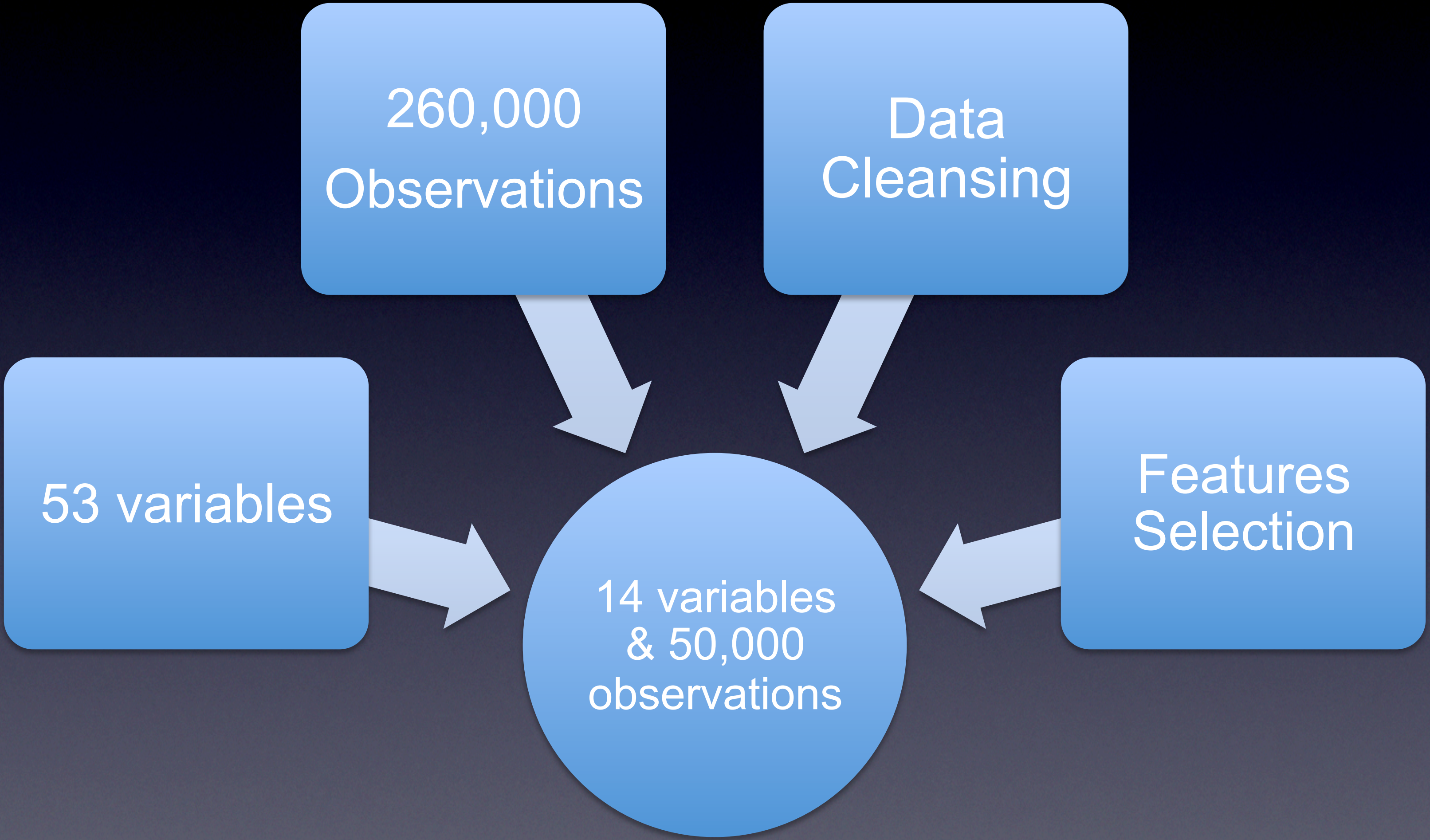
Will A Loaner Default?

Objectives

- Lending Club Peer-to-peer Loans
- High Risk Predict Default
- Overall Risk Control
- Secure Long Term Growth

Data Source



260,000 Observations

Data Cleansing

53 variables

Features Selection

14 variables & 50,000 observations

Variable Description

Variable Name	Description	Note
loan_status	Current status of the loan	1: Default and Chargeoff; 0: Fully Paid
annual_inc	Annual income	
dti	Debt to income ratio	[0,1]
int_rate	Interest rate on the loan	[0,1]
loan_amnt	Loan amount	$
pub_rec	Number of derogatory public records	
emp_length	Employment length in years	[0,10]; <1 = 0 and >10 = 10
grade	LC assigned loan grade	A-G in alphabetical order (A represents highest rating)
purpose	Purpose of the loan	1: Debt Consolidation; 0: Others
term	The number of payments on the loan	Either 36 months or 60 months
home_ownership	Home Ownership Status	Rent, Own or Mortgage

Exploratory Data Analysis

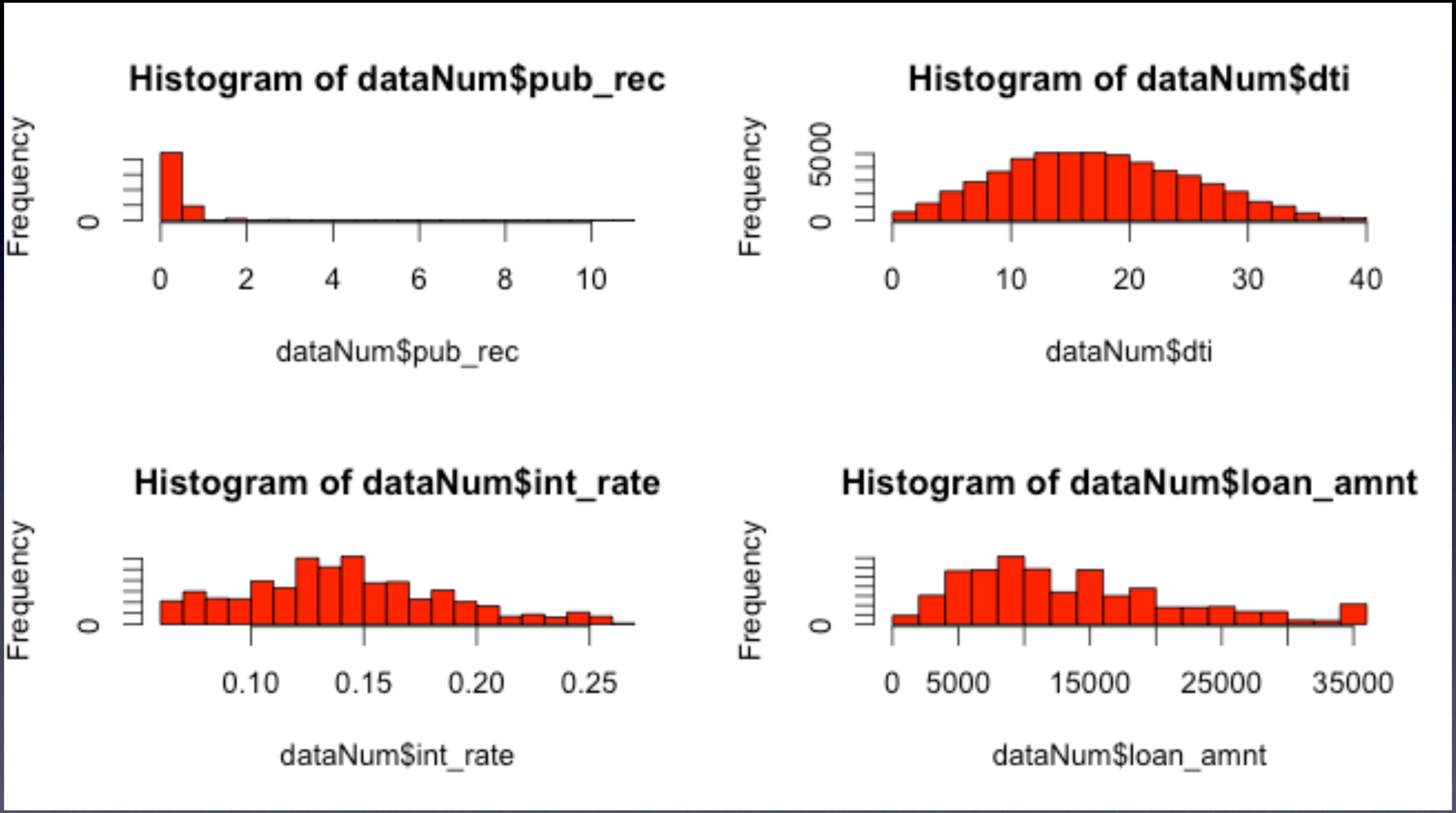
Step 1: Response Variable (Y) -- Loan Status

	Y=1 Default and Charge-off	Y=0 Fully Paid	Total
Number	10,494	44,629	55,123
%	19%	81%	100%

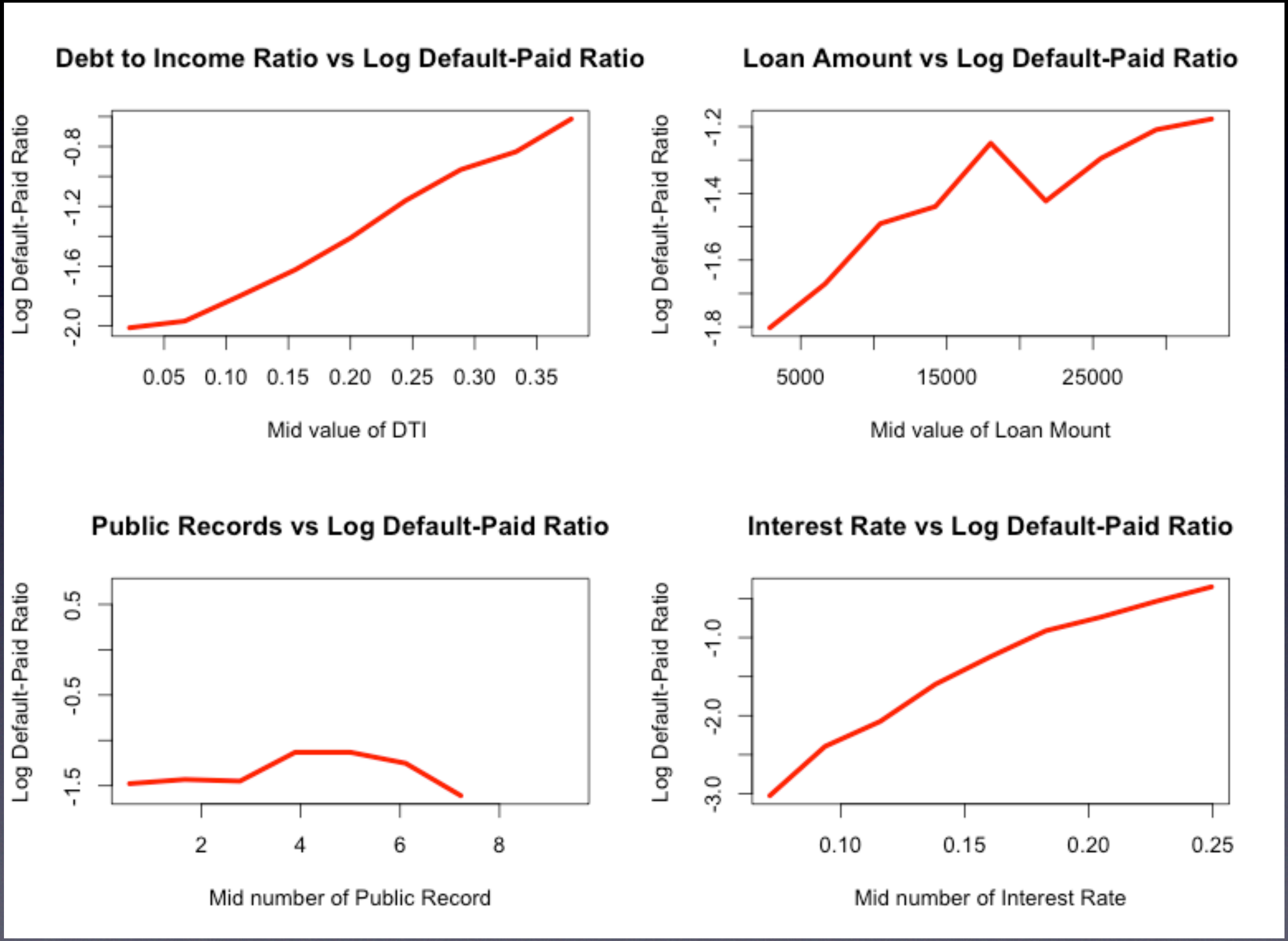
Step 2: Response Variable and Numeric Variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Annual Income($)	3,000	47,000	65,000	74,710	90,000	4,900,000
Debt to Income Ratio	0	11.4	16.86	17.26	22.86	39.99
Interest Rate	0.06	0.11	0.14	0.14	0.17	0.26
Loan Amount($)	1,000	8,000	12,000	14,090	19,200	35,000
Public Record	0	0	0	0.24	0	11

Distribution of Numeric Variables



Relationships between Y and numeric variables



Debt to Income Ratio vs Log Default-Paid Ratio

Loan Amount vs Log Default-Paid Ratio

Public Records vs Log Default-Paid Ratio

Interest Rate vs Log Default-Paid Ratio

Step 3: Response Variable and Category Variables

Loan Status and Loan Grade

Loan Status and Term

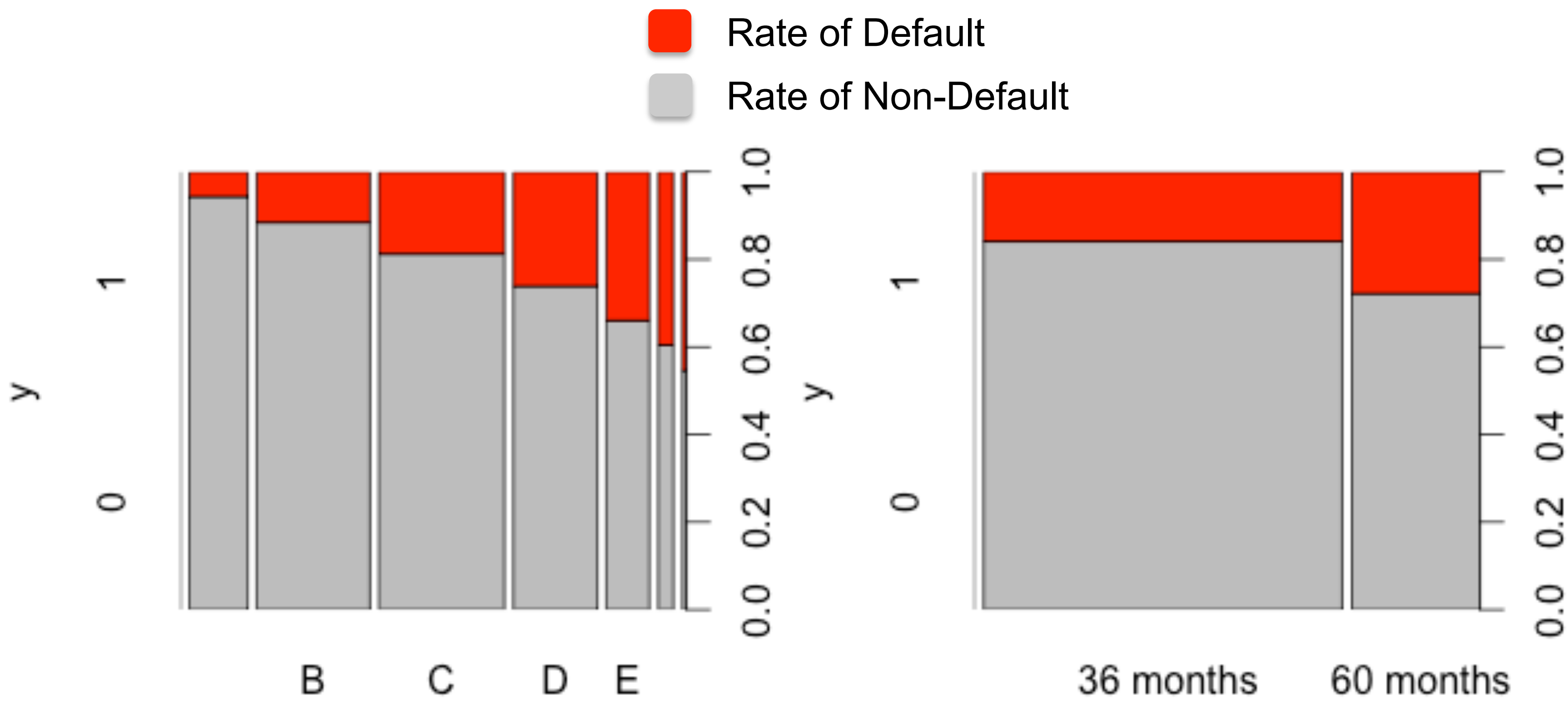


Figure 2a.

Figure 2b.

Loan Status and
House Ownership / Employment Years / Purpose of Loan:
No strong relationships

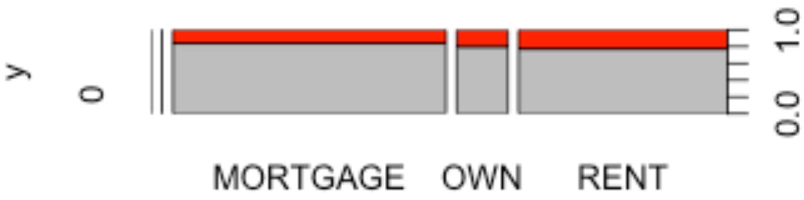


Figure 3a.

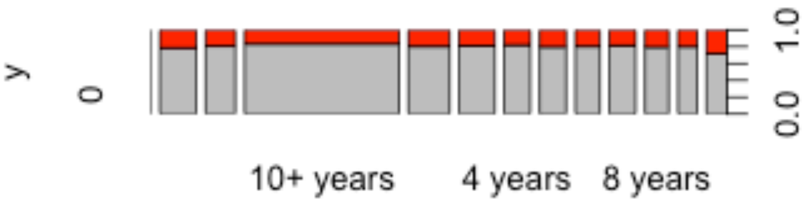


Figure 3b.

🟥 Rate of Default

⬜ Rate of Non-Default

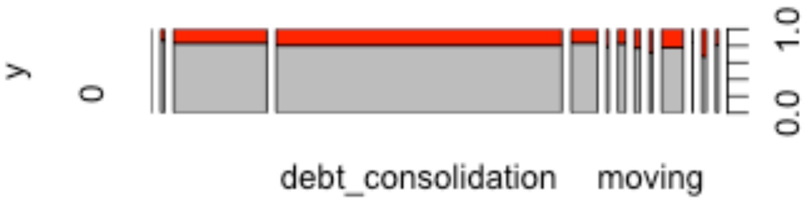


Figure 3c.

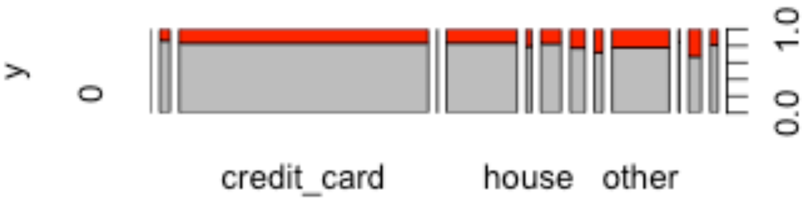


Figure 3d.

Logistic Regression

Why Logistic Regression?

- Binary response
- Interpretation
- Flexible thresholds

Randomly select 80% of the data set as training data.

Use variables according to EDA.

- Usually:

probability > 0.5 → default
probability <= 0.5 → non-default

- Effect:

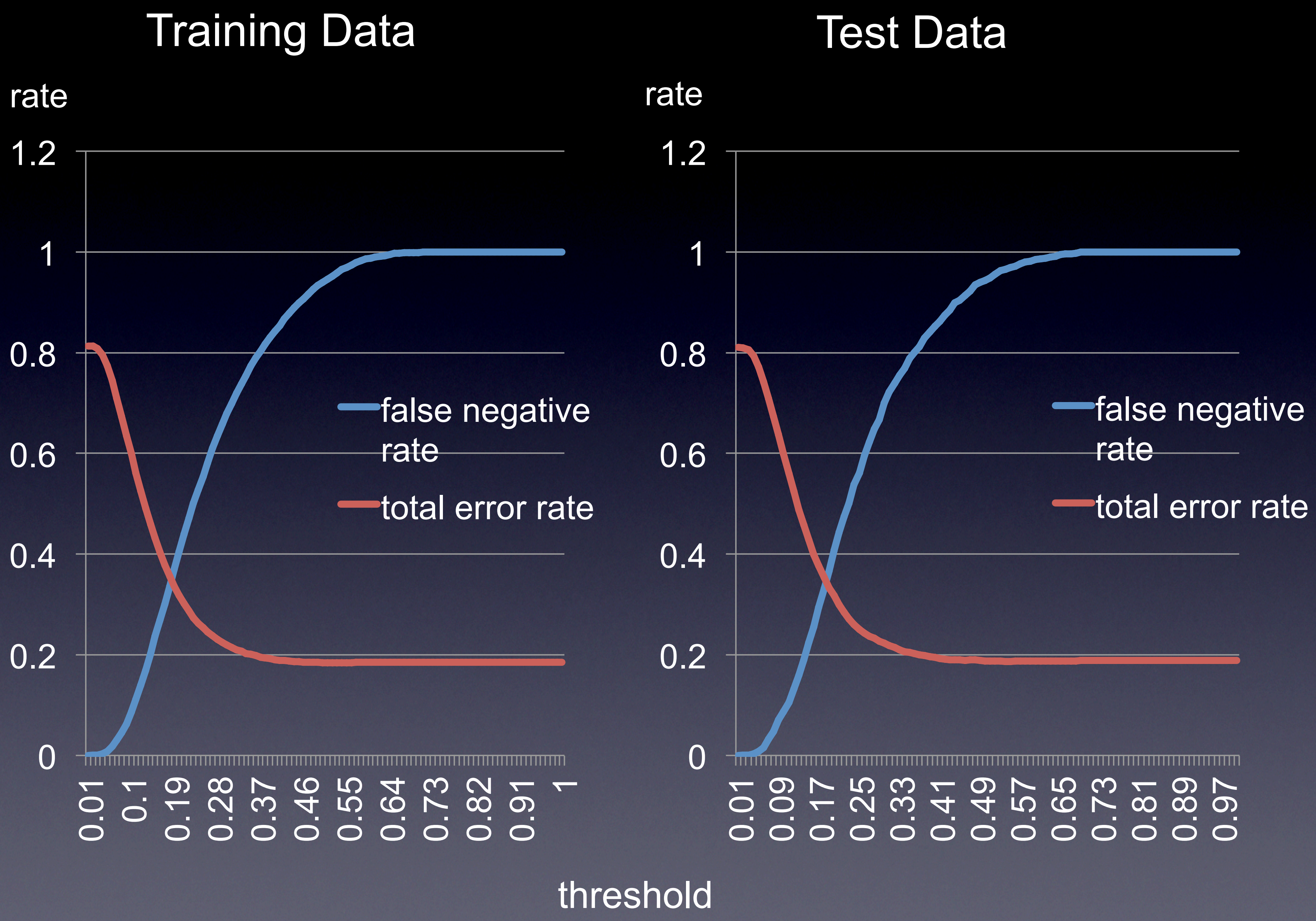
Low total error rate, high false negative rate

- Improvement:

Decrease false negative rate

- Method:

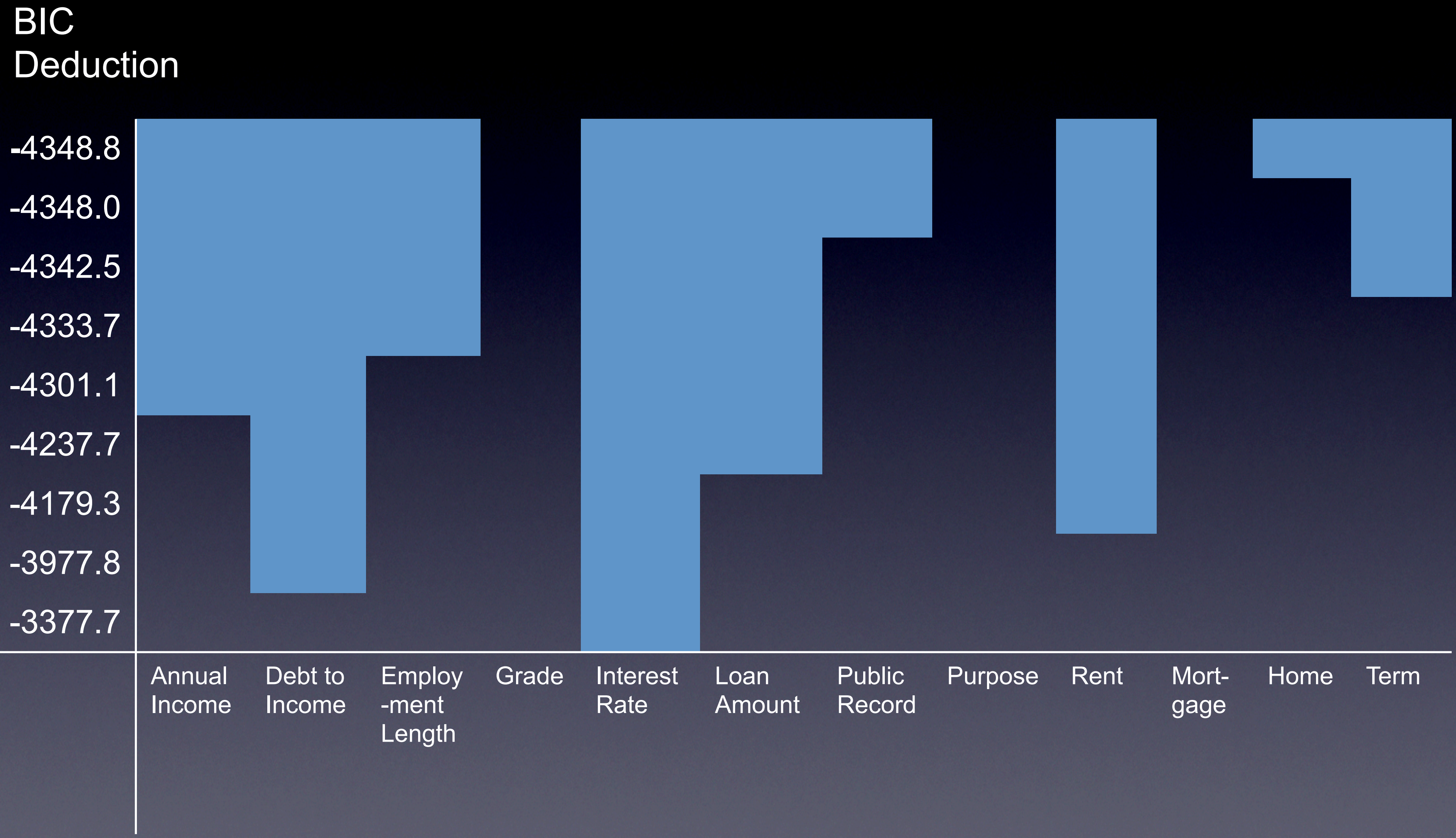
Change thresholds


Training Data
rate
1.2
1
0.8
0.6
0.4
0.2
0
false negative rate
total error rate
0.01
0.1
0.19
0.28
0.37
0.46
0.55
0.64
0.73
0.82
0.91
1
Test Data
rate
1.2
1
0.8
0.6
0.4
0.2
0
false negative rate
total error rate
0.01
0.09
0.17
0.25
0.33
0.41
0.49
0.57
0.65
0.73
0.81
0.89
0.97
threshold

Model Selection (Feature Importance)

Model Size	Annaul Income	Debt to Income	Employment Length	Grade	Interest Rate	Loan Amount	Public Record	Purpose	Rent	Mortgage	Home	Term
1					*							
2		*			*							
3		*			*				*			
4		*			*	*			*			
5	*	*			*	*			*			
6	*	*	*		*	*			*			
7	*	*	*		*	*			*			*
8	*	*	*		*	*	*		*			*
9	*	*	*		*	*	*		*		*	*

Model Selection with BIC



Random Forest

Why Random Forest?

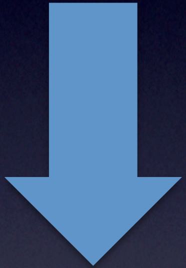
- Unbalanced data
- Predictor correlation reduction
- Feature importance
- No assumption about data

Normal Random Forest

-- unbalanced data

Total error rate: 18.74%

False negative rate: 93.78%



Model Improvement

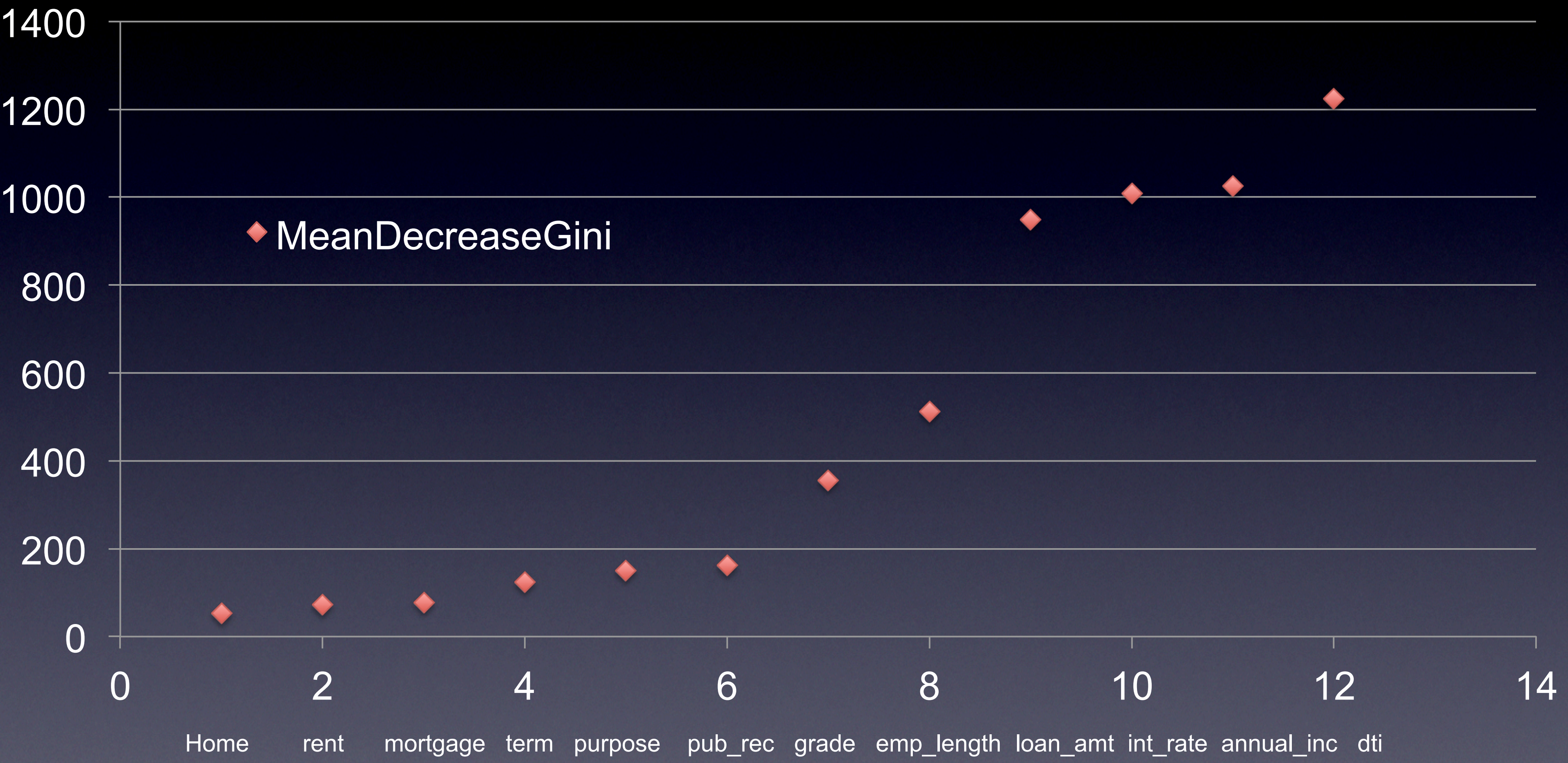
Random Forest with Down-Sampling

-- balanced data

Total error rate: 30.62%

False negative rate： 42.39%

Feature Importance



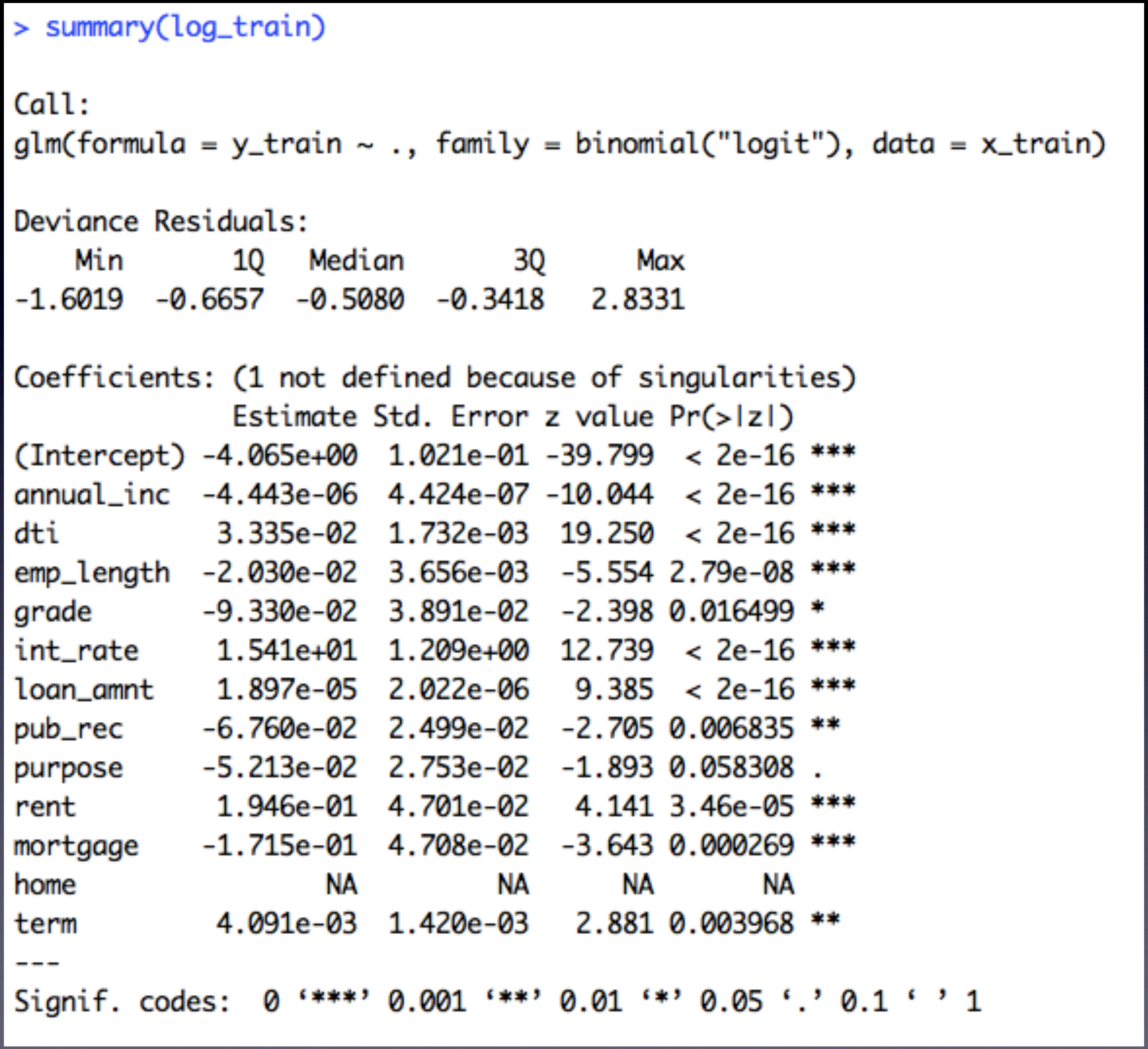
Drawbacks

Logistic Regression	Non-normal: unstable
Random Forest	Down-Sampling: information lost
General	Rough classification

Q & A

Appendix

The result of Logistic Regression



```
> summary(log_train)

Call:
glm(formula = y_train ~ ., family = binomial("logit"), data = x_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6019  -0.6657  -0.5080  -0.3418   2.8331

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.065e+00  1.021e-01 -39.799  < 2e-16 ***
annual_inc  -4.443e-06  4.424e-07 -10.044  < 2e-16 ***
dti          3.335e-02  1.732e-03  19.250  < 2e-16 ***
emp_length  -2.030e-02  3.656e-03  -5.554 2.79e-08 ***
grade       -9.330e-02  3.891e-02  -2.398 0.016499 *
int_rate     1.541e+01  1.209e+00  12.739  < 2e-16 ***
loan_amnt    1.897e-05  2.022e-06   9.385  < 2e-16 ***
pub_rec     -6.760e-02  2.499e-02  -2.705 0.006835 **
purpose     -5.213e-02  2.753e-02  -1.893 0.058308 .
rent         1.946e-01  4.701e-02   4.141 3.46e-05 ***
mortgage    -1.715e-01  4.708e-02  -3.643 0.000269 ***
home                NA         NA      NA       NA
term         4.091e-03  1.420e-03   2.881 0.003968 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The result of Random Forest

	Error Rate	False Negative Rate
Training Data Without Down-Sampling	18.735%	93.778%
Test Data Without Down-Sampling	18.285%	93.980%
Training Data With Down-Sampling	30.621%	42.387%
Test Data With Down-Sampling	30.466%	43.695%