

Lending Club Analysis

Max Kutschinski

Shree Karimkolikuzhiyil

Bradley Gravitt

3/9/2022

Missing Values (Part I)

The goal of this section is to identify features that are eligible for feature wise deletion in order to make the data set easier to navigate. Part II will discuss how to handle any remaining missing values.

```
anyNA(data)
```

```
## [1] TRUE
```

```
dim(data)
```

```
## [1] 42537 111
```

```
ggplot_missing(data)
```

```
## Loading required package: reshape2
```

```
## Warning: package 'reshape2' was built under R version 4.0.3
```


##	0.005	0.031
##	zip_code	addr_state
##	0.005	0.005
##	dti	delinq_2yrs
##	0.005	0.073
##	earliest_cr_line	inq_last_6mths
##	0.073	0.073
##	mths_since_last_delinq	mths_since_last_record
##	63.305	91.417
##	open_acc	pub_rec
##	0.073	0.073
##	revol_bal	revol_util
##	0.005	0.216
##	total_acc	initial_list_status
##	0.073	0.005
##	out_prncp	out_prncp_inv
##	0.005	0.005
##	total_pymnt	total_pymnt_inv
##	0.005	0.005
##	total_rec_prncp	total_rec_int
##	0.005	0.005
##	total_rec_late_fee	recoveries
##	0.005	0.005
##	collection_recovery_fee	last_pymnt_d
##	0.005	0.200
##	last_pymnt_amnt	next_pymnt_d
##	0.005	91.581
##	last_credit_pull_d	collections_12_mths_ex_med
##	0.014	0.346
##	mths_since_last_major_derog	policy_code
##	100.000	0.005
##	application_type	annual_inc_joint
##	0.005	100.000
##	dti_joint	verification_status_joint
##	100.000	100.000
##	acc_now_delinq	tot_coll_amt
##	0.073	100.000
##	tot_cur_bal	open_acc_6m
##	100.000	100.000
##	open_il_6m	open_il_12m
##	100.000	100.000
##	open_il_24m	mths_since_rcnt_il
##	100.000	100.000
##	total_bal_il	il_util
##	100.000	100.000
##	open_rv_12m	open_rv_24m
##	100.000	100.000
##	max_bal_bc	all_util
##	100.000	100.000
##	total_rev_hi_lim	inq_fi
##	100.000	100.000
##	total_cu_tl	inq_last_12m
##	100.000	100.000
##	acc_open_past_24mths	avg_cur_bal

```
##          100.000          100.000
##          bc_open_to_buy          bc_util
##          100.000          100.000
##          chargeoff_within_12_mths          delinq_amnt
##          0.346          0.073
##          mo_sin_old_il_acct          mo_sin_old_rev_tl_op
##          100.000          100.000
##          mo_sin_rcnt_rev_tl_op          mo_sin_rcnt_tl
##          100.000          100.000
##          mort_acc          mths_since_recent_bc
##          100.000          100.000
##          mths_since_recent_bc_dlq          mths_since_recent_inq
##          100.000          100.000
##          mths_since_recent_revol_delinq          num_accts_ever_120_pd
##          100.000          100.000
##          num_actv_bc_tl          num_actv_rev_tl
##          100.000          100.000
##          num_bc_sats          num_bc_tl
##          100.000          100.000
##          num_il_tl          num_op_rev_tl
##          100.000          100.000
##          num_rev_accts          num_rev_tl_bal_gt_0
##          100.000          100.000
##          num_sats          num_tl_120dpd_2m
##          100.000          100.000
##          num_tl_30dpd          num_tl_90g_dpd_24m
##          100.000          100.000
##          num_tl_op_past_12m          pct_tl_nvr_dlq
##          100.000          100.000
##          percent_bc_gt_75          pub_rec_bankruptcies
##          100.000          3.214
##          tax_liens          tot_hi_cred_lim
##          0.252          100.000
##          total_bal_ex_mort          total_bc_limit
##          100.000          100.000
##          total_il_high_credit_limit
##          100.000
```

Thus, it makes sense to delete features that have a missingness percentage above 30%.

```
data = data %>% select_if(~mean(is.na(.))<=0.3)
dim(data)
```

```
## [1] 42537    53
```

Data structures

```
str(data)
```

```
## tibble [42,537 x 53] (S3: tbl_df/tbl/data.frame)
## $ id          : num [1:42537] 1077501 1077430 1077175 1076863 1075358 ...
```

```

## $ member_id      : num [1:42537] 1296599 1314167 1313524 1277178 1311748 ...
## $ loan_amnt      : num [1:42537] 5000 2500 2400 10000 3000 ...
## $ funded_amnt    : num [1:42537] 5000 2500 2400 10000 3000 ...
## $ funded_amnt_inv : num [1:42537] 4975 2500 2400 10000 3000 ...
## $ term           : chr [1:42537] "36 months" "60 months" "36 months" "36 months" ...
## $ int_rate       : chr [1:42537] "10.65%" "15.27%" "15.96%" "13.49%" ...
## $ installment    : num [1:42537] 162.9 59.8 84.3 339.3 67.8 ...
## $ grade          : chr [1:42537] "B" "C" "C" "C" ...
## $ sub_grade      : chr [1:42537] "B2" "C4" "C5" "C1" ...
## $ emp_title      : chr [1:42537] NA "Ryder" NA "AIR RESOURCES BOARD" ...
## $ emp_length     : chr [1:42537] "10+ years" "< 1 year" "10+ years" "10+ years" ...
## $ home_ownership : chr [1:42537] "RENT" "RENT" "RENT" "RENT" ...
## $ annual_inc     : num [1:42537] 24000 30000 12252 49200 80000 ...
## $ verification_status : chr [1:42537] "Verified" "Source Verified" "Not Verified" "Source Ver
## $ issue_d        : chr [1:42537] "Dec-11" "Dec-11" "Dec-11" "Dec-11" ...
## $ loan_status    : chr [1:42537] "Fully Paid" "Charged Off" "Fully Paid" "Fully Paid" ..
## $ pymnt_plan     : chr [1:42537] "n" "n" "n" "n" ...
## $ url           : chr [1:42537] "https://lendingclub.com/browse/loanDetail.action?loan_
## $ purpose        : chr [1:42537] "credit_card" "car" "small_business" "other" ...
## $ title         : chr [1:42537] "Computer" "bike" "real estate business" "personel" ...
## $ zip_code       : chr [1:42537] "860xx" "309xx" "606xx" "917xx" ...
## $ addr_state     : chr [1:42537] "AZ" "GA" "IL" "CA" ...
## $ dti            : num [1:42537] 27.65 1 8.72 20 17.94 ...
## $ delinq_2yrs    : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ earliest_cr_line : chr [1:42537] "Jan-85" "Apr-99" "Nov-01" "Feb-96" ...
## $ inq_last_6mths  : num [1:42537] 1 5 2 1 0 3 1 2 2 0 ...
## $ open_acc       : num [1:42537] 3 3 2 10 15 9 7 4 11 2 ...
## $ pub_rec        : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ revol_bal      : num [1:42537] 13648 1687 2956 5598 27783 ...
## $ revol_util     : chr [1:42537] "83.70%" "9.40%" "98.50%" "21%" ...
## $ total_acc      : num [1:42537] 9 4 10 37 38 12 11 4 13 3 ...
## $ initial_list_status : logi [1:42537] FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ out_prncp      : num [1:42537] 0 0 0 0 335 ...
## $ out_prncp_inv   : num [1:42537] 0 0 0 0 335 ...
## $ total_pymnt     : num [1:42537] 5863 1009 3006 12232 3717 ...
## $ total_pymnt_inv : num [1:42537] 5834 1009 3006 12232 3717 ...
## $ total_rec_prncp : num [1:42537] 5000 456 2400 10000 2665 ...
## $ total_rec_int   : num [1:42537] 863 435 606 2215 1052 ...
## $ total_rec_late_fee : num [1:42537] 0 0 0 17 0 ...
## $ recoveries      : num [1:42537] 0 117 0 0 0 ...
## $ collection_recovery_fee : num [1:42537] 0 1.11 0 0 0 0 0 0 2.09 2.52 ...
## $ last_pymnt_d    : chr [1:42537] "Jan-15" "Apr-13" "Jun-14" "Jan-15" ...
## $ last_pymnt_amnt : num [1:42537] 171.6 119.7 649.9 357.5 67.8 ...
## $ last_credit_pull_d : chr [1:42537] "Jul-16" "Sep-13" "Jul-16" "Apr-16" ...
## $ collections_12_mths_ex_med : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ policy_code     : num [1:42537] 1 1 1 1 1 1 1 1 1 1 ...
## $ application_type : chr [1:42537] "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" "INDIVIDUAL" ...
## $ acc_now_delinq  : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ chargeoff_within_12_mths : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ delinq_amnt     : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ pub_rec_bankruptcies : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## $ tax_liens       : num [1:42537] 0 0 0 0 0 0 0 0 0 0 ...
## - attr(*, "problems")= tibble [1 x 5] (S3: tbl_df/tbl/data.frame)
## ..$ row      : int 39788

```

```
## ..$ col      : chr "id"
## ..$ expected: chr "a double"
## ..$ actual   : chr "Loans that do not meet the credit policy"
## ..$ file     : chr "'Data/LoanStats3a.csv'"
```

```
sapply(data,function(x){ length(unique(x)) })
```

```
##          id          member_id
##      42536          42536
##      loan_amnt      funded_amnt
##      899           1052
##      funded_amnt_inv      term
##      9242           3
##      int_rate      installment
##      395           16460
##      grade      sub_grade
##      8           36
##      emp_title      emp_length
##      30449          13
##      home_ownership      annual_inc
##      6           5598
##      verification_status      issue_d
##      4           56
##      loan_status      pymnt_plan
##      10           3
##      url      purpose
##      42536          15
##      title      zip_code
##      20965          838
##      addr_state      dti
##      51           2895
##      delinq_2yrs      earliest_cr_line
##      13           531
##      inq_last_6mths      open_acc
##      29           45
##      pub_rec      revol_bal
##      7           22710
##      revol_util      total_acc
##      1120          84
##      initial_list_status      out_prncp
##      2           831
##      out_prncp_inv      total_pymnt
##      833           40610
##      total_pymnt_inv      total_rec_prncp
##      40118          8471
##      total_rec_int      total_rec_late_fee
##      37582          1562
##      recoveries      collection_recovery_fee
##      4518           2843
##      last_pymnt_d      last_pymnt_amnt
##      106           37080
##      last_credit_pull_d      collections_12_mths_ex_med
##      111           2
##      policy_code      application_type
```

```
##                2                2
##      acc_now_delinq  chargeoff_within_12_mths
##                3                2
##      delinq_amnt      pub_rec_bankruptcies
##                4                4
##      tax_liens
##                3
```

There are some qualitative features that need to be converted to factors.

Missing Values (Part II)