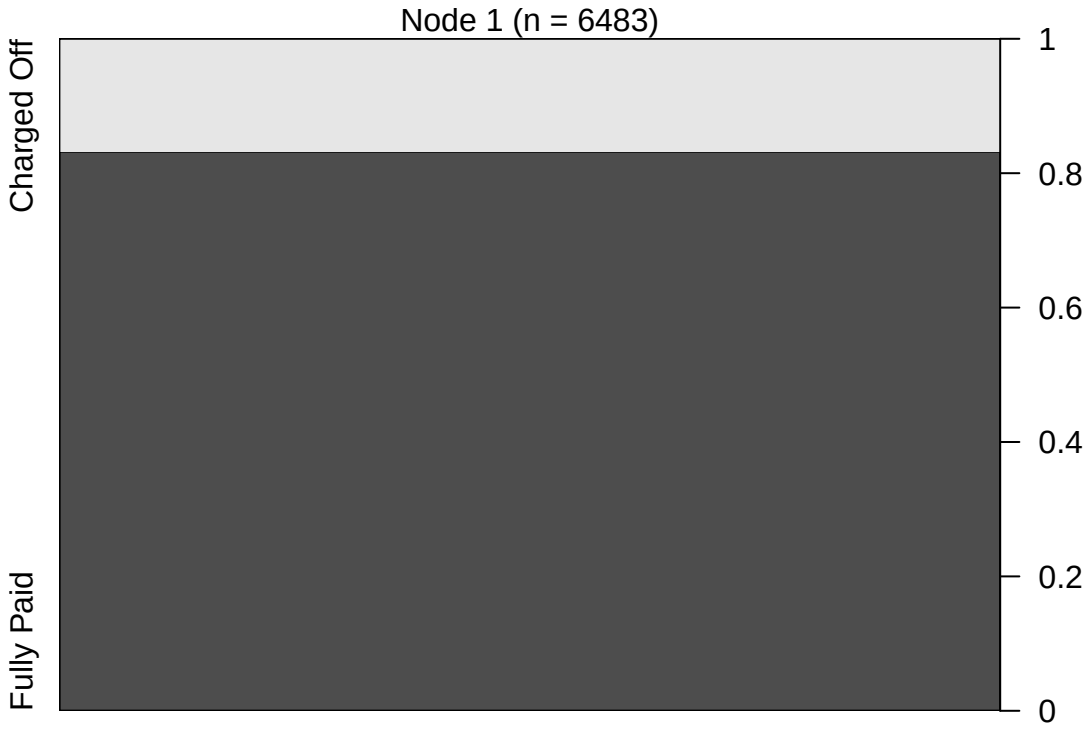
```
lend_dtree <- decision_tree() %>%
  set_engine("rpart", control = rpart.control(cp = 0.003)) %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train)

lend_dtree
```

```
## parsnip model object
##
## n= 6483
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 6483 1089 Fully Paid (0.1679778 0.8320222) *
```

```
lend_dtree %>%
  predict(test) %>%
  bind_cols(test) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.849
## 2 kap      binary         0
```

```
plot(as.party(lend_dtree$fit))
```




```
lend_dtree %>%
  predict(test) %>%
  bind_cols(test) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Charged Off Fully Paid
##   Charged Off           0          0
##   Fully Paid          326       1836
```

```
lend_dtree %>%
  predict(test, type = "prob") %>%
  bind_cols(test) %>%
  roc_curve(loan_status, `.pred_Charged Off`)
```

```
## # A tibble: 3 x 3
##   .threshold specificity sensitivity
##        <dbl>       <dbl>       <dbl>
## 1   -Inf               0           1
## 2      0.168           0           1
## 3    Inf               1           0
```

Building the model with the Subsampled Dataset

As we have observed in the distribution plot of loan_status above, we are dealing with a highly imbalanced dataset. There are only 1415 observed rows of Charged Off and 7230 observed rows of Fully Paid. The predictions we have developed are biased. We are going to create a new dataset using the ROSE and caret package to have an equal number of both occurences to see if it contributes to a better prediction model. Although the data is still overfitted, the percentage of Charged Off has gotten better and will give our model better training data.

```
over_lend <- ovun.sample(loan_status ~ ., data = train, method = "over", N=9000)$data
head(over_lend)
```

```
##   annual_inc loan_amnt verification_status fico_range_high grade total_acc
## 1      53000     15000     Source Verified             669     B        24
## 2     130000     32000            Verified             729     A        25
## 3      23000      5400        Not Verified             679     C        37
## 4      20000      6000        Not Verified             664     C        32
## 5      62500      8000        Not Verified             774     A        41
## 6      75000      8100     Source Verified             674     D        28
##   loan_status inq_last_6mths emp_length home_ownership            purpose
## 1  Fully Paid              0  4-7 years           RENT debt_consolidation
## 2  Fully Paid              0  0-3 years       MORTGAGE debt_consolidation
## 3  Fully Paid              0  0-3 years           RENT debt_consolidation
## 4  Fully Paid              1  0-3 years           RENT debt_consolidation
## 5  Fully Paid              0  4-7 years           RENT        credit_card
## 6  Fully Paid              1  4-7 years           RENT debt_consolidation
##   int_rate tot_cur_bal
## 1    12.85       59430
## 2     8.90      243971
## 3    13.35      131220
## 4    14.49        9200
## 5     6.03       66475
## 6    16.99       57738
```

```
table(over_lend$loan_status)
```

```
##
##  Fully Paid Charged Off
##        5394        3606
```

```
over_lend_parts <- over_lend %>%
  initial_split(prop = 0.8)

train2 <- over_lend_parts %>%
  training()

test2 <- over_lend_parts %>%
  testing()
```

Null Model

```
lend_null2 <- logistic_reg(mode = "classification") %>%
  set_engine("glm") %>%
  fit(loan_status ~  1., data = train2)
```

```
lend_null2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.597
## 2 kap      binary         0
```

kNN

```
lend_knn2 <- nearest_neighbor(neighbors = 25) %>%
  set_engine("kknn", scale = TRUE) %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)
```

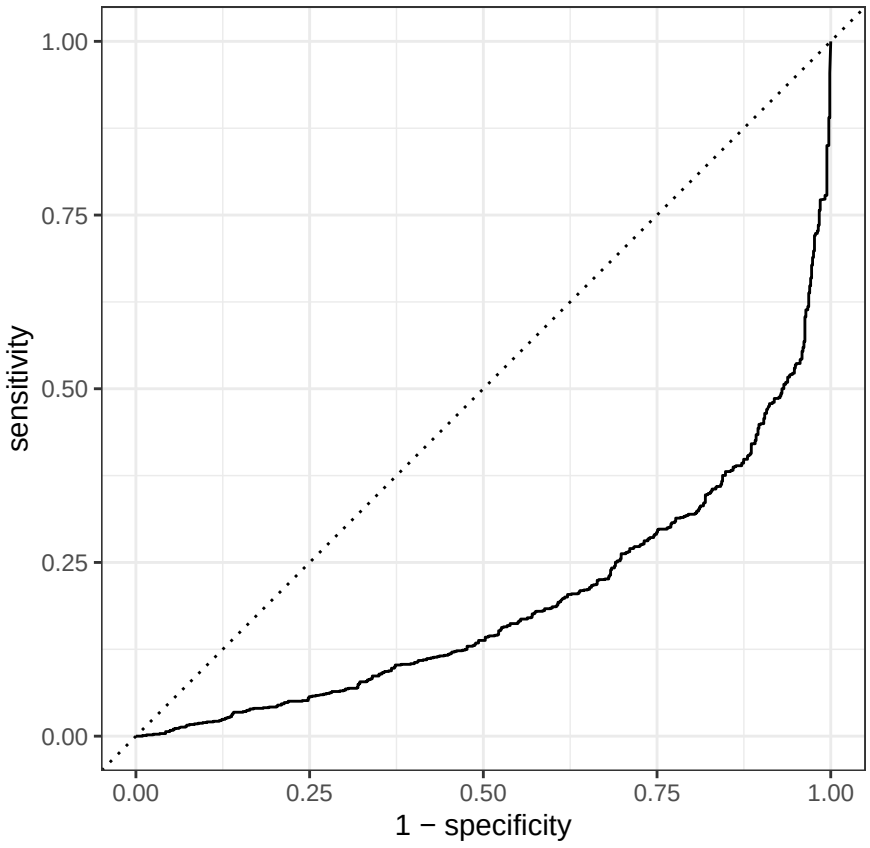
```
lend_knn2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.73
## 2 kap      binary         0.443
```

```
lend_knn2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid         814         226
##   Charged Off        260         500
```

```
lend_knn2 %>%
  predict(test2, type = "prob") %>%
  bind_cols(test2) %>%
  roc_curve(loan_status, `.pred_Charged Off`) %>%
  autoplot()
```



Random Forest

```
lend_rf2 <- rand_forest(trees = 100) %>%
  set_engine("ranger") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)
```

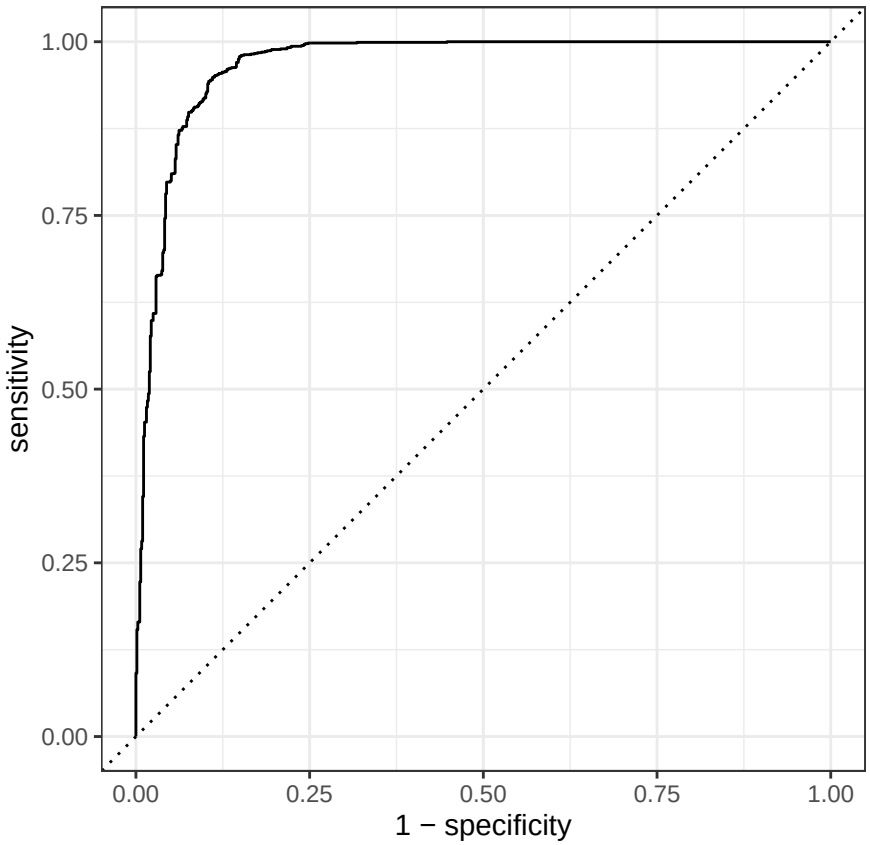
```
lend_rf2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.924
## 2 kap      binary         0.841
```

```
lend_rf2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid        1017          80
##   Charged Off         57         646
```

```
lend_rf2 %>%
  predict(test2, type = "prob") %>%
  bind_cols(test2) %>%
  roc_curve(loan_status, `.pred_Fully Paid`) %>%
  autoplot()
```



Naive Bayes

```
lend_nb2 <- naive_Bayes(Laplace = 1) %>%
  set_engine("klaR") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)
```

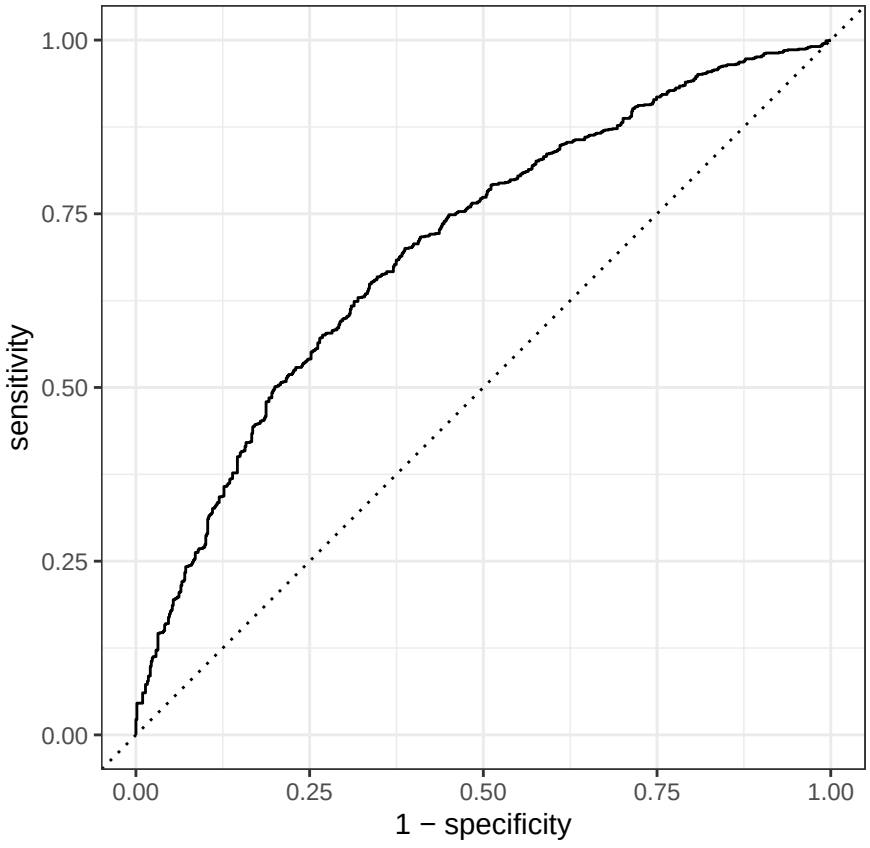
```
suppressWarnings({lend_nb2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)})
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.662
## 2 kap      binary         0.301
```

```
suppressWarnings({lend_nb2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)})
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid         761         295
##   Charged Off        313         431
```

```
suppressWarnings({lend_nb2 %>%
  predict(test2, type = "prob") %>%
  bind_cols(test2) %>%
  roc_curve(loan_status, `.pred_Fully Paid`) %>%
    autoplot()})
```



GLM using Regularlization

```
lend_glm2 <- logistic_reg(penalty = .00001, mixture = 0.1) %>%
  set_engine("glmnet") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)
```

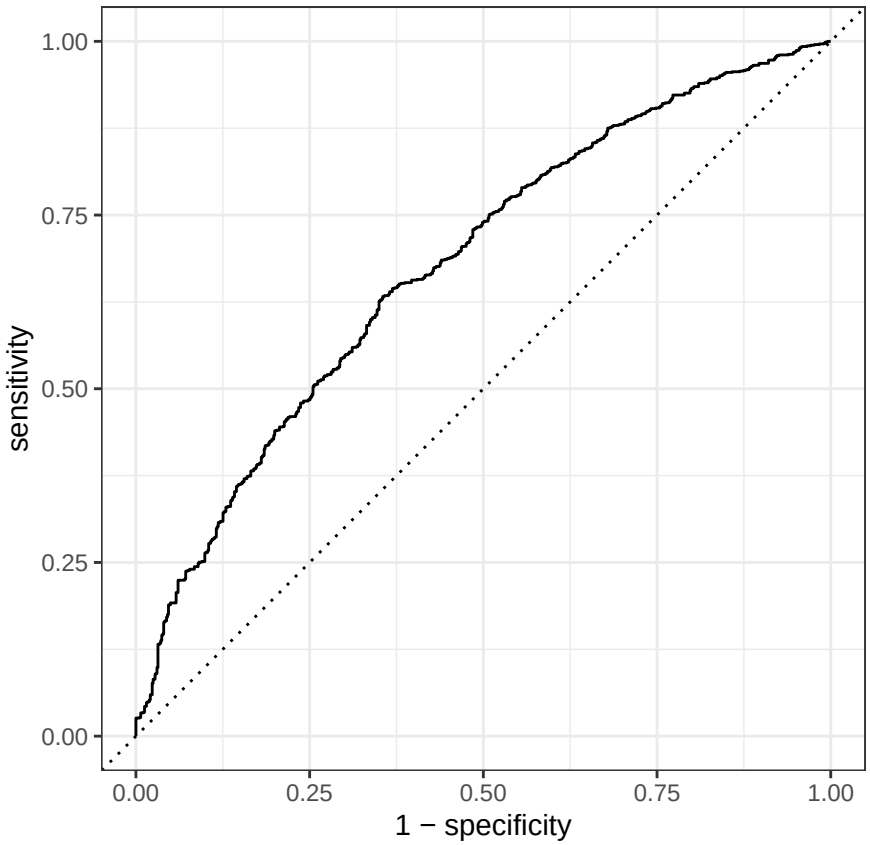
```
lend_glm2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.65
## 2 kap      binary         0.232
```

```
lend_glm2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid         877         433
##   Charged Off        197         293
```

```
lend_glm2 %>%
  predict(test2, type = "prob") %>%
  bind_cols(test2) %>%
  roc_curve(loan_status, `.pred_Fully Paid`) %>%
  autoplot()
```



XGBoost

```
lend_xgb2 <- boost_tree(trees = 55) %>%
  set_engine("xgboost") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)
```

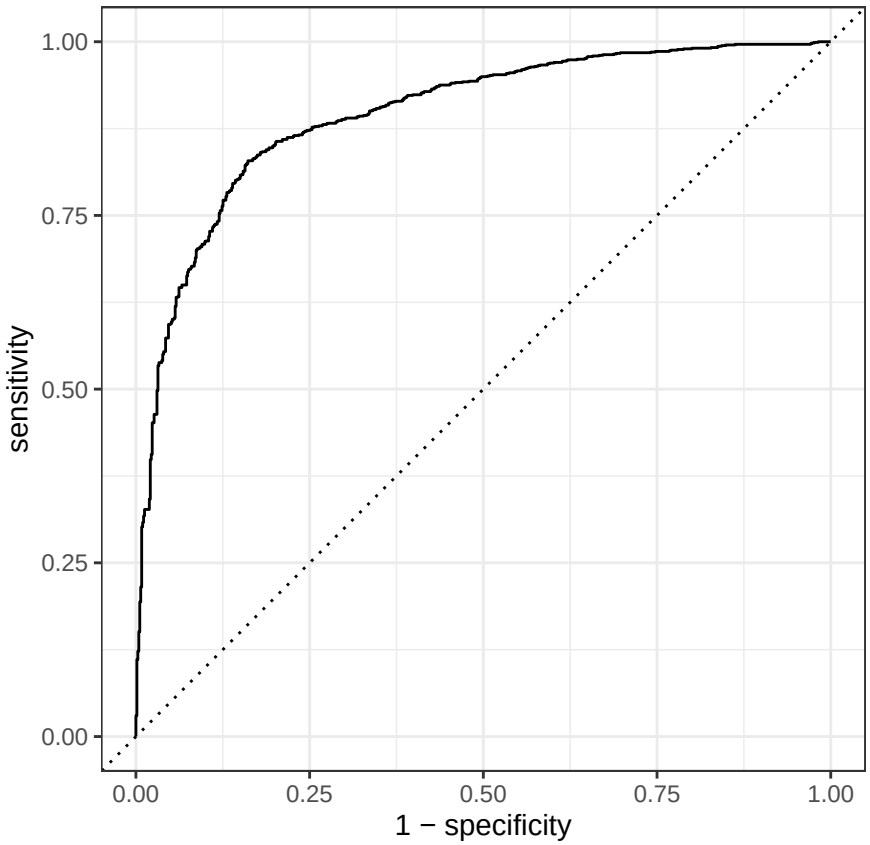
```
lend_xgb2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.831
## 2 kap      binary         0.648
```

```
lend_xgb2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid         920         151
##   Charged Off        154         575
```

```
lend_xgb2 %>%
  predict(test2, type = "prob") %>%
  bind_cols(test2) %>%
  roc_curve(loan_status, `.pred_Fully Paid`) %>%
  autoplot()
```



C5.0

```
lend_c502 <- boost_tree(trees = 55) %>%
  set_engine("C5.0") %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)
```

```
lend_c502 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.931
## 2 kap      binary         0.858
```

```
lend_c502 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid         999          49
##   Charged Off         75         677
```

Decision Tree

```
lend_dtree2 <- decision_tree() %>%
  set_engine("rpart", control = rpart.control(cp = 0.003)) %>%
  set_mode("classification") %>%
  fit(loan_status ~ ., data = train2)

lend_dtree
```

```
## parsnip model object
##
## n= 6483
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
## 1) root 6483 1089 Fully Paid (0.1679778 0.8320222) *
```

```
lend_dtree2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  metrics(truth = loan_status, estimate = .pred_class)
```

```
## # A tibble: 2 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.673
## 2 kap      binary         0.302
```

```
lend_dtree2 %>%
  predict(test2) %>%
  bind_cols(test2) %>%
  conf_mat(truth = loan_status, estimate = .pred_class)
```

```
##              Truth
## Prediction    Fully Paid Charged Off
##   Fully Paid         839         354
##   Charged Off        235         372
```

With the resampled data set, our best model is the C5.0 Decision Tree with a 93.11% accuracy. This model has the least false positive and false negative ratio out of all of the algorithms.

Tuning the Model

```
m_c50_bst <- C5.0(loan_status ~ ., data = train2, trials = 100)
```

```
pred <- predict(m_c50_bst, test2)
confusionMatrix(data=pred, test2$loan_status)
```

```
## Confusion Matrix and Statistics
##
##              Reference
## Prediction    Fully Paid Charged Off
##   Fully Paid        1006          43
##   Charged Off         68         683
##
##                Accuracy : 0.9383
##                  95% CI : (0.9262, 0.949)
##     No Information Rate : 0.5967
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8726
##
##  Mcnemar's Test P-Value : 0.02273
##
##             Sensitivity : 0.9367
##             Specificity : 0.9408
##          Pos Pred Value : 0.9590
##          Neg Pred Value : 0.9095
##              Prevalence : 0.5967
##          Detection Rate : 0.5589
##    Detection Prevalence : 0.5828
##       Balanced Accuracy : 0.9387
##
##        'Positive' Class : Fully Paid
##
```

Conclusion

We have determined that that the C5.0 decision tree model provided the most accurate predictions out of the other machine learning algorithms. Initially our dataset had a problem with an imbalance with more loans that were fully paid versus loans that were charged off. We then used caret to offset the imbalance, not perfectly balanced, but better than our initial data. With the model tuned, we have an accuracy of 94.44%! If the model were to be done differently, instead of using a condensed dataset, the full one should be used and should be compared to see if the same algorithm would be chosen. We can also try experimenting with other withheld variables as well.