Lending Club Loan Data

Analyze Lending Club's issued loans

Tanawit Pattanaveerangkoon
Setthavuth Tsoi
Peerasin Nilsitikul
Adinun Rungratmaneemas
Anupon Khunarongnunthakul

Agenda

- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- Modeling
- Evaluation
- Visualization of results

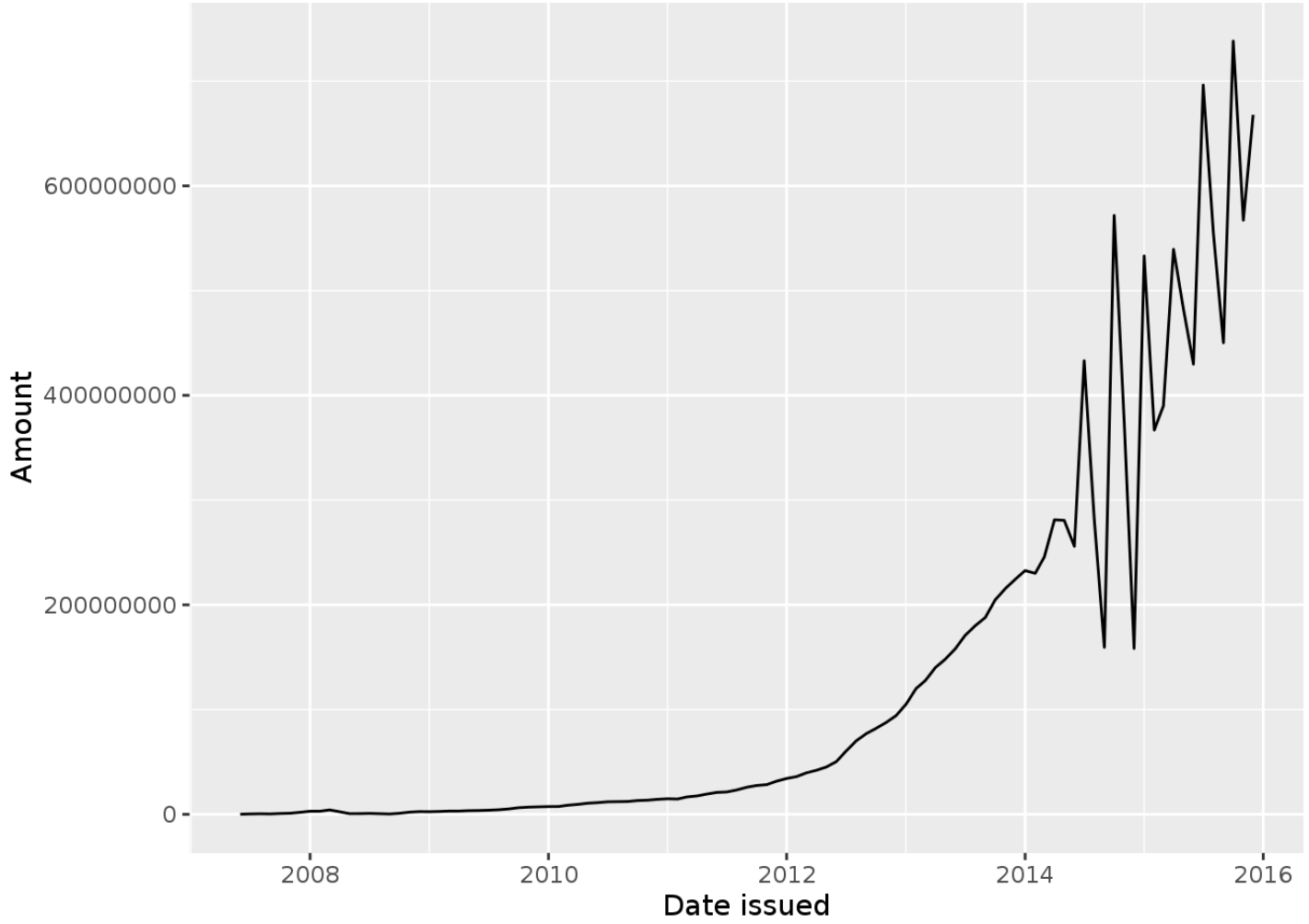
Data Description

Data Description

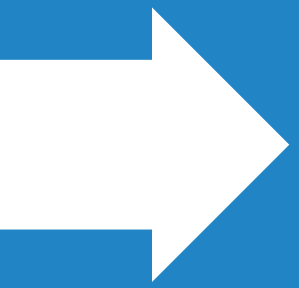
- loan data for all loans issued through the 2007-2015
- including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information.
- 887,383 observations
- 75 variables

Data source URL: https://www.kaggle.com/wendykan/lending-club-loan-data

The amount of loan by date



- **Data description**
- Objective
- Data selection
- Data cleaning
- Data transformation
- Modeling
- Evaluation
- Visualization of results



Objective

Objective

- to predict which customers' loan are likely to be bad loan
- to investigate which factors have an effect on classify as a bad loan

Data Selection

Loan_status variable is target class

- Current status of the loan

  - Fully Paid
  - Current

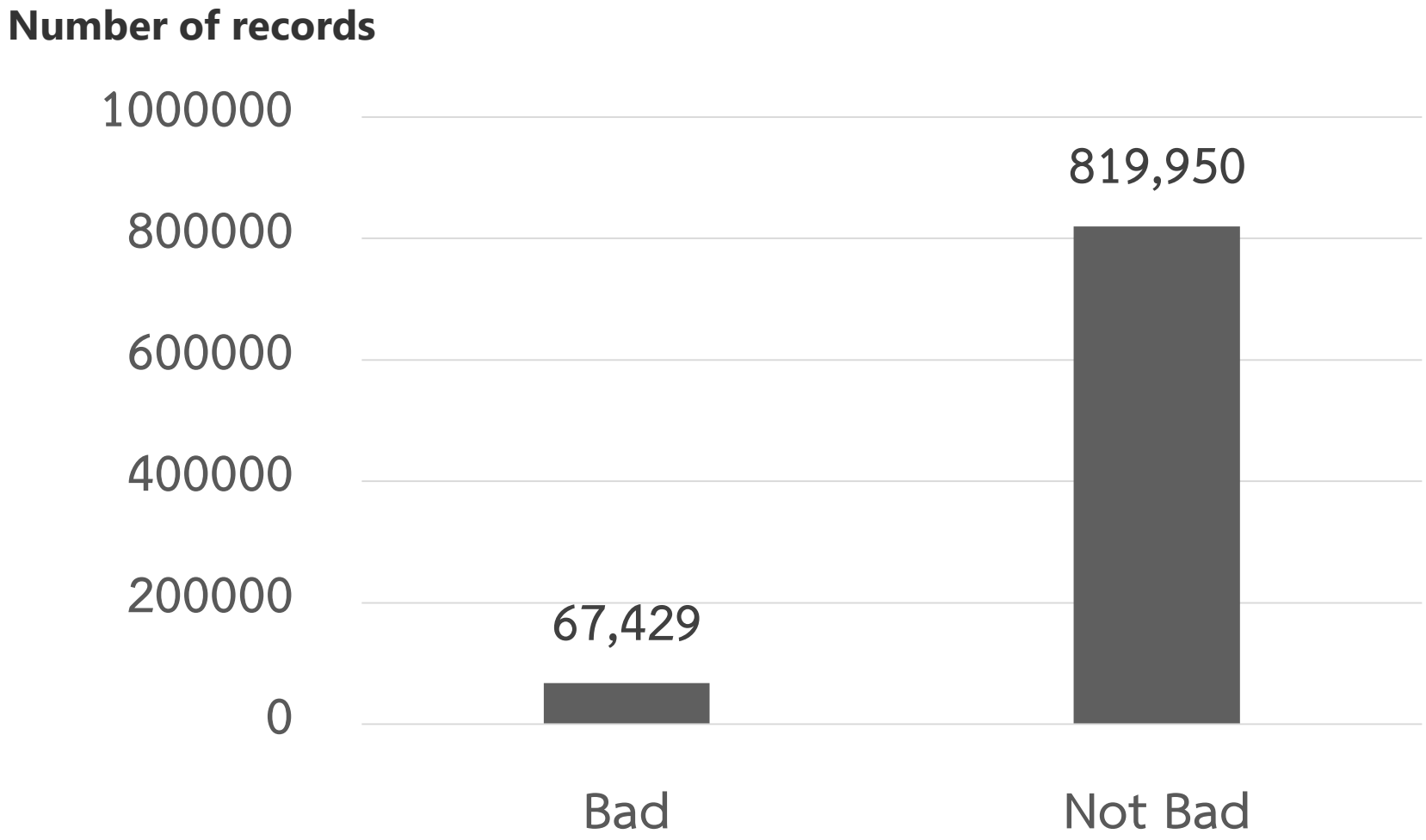
  - Charged Off
  - Default
  - Late (31-120 days)
  - In Grace Period
  - Late (16-30 days)

Loan_status variable is target class

- Current status of the loan

Good loan	Bad loan
Fully Paid	Charged Off
Current	Default
	Late (31-120 days)
	In Grace Period
	Late (16-30 days)

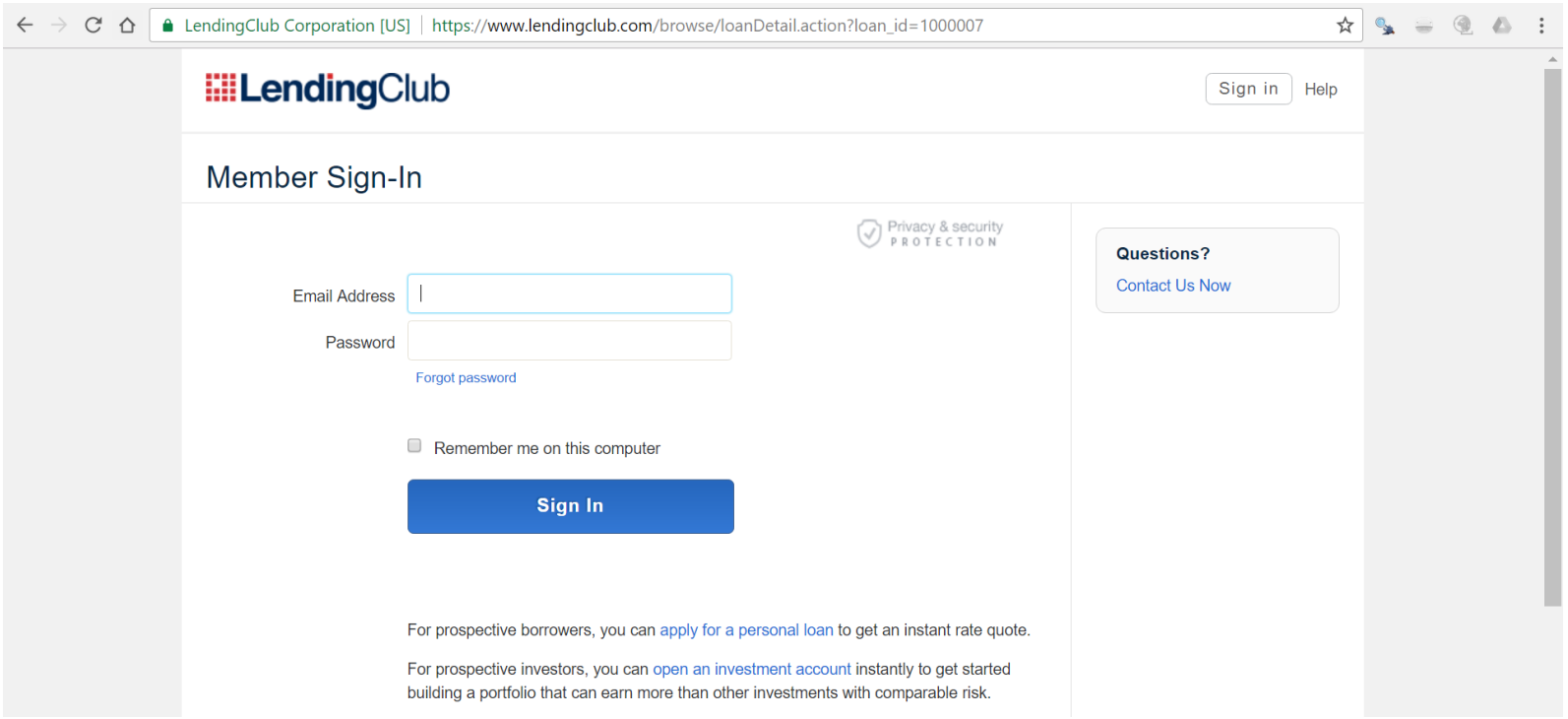
The number of bad/not bad loan



- Data description
- Objective
- **Data selection**
- Data cleaning
- Data transformation
- Modeling
- Evaluation
- Visualization of results

Remove url variables

- URL for the LC page with listing data.

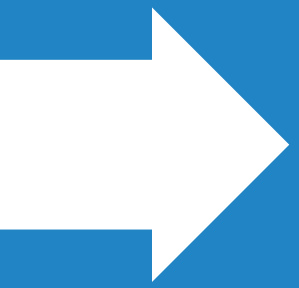


- Data description
- Objective
- **Data selection**
- Data cleaning
- Data transformation
- Modeling
- Evaluation
- Visualization of results

Remove desc variables

- loan description provided by the borrower

```
                                                                desc
                                                                  :761350
                                                                  :   246
Debt Consolidation                                                :    13
  Borrower added on 03/17/14 > Debt consolidation<br>:            :    11
  Borrower added on 03/10/14 > Debt consolidation<br>:            :    10
  Borrower added on 02/19/14 > Debt consolidation<br>:            :     9
(Other)                                                           :125740
```



Data Cleaning

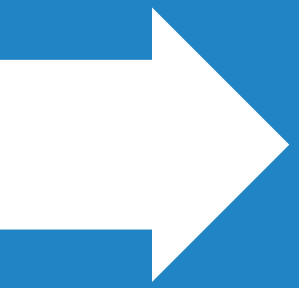
Remove variables where more than 20% of the observations are missing values

- remove 19 variables
  - mths_since_last_delinq
  - mths_since_last_record
  - mths_since_last_major_derog
  - annual_inc_joint
  - dti_joint
  - open_acc_6m
  - open_il_6m
  - open_il_12m
  - open_il_24m
  - mths_since_rcnt_il
  - total_bal_il
  - il_util
  - open_rv_12m
  - open_rv_24m
  - max_bal_bc
  - all_util
  - inq_fi
  - total_cu_tl
  - inq_last_12m

Remove variables where more than 20% of the observations are missing values

```
 open_il_24m     mths_since_rcnt_il  total_bal_il       il_util
Min.   : 0.0     Min.   :  0.0      Min.   :     0   Min.   :  0.0
1st Qu.: 0.0     1st Qu.:  6.0      1st Qu.: 10252   1st Qu.: 58.6
Median : 1.0     Median : 12.0      Median : 24685   Median : 74.9
Mean   : 1.7     Mean   : 20.9      Mean   : 36553   Mean   : 71.5
3rd Qu.: 2.0     3rd Qu.: 23.0      3rd Qu.: 47858   3rd Qu.: 87.6
Max.   :19.0     Max.   :363.0      Max.   :878459   Max.   :223.3
NA's   :866007   NA's   :866569     NA's   :866007   NA's   :868762
```

- Data description
- Objective
- Data selection
- **Data cleaning**
- Data transformation
- Modeling
- Evaluation
- Visualization of results



Data Transformation

Loan_status variable is target class

- Current status of the loan

- Fully Paid
- Current

loan_status: 0

Good loan

- Charged Off
- Default
- Late (31-120 days)
- In Grace Period
- Late (16-30 days)

loan_status: 1

Bad loan

issue_d, earliest_cr_line

- create date variables that R recognizes as dates

Jul-2015		2015-7-15
Dec-2015		2015-7-15
Oct-2014		2015-7-15
Factor		Date

term.months, term

- identify loans that have already come to term
- remove the "term" variable because it is redundant with the term.months variable

- Data description
- Objective
- Data selection
- Data cleaning
- **Data transformation**
- Modeling
- Evaluation
- Visualization of results

emp_length

- convert character employment length to numeric variable

Factor		Numeric
1 year		1
2 years	→	2
3 years		3

- Data description
- Objective
- Data selection
- Data cleaning
- **Data transformation**
- Modeling
- Evaluation
- Visualization of results

grade

- convert character to ordinal variable

A > B > C > D > E > F

- Data description
- Objective
- Data selection
- Data cleaning
- **Data transformation**
- Modeling
- Evaluation
- Visualization of results

Generate paid.back variable

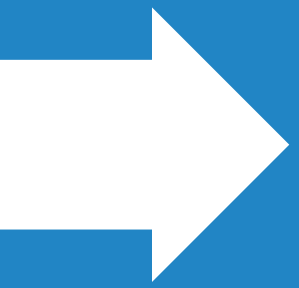
- Calculate the percentage of loan paid back

$$\text{paid.back} = \frac{\text{funded\_amnt - out\_prncp}}{\text{funded\_amnt}}$$

- remove NA values

Remove factor variables with too many levels

- Random Forest Algorithm in R cannot handle variables with more than 32 levels
- remove 5 variables
  - sub_grade
  - emp_title
  - title
  - zip_code
  - addr_state



Modeling

Remaining input variables

- 41 variables

- loan_amnt
- funded_amnt
- funded_amnt_inv
- int_rate
- installment
- grade
- emp_length
- home_ownership
- annual_inc
- verification_status
- issue_d
- pymnt_plan
- purpose
- dti
- delinq_2yrs
- earliest_cr_line
- inq_last_6mths
- open_acc
- pub_rec
- revol_bal
- revol_util
- total_acc
- initial_list_status
- out_prncp
- out_prncp_inv
- total_pymnt
- total_pymnt_inv
- total_rec_prncp
- total_rec_int
- total_rec_late_fee
- recoveries
- collection_recovery_fee
- last_pymnt_amnt
- collections_12_mths_ex_med
- application_type
- acc_now_delinq
- tot_coll_amt
- tot_cur_bal
- total_rev_hi_lim
- term.months
- paid.back

Split data into 2 parts – training and test set

Training set

107,933 records (80%)
for train model

Bad Loan: 53,851 records
Good Loan: 54,082 records

Test set

26,925 records (20%)
for evaluate model

Total: 134,858 records

- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- **Modeling**
- Evaluation
- Visualization of results

Use Random Forest algorithm to generate model

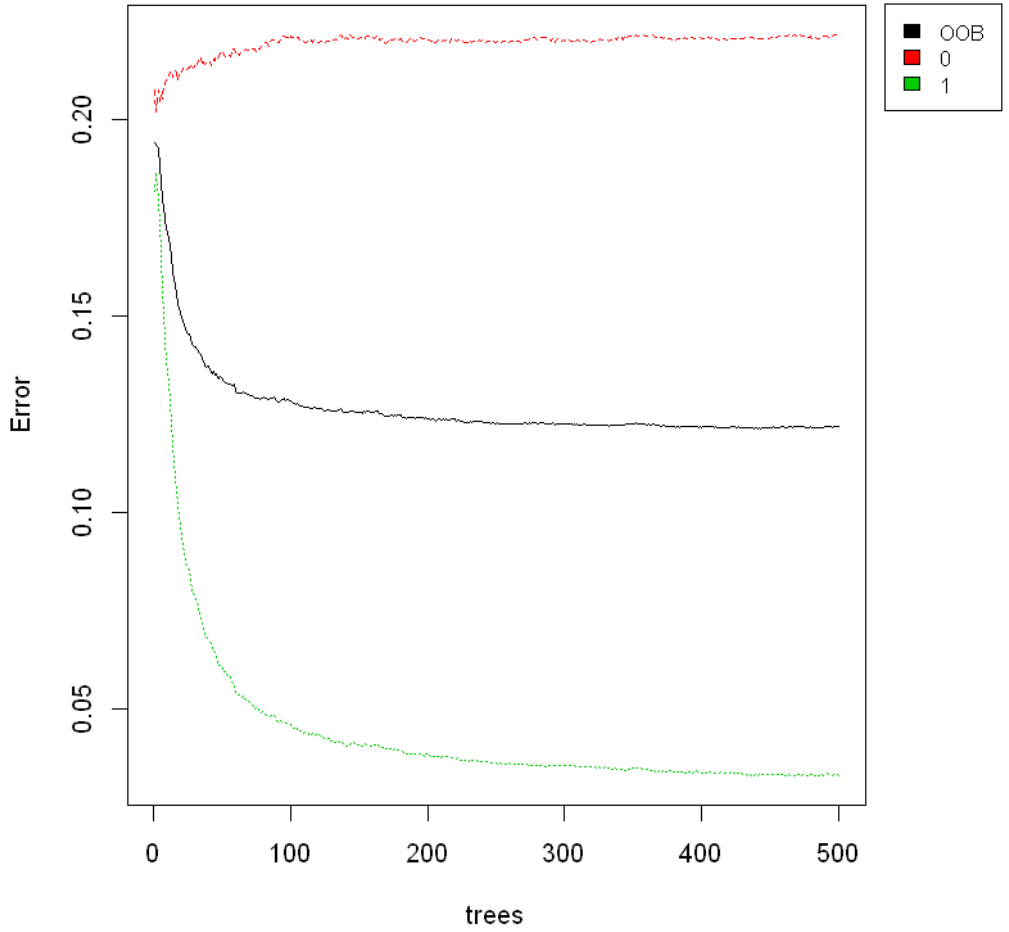
Training set
107,933 records (80%)

Random Forest algorithm

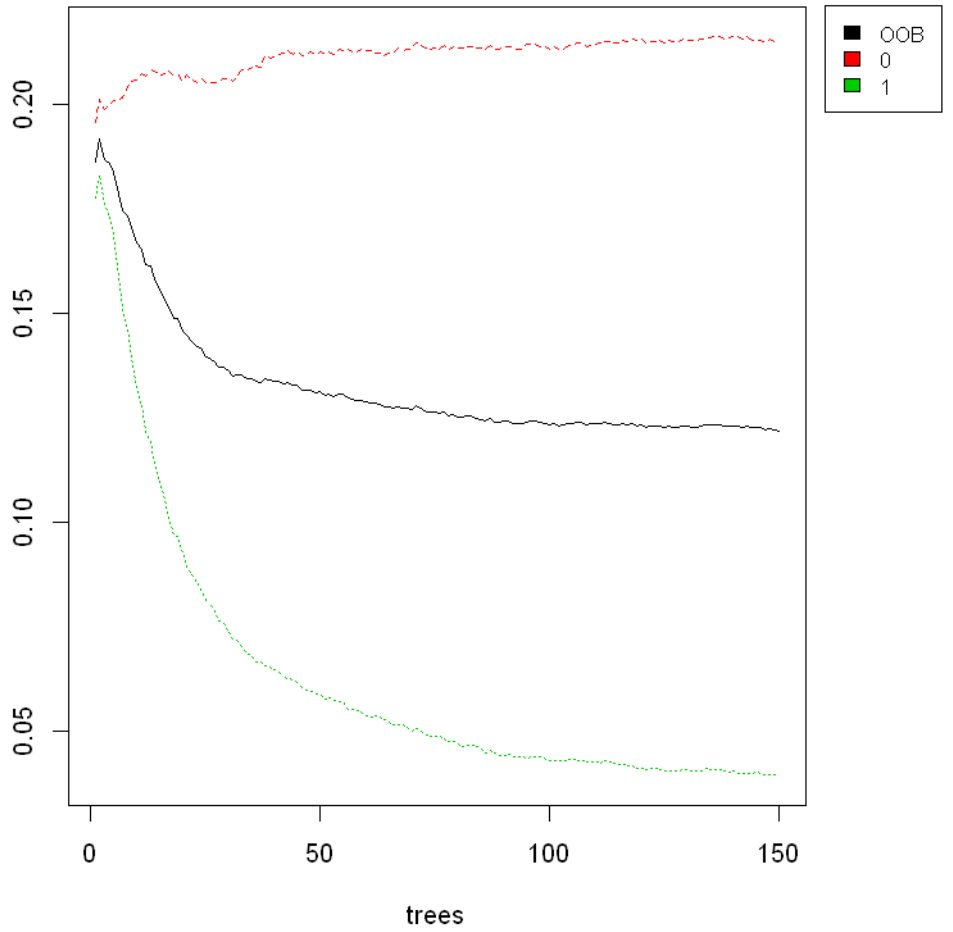
Model

- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- **Modeling**
- Evaluation
- Visualization of results

Plot model error by the number of trees



ntree = 500



ntree = 150

- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- **Modeling**
- Evaluation
- Visualization of results

Evaluation

Area Under Cruve (AUC)

0.8703

Confusion Metrix

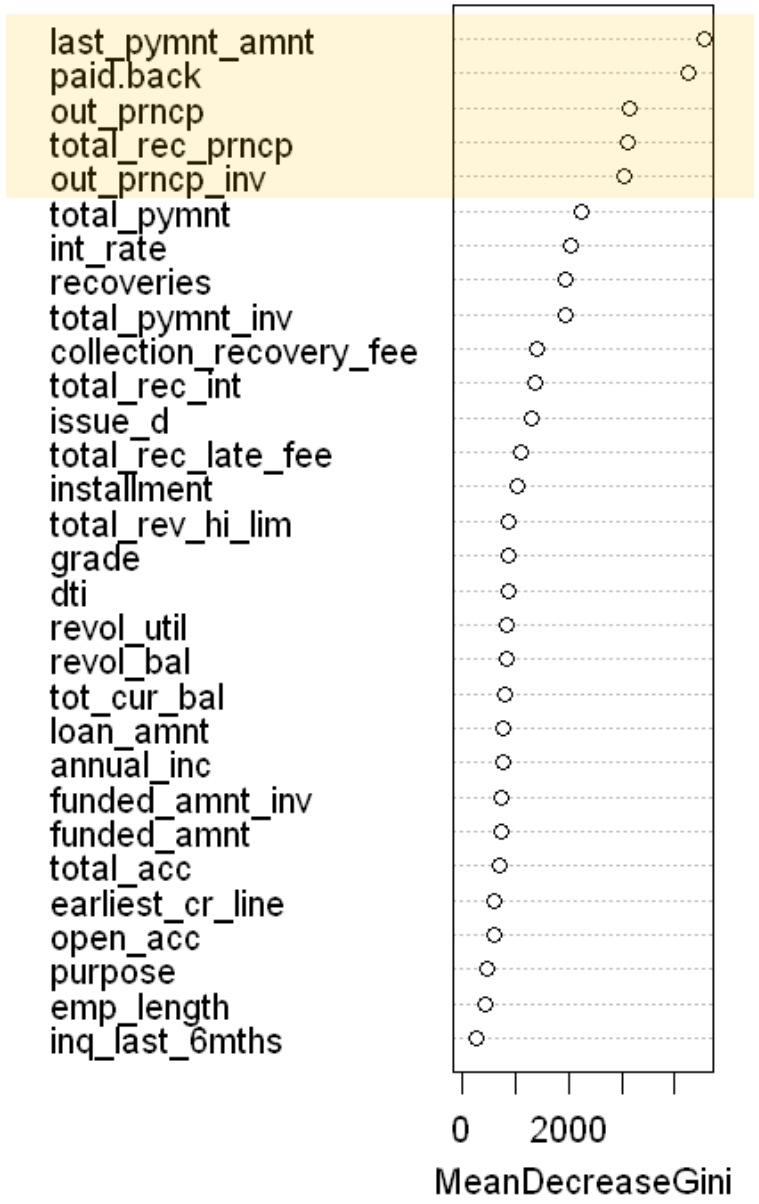
```
Confusion Matrix and Statistics

          Reference
Prediction     1     2
         1  8283   402
         2  2404 11289

               Accuracy : 0.8746
                 95% CI : (0.8702, 0.8789)
    No Information Rate : 0.5224
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7467
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7751
            Specificity : 0.9656
         Pos Pred Value : 0.9537
         Neg Pred Value : 0.8244
             Prevalence : 0.4776
         Detection Rate : 0.3701
   Detection Prevalence : 0.3881
      Balanced Accuracy : 0.8703

       'Positive' Class : 1
```

- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- Modeling
- **Evaluation**
- Visualization of results

Visualization of result

Factor important

- Last_pymnt_amnt is the most important factor for classify loan quality



- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- Modeling
- Evaluation
- **Visualization of results**

last_pymnt_amnt

- last total payment amount received

paid.back

- the percentage of loan paid back

out_prncp

- remaining outstanding principal for total amount funded

total_rec_int

- interest received to date

out_prncp_inv

- remaining outstanding principal for portion of total amount funded by investors

- Data description
- Objective
- Data selection
- Data cleaning
- Data transformation
- Modeling
- Evaluation
- **Visualization of results**

Instructor Feedback

Version 1

Instructor Feedback (Version 1)

- Only 2 levels of target class
  - Fully Paid
  - Charged Off

Split data into 2 parts – training and test set

Training set

107,933 records (80%)
for train model

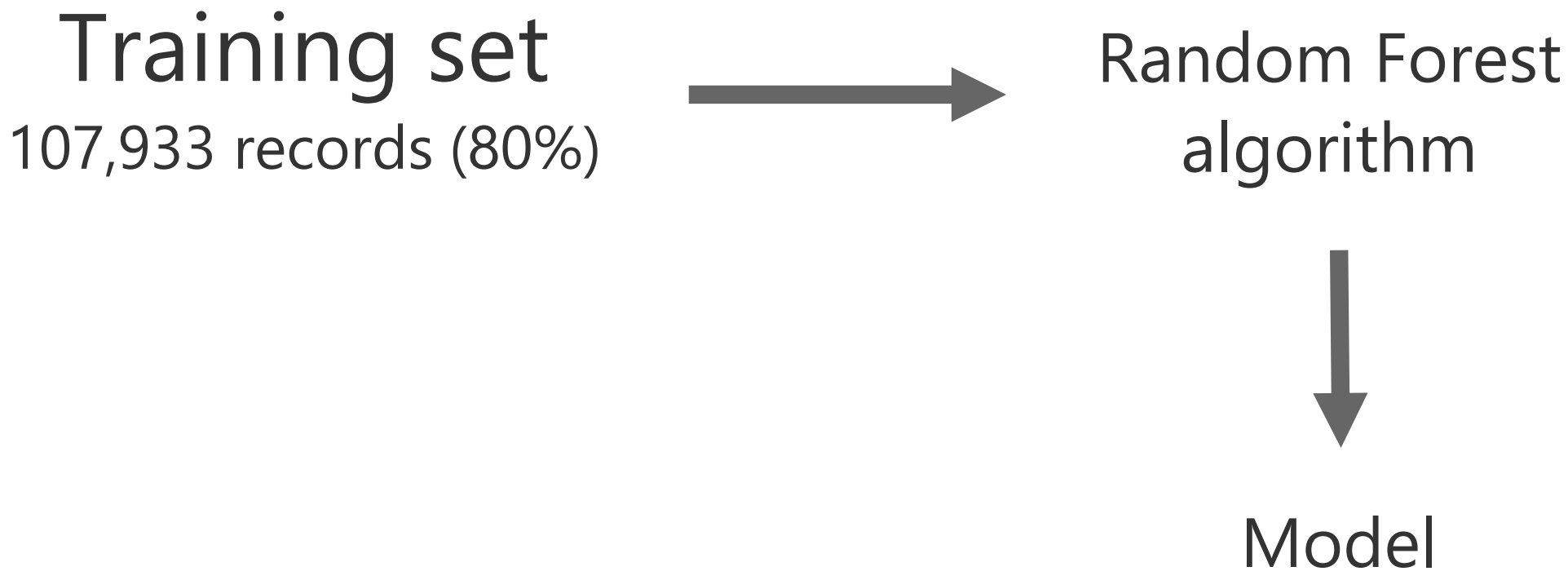
Bad Loan: 53,851 records
Good Loan: 54,082 records

Test set

26,925 records (20%)
for evaluate model

Total: 134,858 records

Use Random Forest algorithm to generate model



Area Under Cruve (AUC)

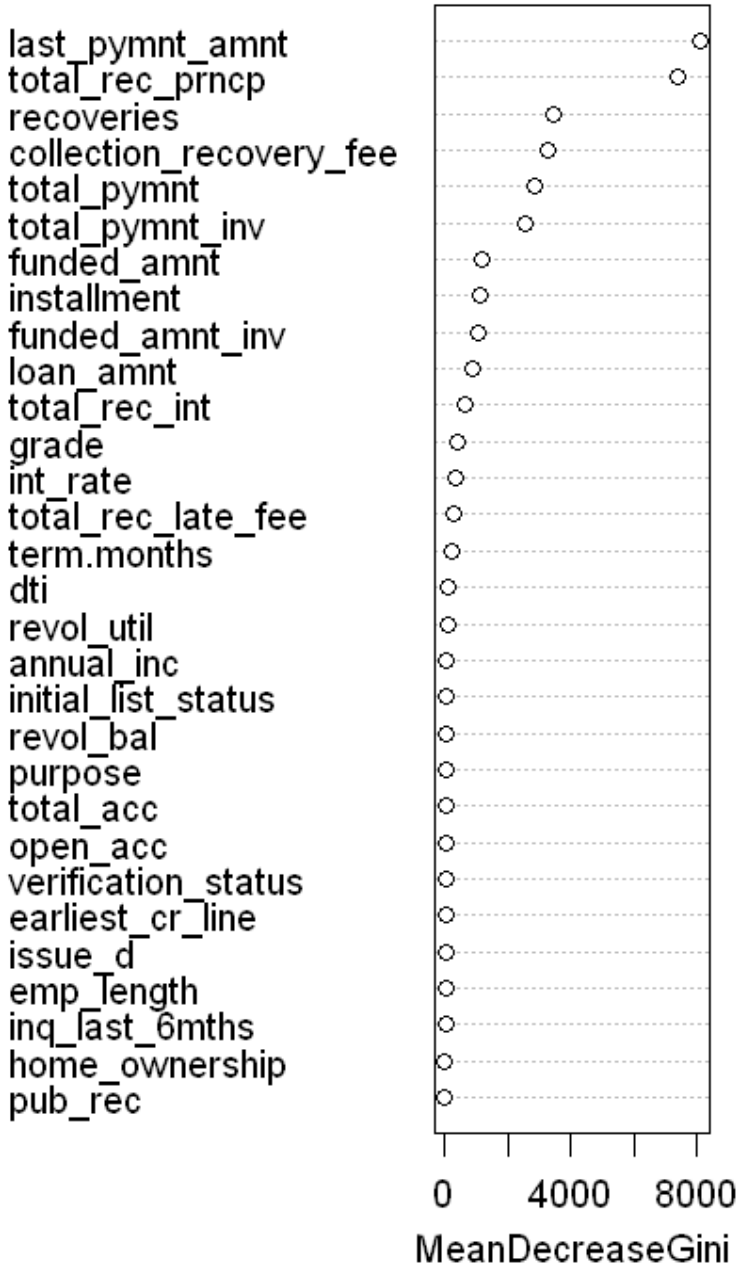
0.9969

Confusion Metrix

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 8587    1
         2   53 8693

               Accuracy : 0.9969
                 95% CI : (0.9959, 0.9977)
    No Information Rate : 0.5016
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9938
 Mcnemar's Test P-Value : 3.915e-12

            Sensitivity : 0.9939
            Specificity : 0.9999
         Pos Pred Value : 0.9999
         Neg Pred Value : 0.9939
             Prevalence : 0.4984
         Detection Rate : 0.4954
   Detection Prevalence : 0.4954
      Balanced Accuracy : 0.9969

       'Positive' Class : 1
```

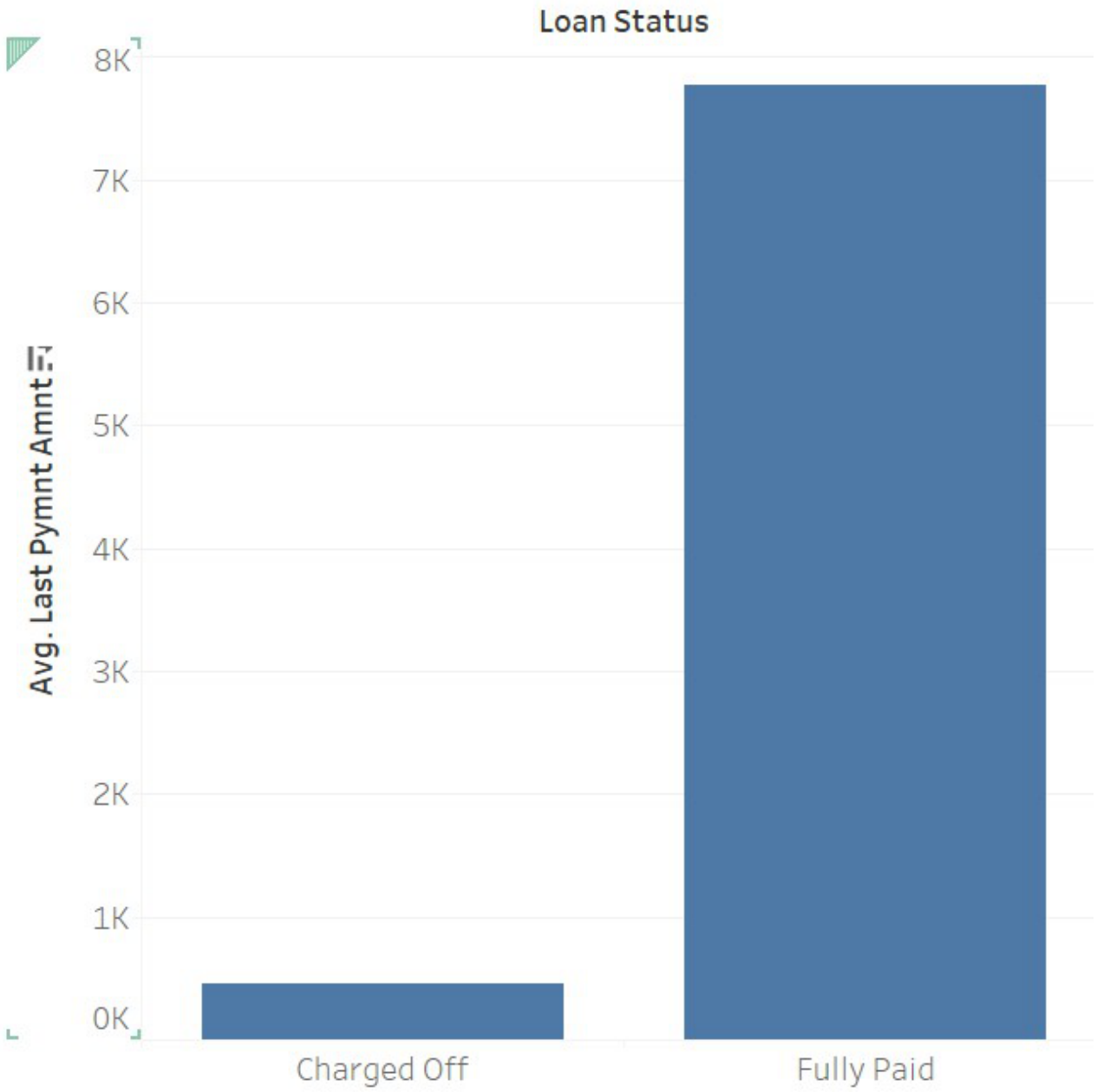
Factor Important

- last_pymnt and total_rec_prncp are the most important factor for classify loan quality



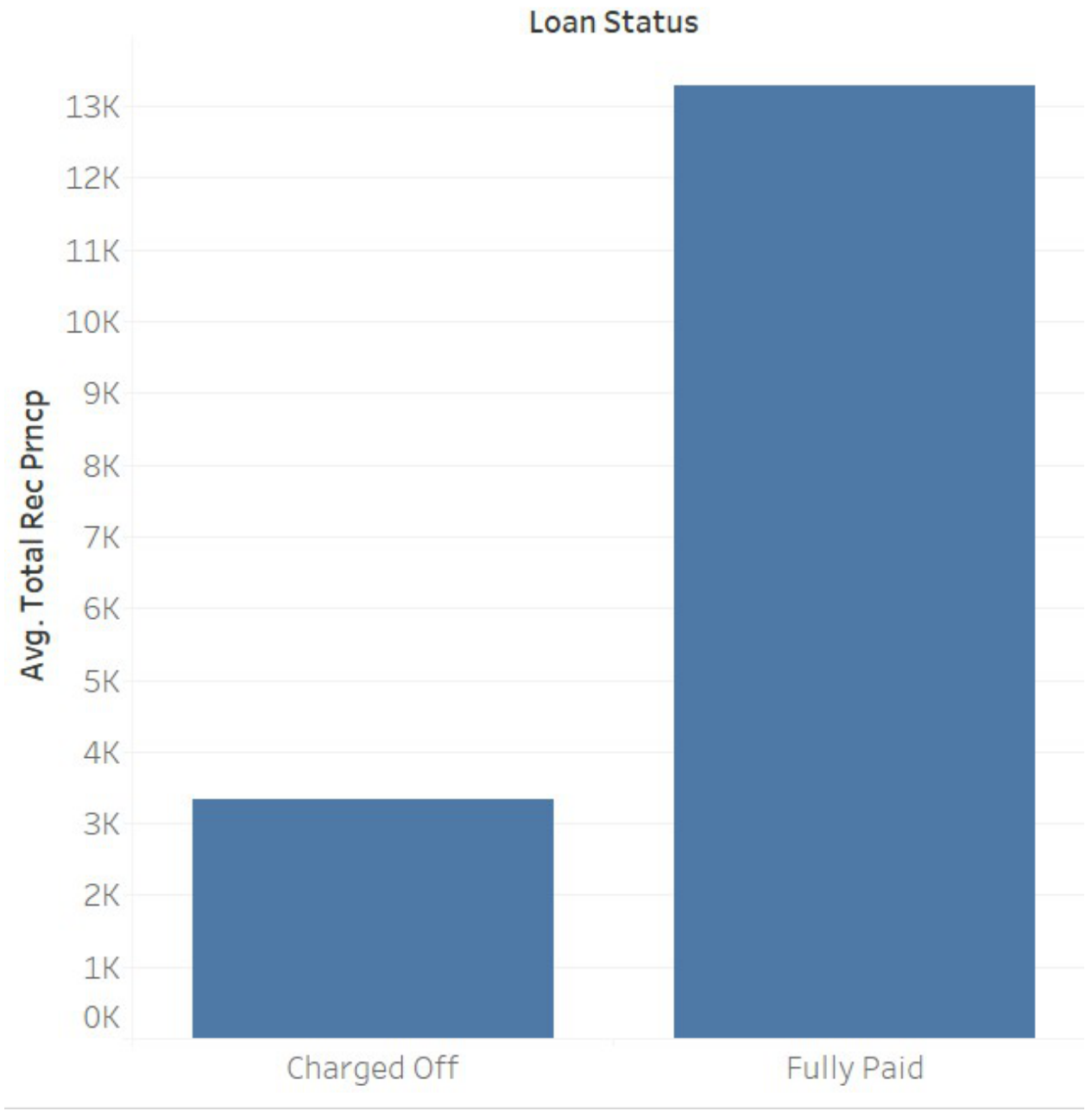
last_pymnt_amnt

- last total payment amount received



total_rec_prncp

- principal received to date



Instructor Feedback

Version 2

Instructor Feedback (Version 2)

- to predict which customers' loan are likely to be bad loan before we loan

Input variables

- 13 variables
- We known all of these variables value before we loan

- annual_inc
- emp_length
- home_ownership
- dti
- delinq_2yrs
- collections_12_mths_ex_med
- inq_last_6mths
- purpose
- application_type
- inq_last_6mths
- loan_amnt
- funded_amnt
- verification_status

annual_inc

- The self-reported annual income provided by the borrower during registration.

emp_length

- Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.

home_ownership

- The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.

dti

- A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

Variable Description

delinq_2yrs

- The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years

collections_12_mths_ex_med

- Number of collections in 12 months excluding medical collections

inq_last_6mths

- The number of inquiries in past 6 months (excluding auto and mortgage inquiries)

purpose

- A category provided by the borrower for the loan request.

Variable Description

verification_status

- Indicates if the co-borrowers' joint income was verified by LC, not verified, or if the income source was verified

application_type

- Indicates whether the loan is an individual application or a joint application with two co-borrowers

inq_last_6mths

- The number of inquiries in past 6 months (excluding auto and mortgage inquiries)

loan_amnt

- The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.

funded_amnt

- The total amount committed to that loan at that point in time.

Split data into 2 parts – training and test set

Training set

107,933 records (80%)
for train model

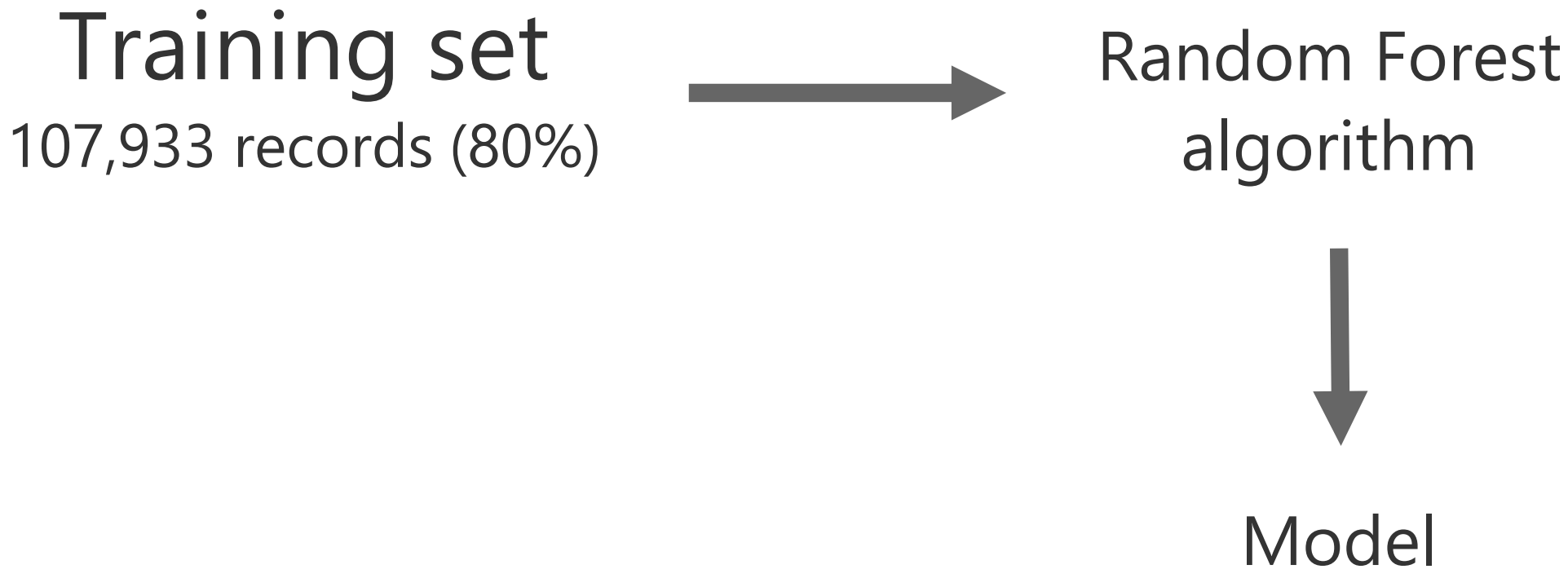
Bad Loan: 53,851 records
Good Loan: 54,082 records

Test set

26,925 records (20%)
for evaluate model

Total: 134,858 records

Use Random Forest algorithm to generate model



Area Under Cruve (AUC)

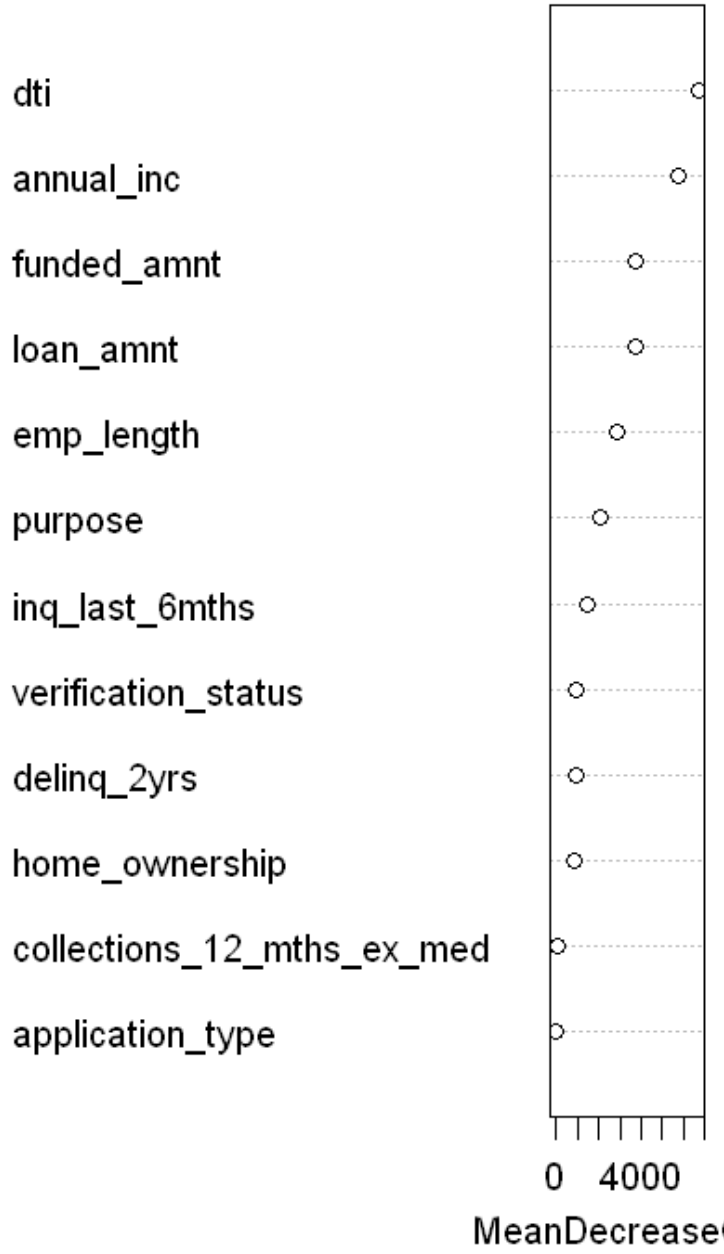
0.5926

Confusion Metrix

```
Confusion Matrix and Statistics

          Reference
Prediction    1    2
         1 4774 3183
         2 3877 5500

               Accuracy : 0.5927
                 95% CI : (0.5853, 0.6)
    No Information Rate : 0.5009
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1853
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5518
            Specificity : 0.6334
         Pos Pred Value : 0.6000
         Neg Pred Value : 0.5865
             Prevalence : 0.4991
         Detection Rate : 0.2754
   Detection Prevalence : 0.4590
      Balanced Accuracy : 0.5926

       'Positive' Class : 1
```

Factor Important

- dti and annual_inc are the most important factor for classify loan quality



Thanks!

Any questions?