LENDING CLUB DATA ANALYSIS

Predicting a NonPerforming Loan

WHY REVISIT?

- Short timeframe, familiar with the dataset, plus lots of unexplored relationships
- Data is nice mix of continuous and categorical variables
- Historical data goes from 2007 to the present
- Will a loan turn out bad? Will look at sub grade variable, description variable and predict with logistic regression and random forest

THE DATASET

- Using dataset from 2007-2011. Data from 2012 to present can added later on.
- 42,485 observations. 110 variables
- Remove columns 85% NA ->down to 56 variables.
- Loans are either Good or Bad (performing or nonperforming).

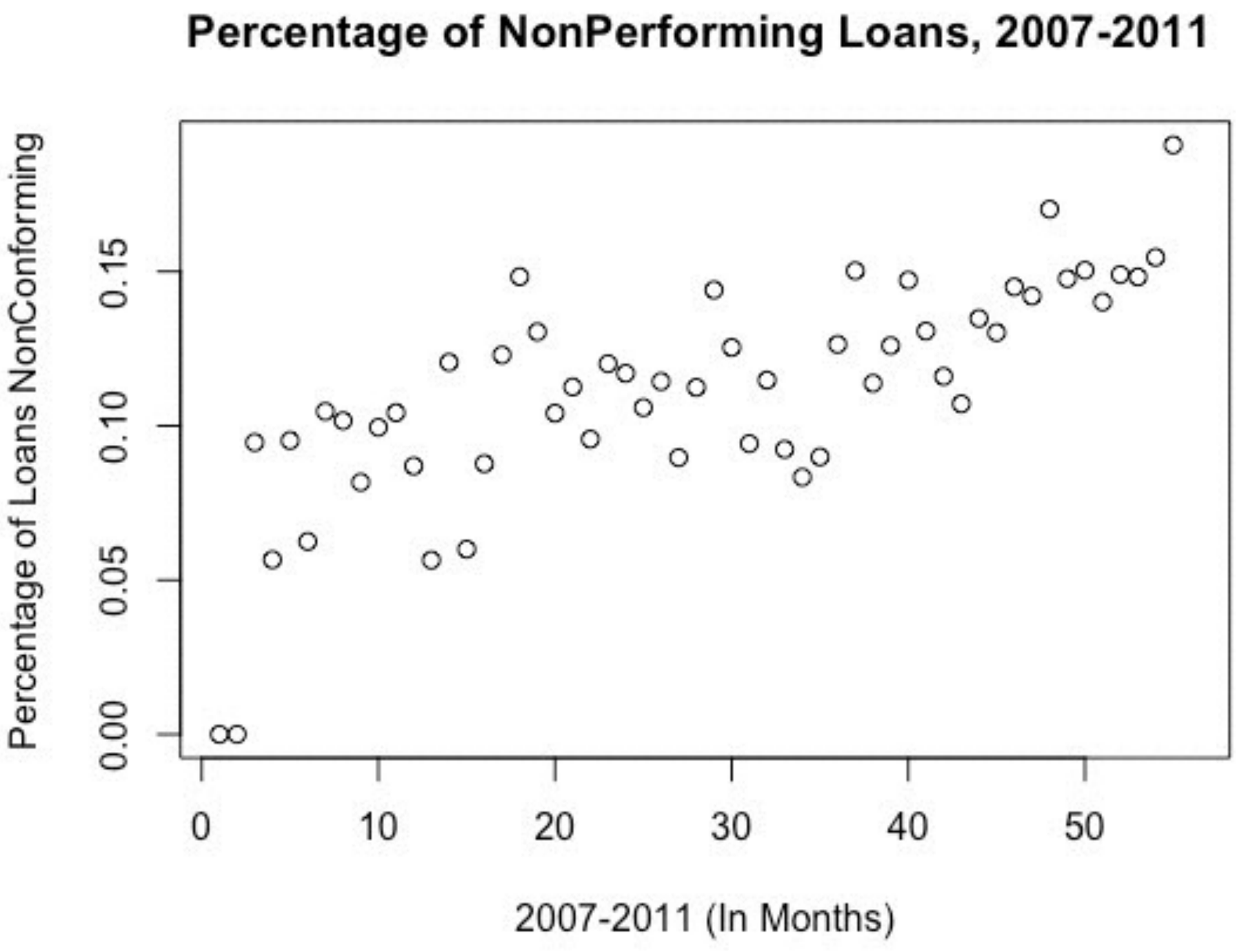
Binary Classification. Mapped 7 categories of loans to 2 categories

Status	Mapping	Description
Fully Paid	Performing	Loan has been fully repaid
Current	Performing	Loan is up to date on all payments
In Grace Period	NonPerforming	Loan is between 0 and 15 days past due
Late (16 - 30 days)	NonPerforming	Loan is between 16 and 30 days past due
Late (31 - 120 days)	NonPerforming	Loan is between 31 and 120 days past due
Default	NonPerforming	Loan is over 121 days past due
Charged Off	NonPerforming	Loan for which there is no reasonable expectation of additional payments

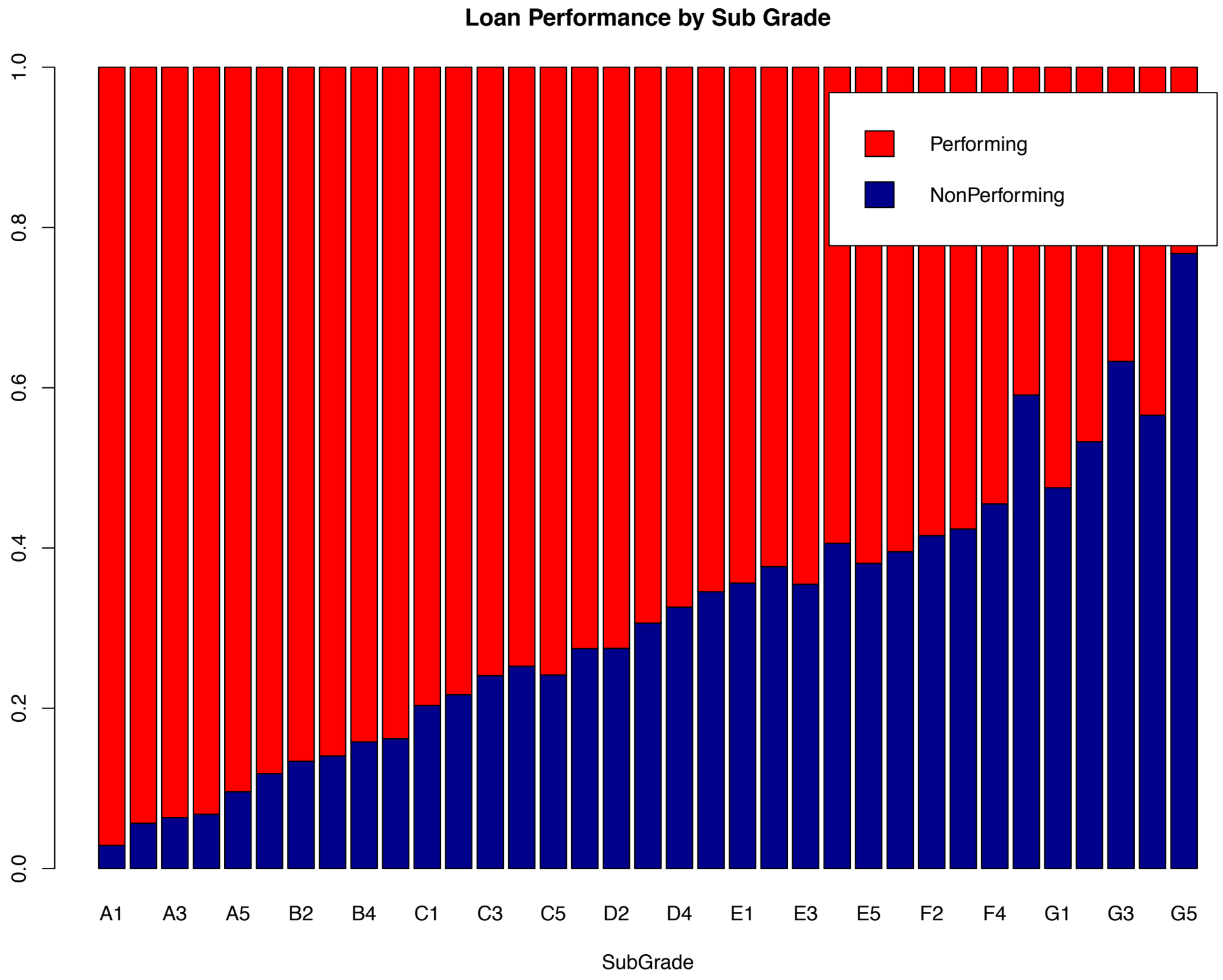
THE DATASET

- Variable Subgrade : Lending Club maps borrowers to a series of grades [A-F] and subgrades [A-F][1-5] based on their **risk profile**. Loans in each subgrade are then given appropriate interest rates. The specific rates will change over time according to market conditions, but generally they will fall within a tight range for each subgrade.

HOW MANY BAD LOANS?



PERCENTAGE BAD LOANS PER SUB GRADE

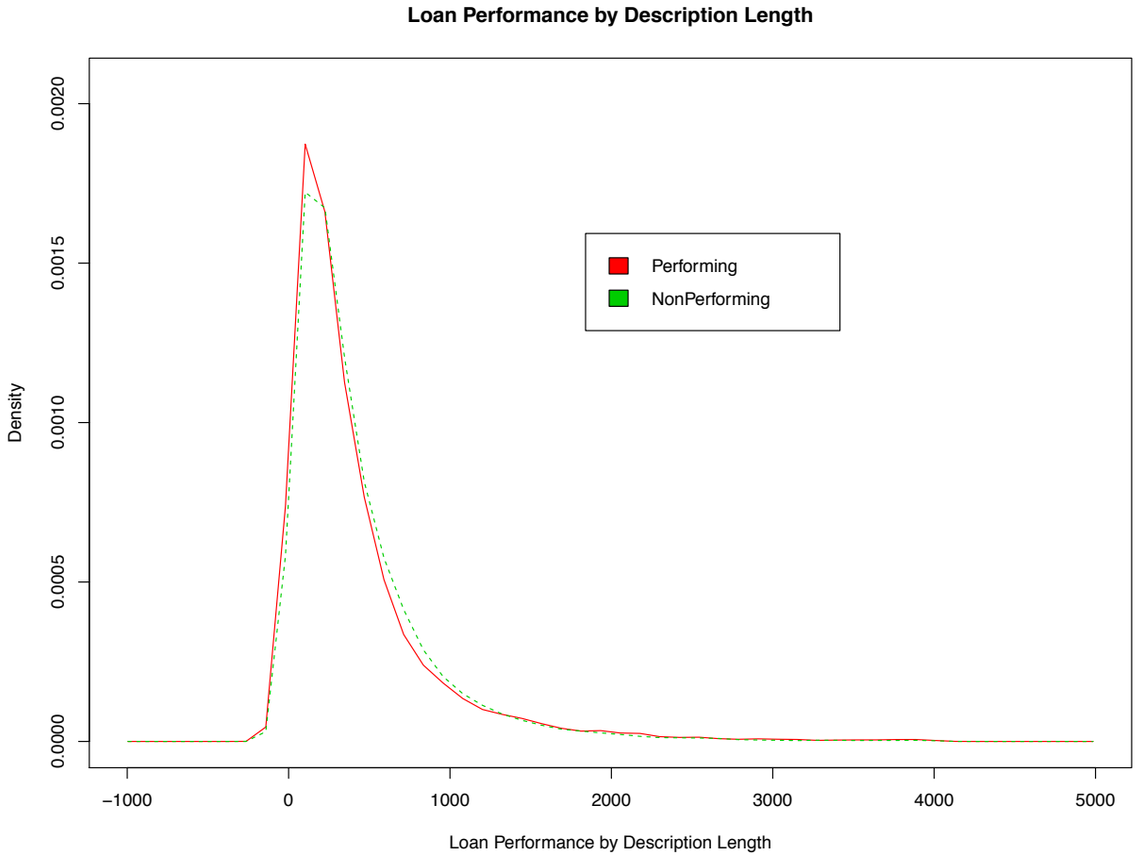


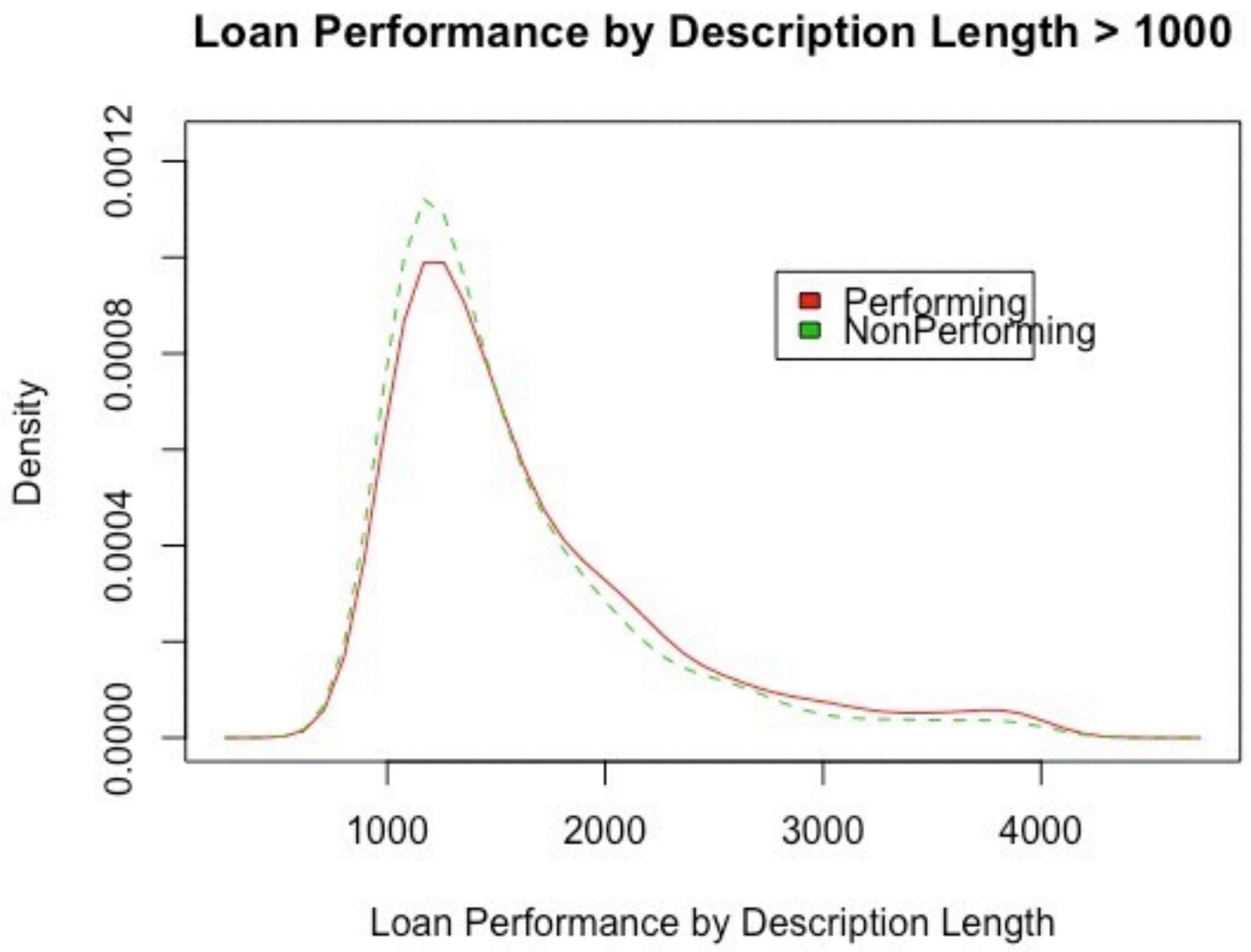
WHAT IS DESCRIPTION FIELD?

I plan on combining three large interest bills together and freeing up some extra each month to pay toward other bills. I've always been a good payor but have found myself needing to make adjustments to my budget due to a medical scare. My job is very stable, I love it.

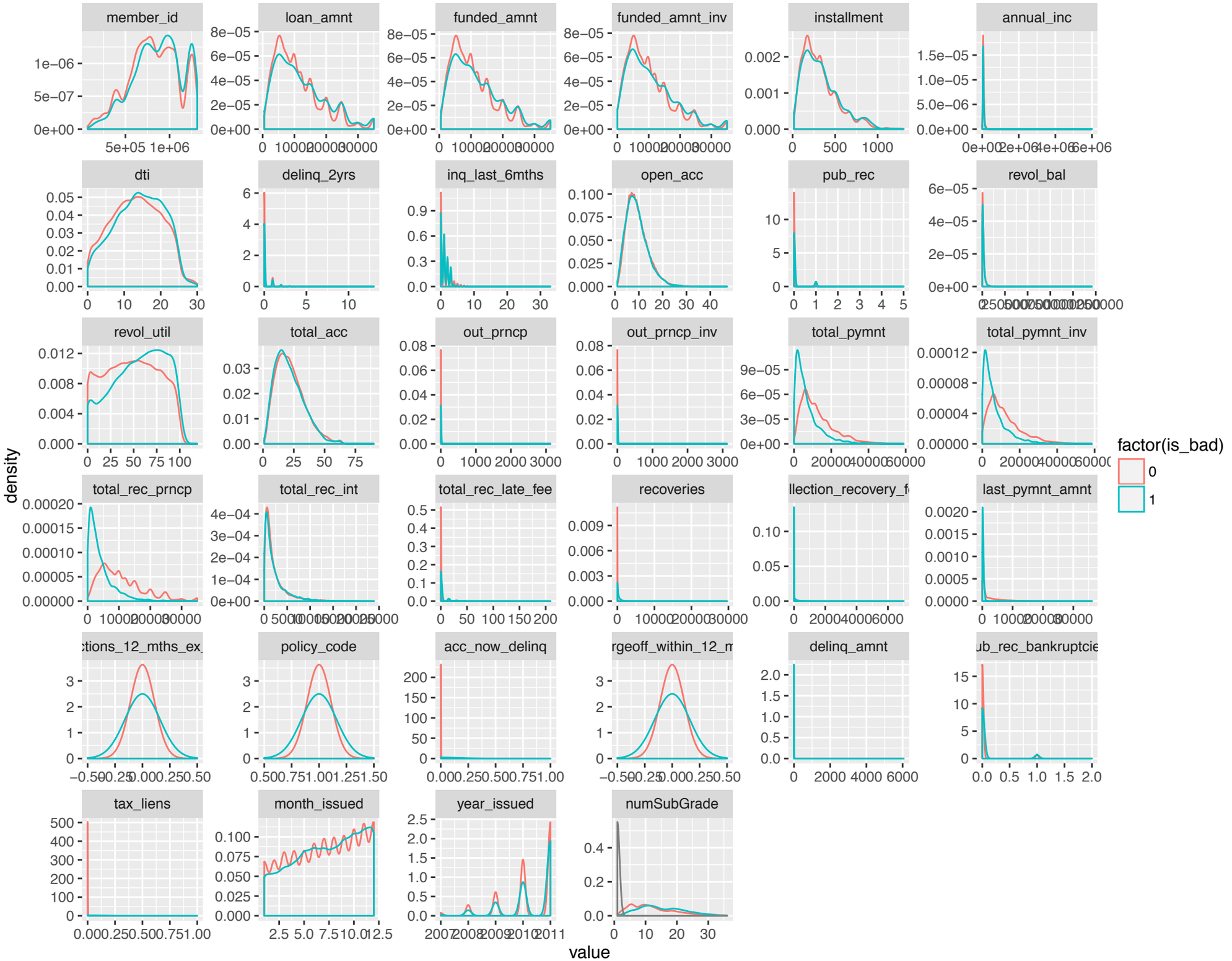
I plan to use this money to finance the motorcycle i am looking at. I plan to have it paid off as soon as possible/when i sell my old bike. I only need this money because the deal im looking at is to good to pass up. I plan to use this money to finance the motorcycle i am looking at. I plan to have it paid off as soon as possible/when i sell my old bike.I only need this money because the deal im looking at is to good to pass up. I have finished college with an associates degree in business and its taking me places

DENSITY OF LOANS VS. DESCRIPTION LENGTH

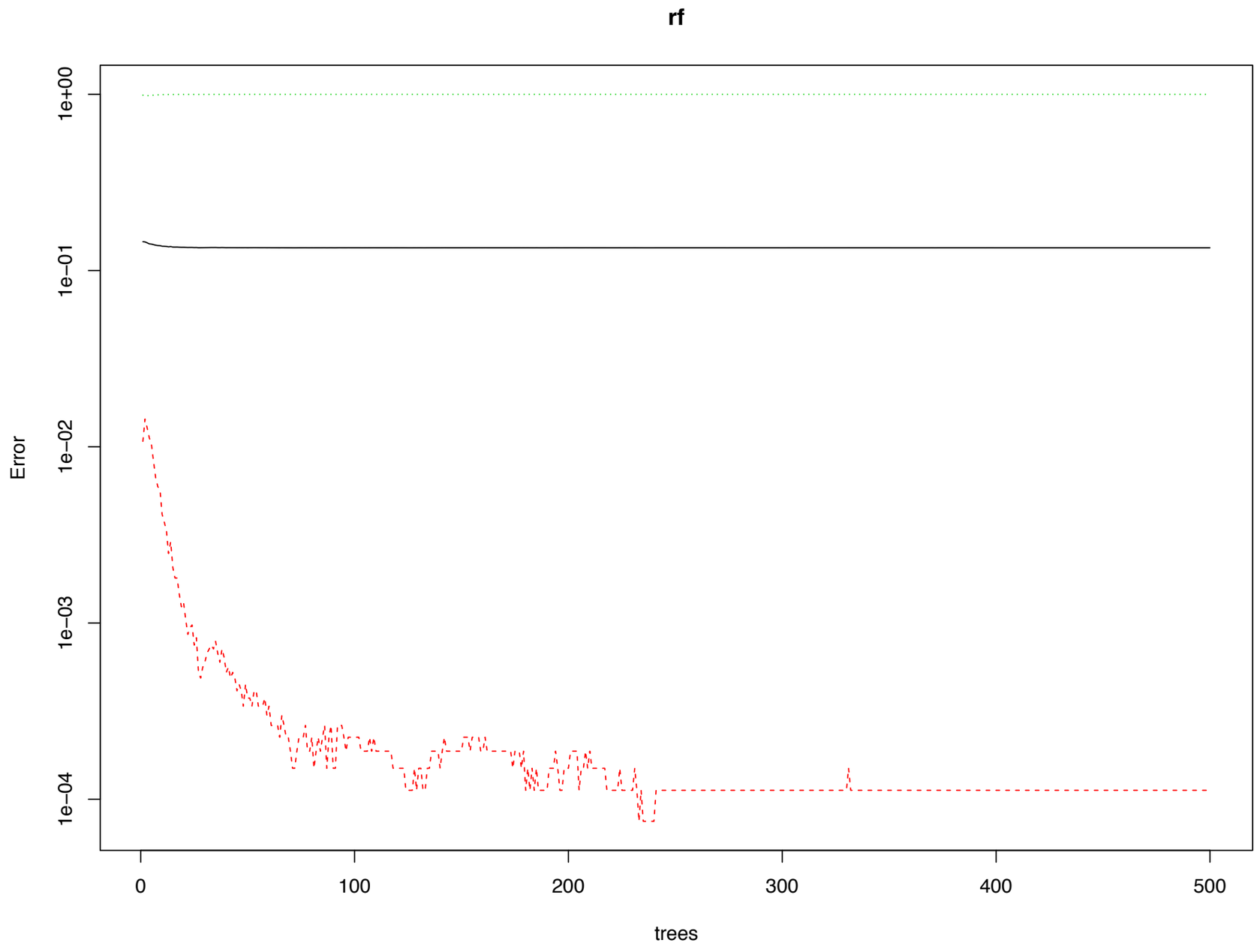




WHAT VARIABLES CONTRIBUTE TO A LOAN GONE BAD?



RANDOM FOREST FOR CLASSIFICATION



No. of variables tried at each split: 2

OOB estimate of error rate: 13.52%

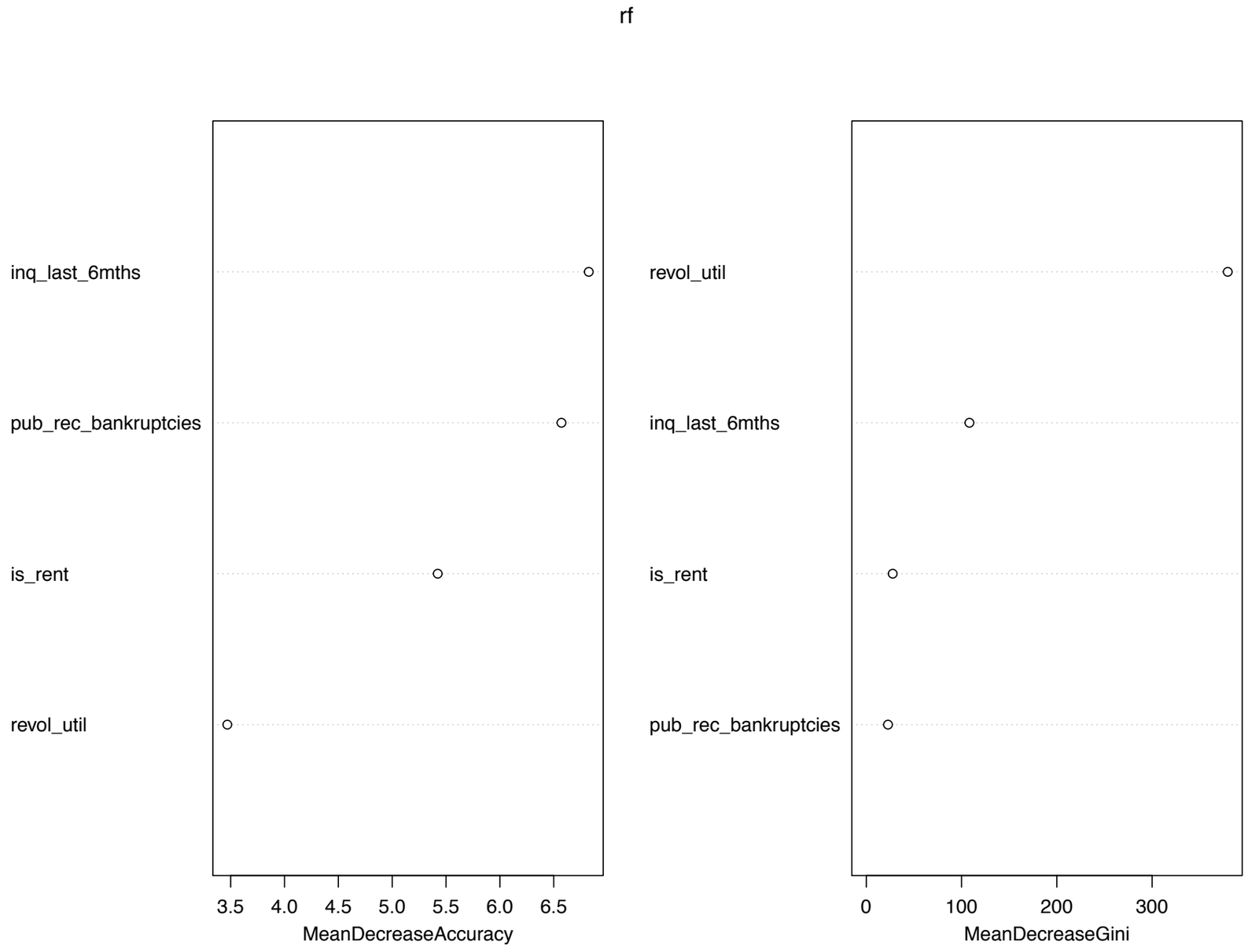
Confusion matrix:

0 1 class.error

0 26746 3 0.0001121537

1 4179 0 1.0000000000

RANDOM FOREST VARIABLE IMPORTANCE



0 1 MeanDecreaseAccuracy MeanDecreaseGini

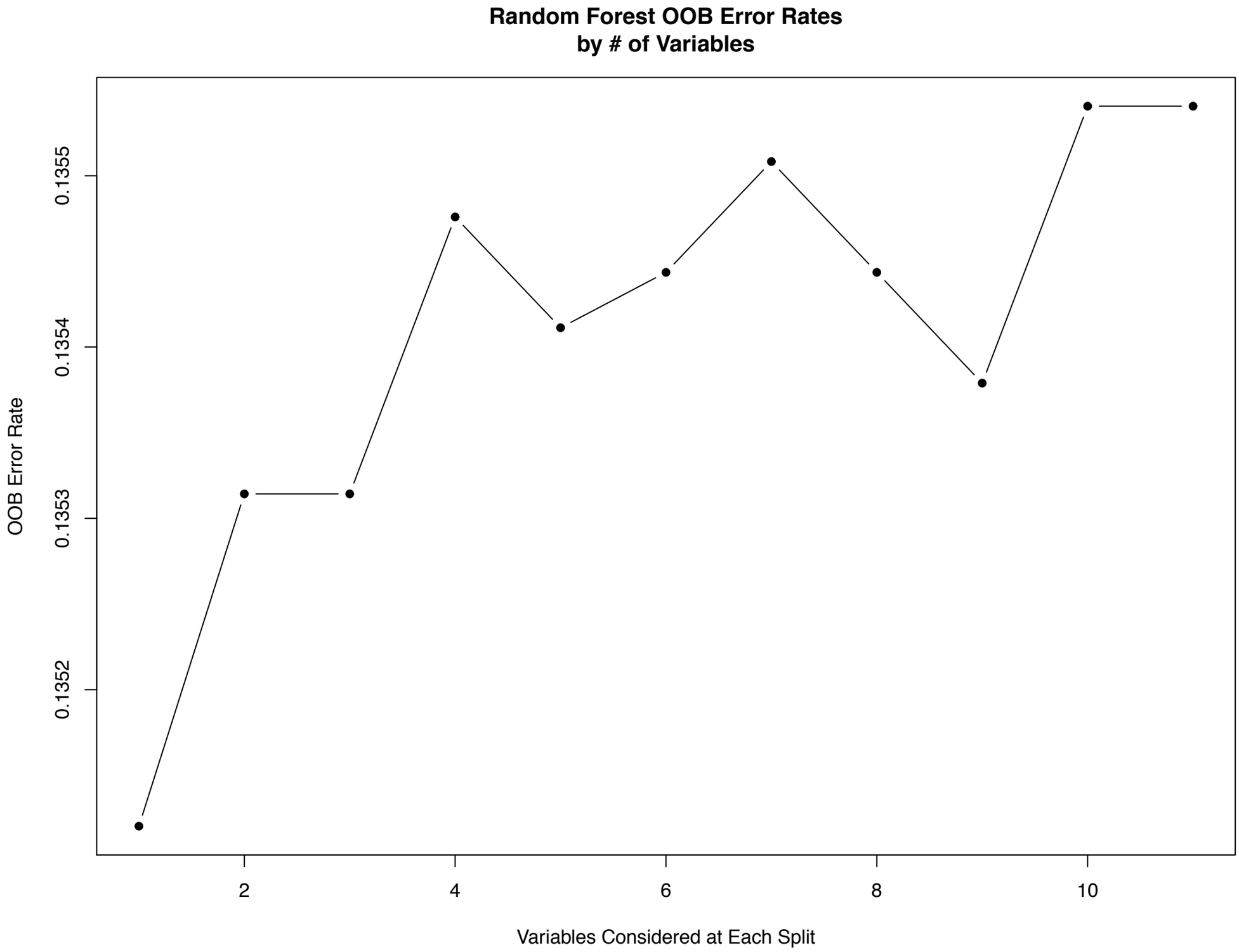
pub_rec_bankruptcies -1.8850001 18.076604 6.571964 22.78685

revol_util 0.3501855 7.642401 3.468845 379.39139

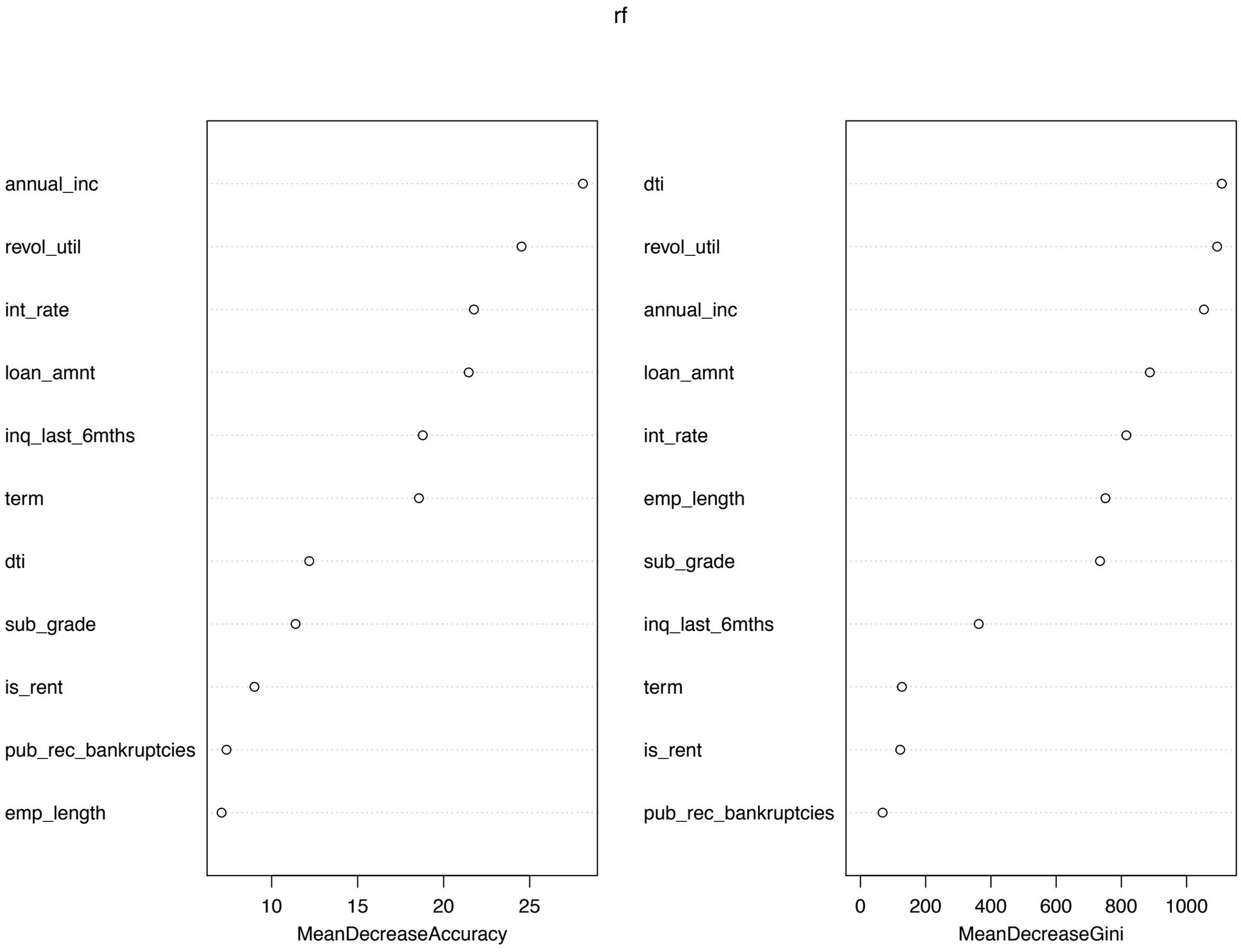
inq_last_6mths 2.8783556 10.495827 6.825414 108.24223

is_rent 5.5685251 0.542448 5.422811 27.73430

RANDOM FOREST FOR CLASSIFICATION

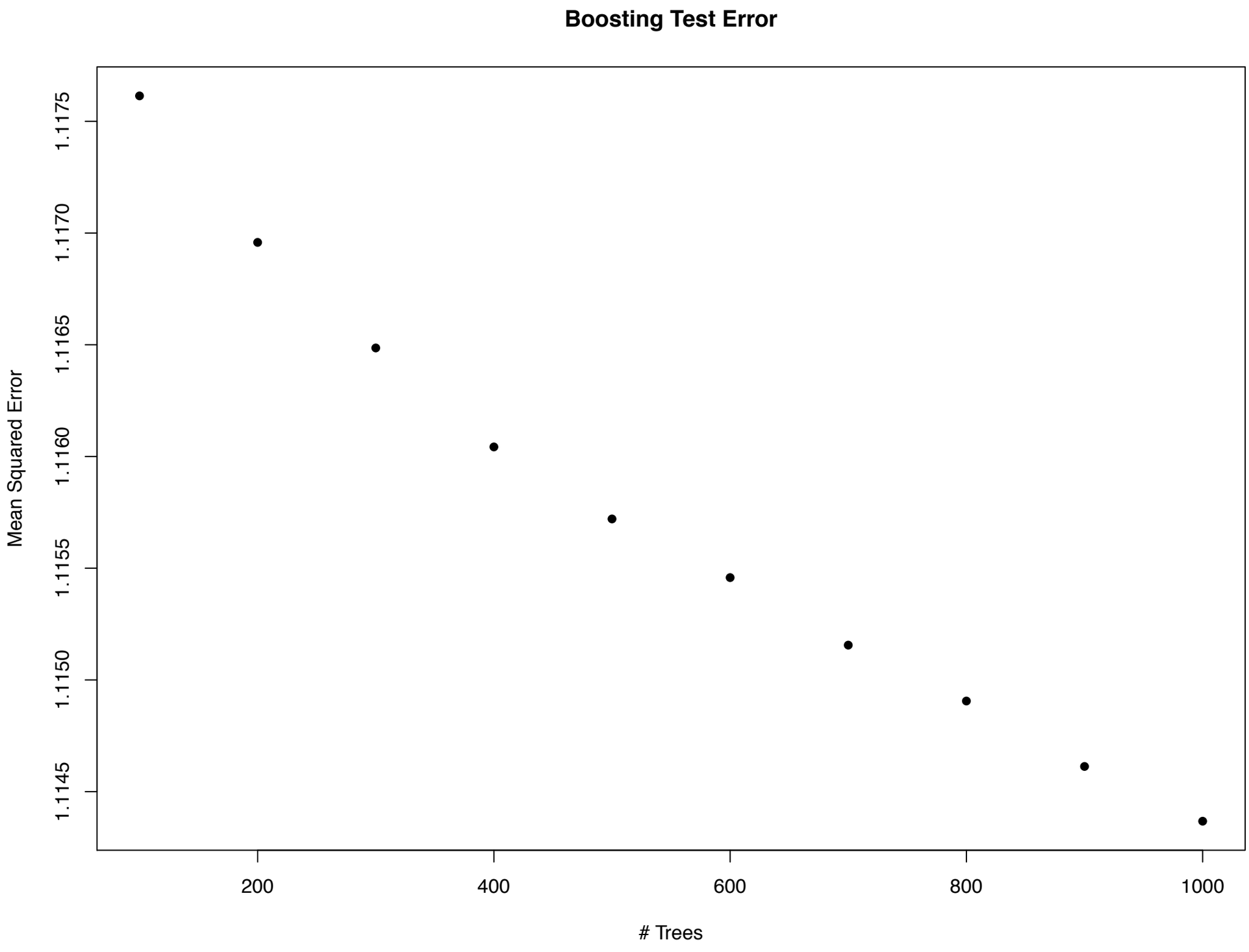


RANDOM FOREST VARIABLE IMPORTANCE

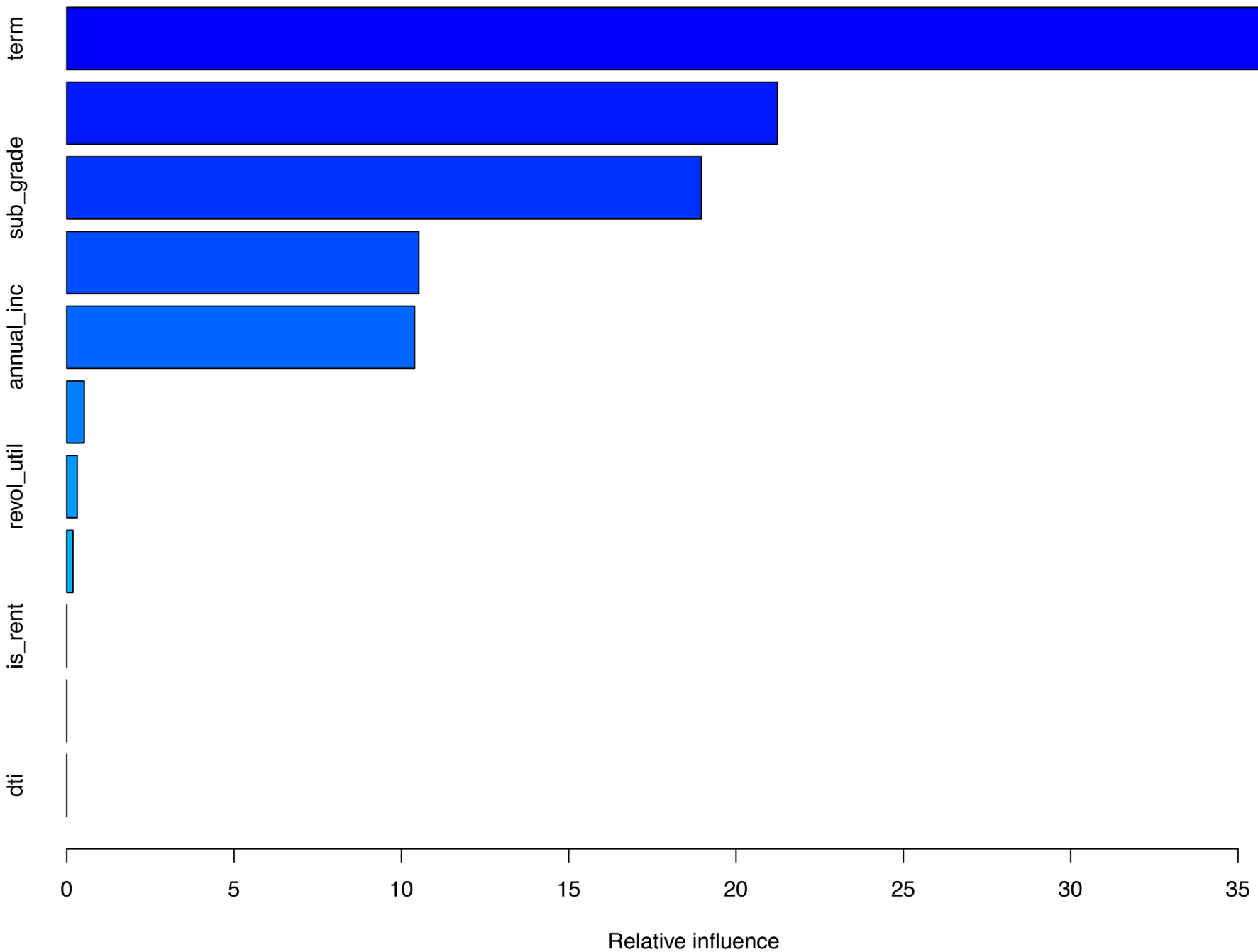


- Mean Decrease Accuracy - MDA - the prediction error on out of bag portion of data
- Mean Decrease Gini - measure of "node impurity" in tree-based classification. A low Gini (higher decrease in Gini) means that a particular predictor variable plays a greater role in partitioning the data into the defined classes.

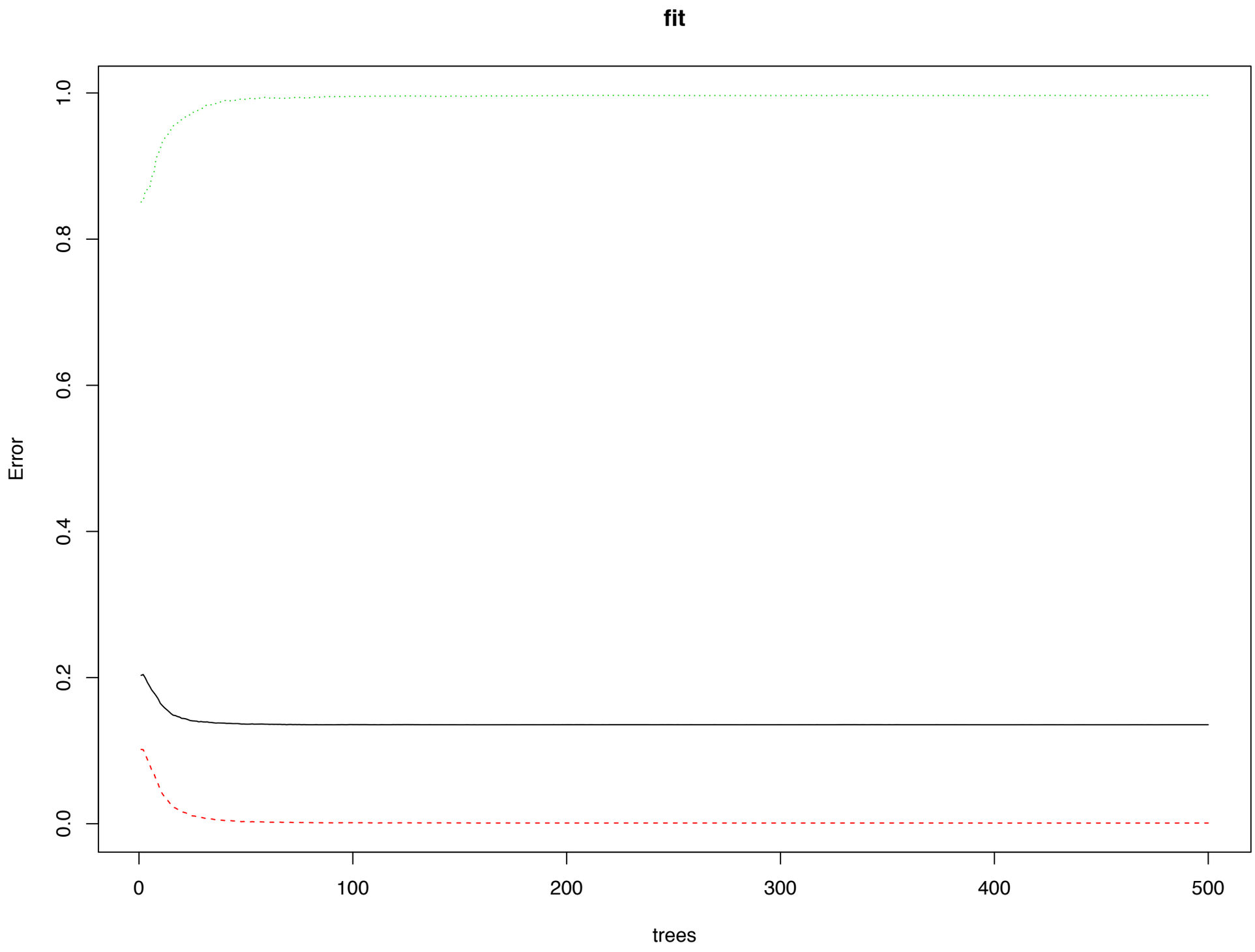
GRADIENT BOOST



GBM RELATIVE INFLUENCE



GBM RESULT



CONCLUSIONS

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 3

OOB estimate of error rate: 13.54%

Confusion matrix:

0 1 class.error

0 26724 25 0.0009346144

1 4163 16 0.9961713329

gbm(formula =

factor(is_bad) ~ pub_rec_bankruptcies + revol_util +

inq_last_6mths + is_rent + loan_amnt + term + sub_grade +

int_rate + emp_length +

annual_inc + dti, distribution = "gaussian",

data = train, n.trees = 1000, interaction.depth = 4)

A gradient boosted model with gaussian loss function.

1000 iterations were performed.

There were 11 predictors of which 8 had non-zero influence.