Midterm

Team 3 in INFO 7390

Team Members: Hongwei Hu, Guangnan Liang

GitHub: <https://github.com/Besimilar/Single_Family_Loan>

Docker Hub: besimilar/advanced-data-analysis:mid

Resources:

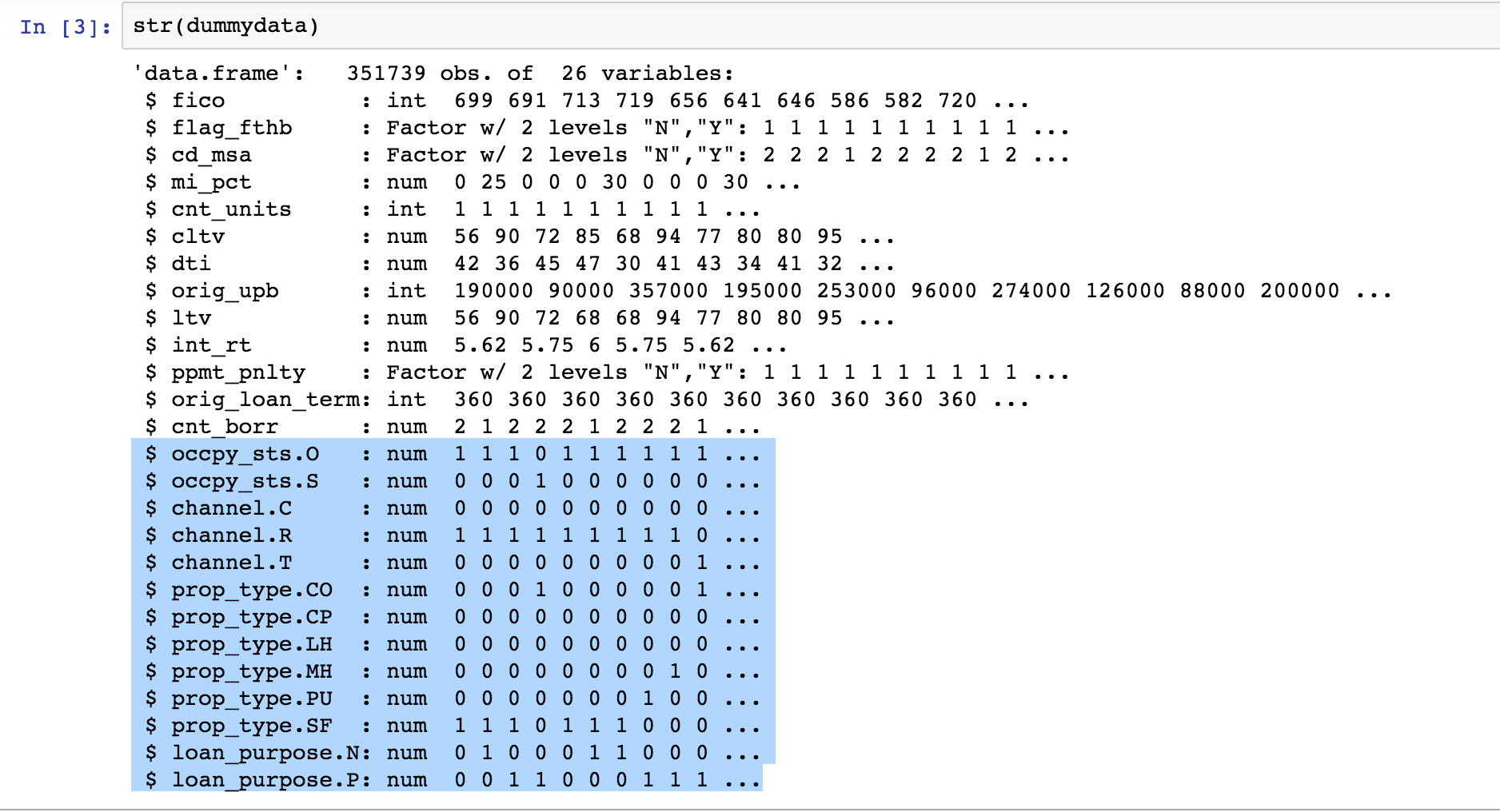
Single-family loan data: <http://www.freddiemac.com/research/datasets/sf_loanlevel_dataset.html>

Note: Red lines means Very Important Steps.

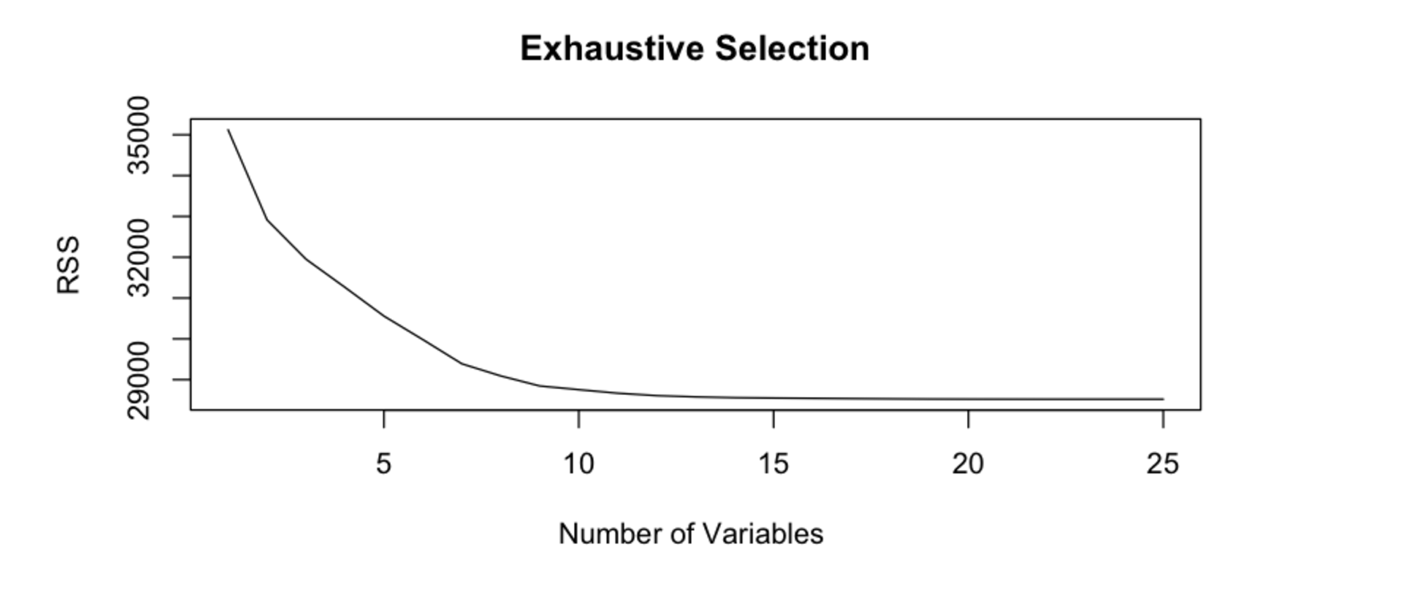
**Building and evaluating models**

**Part 1: Prediction**

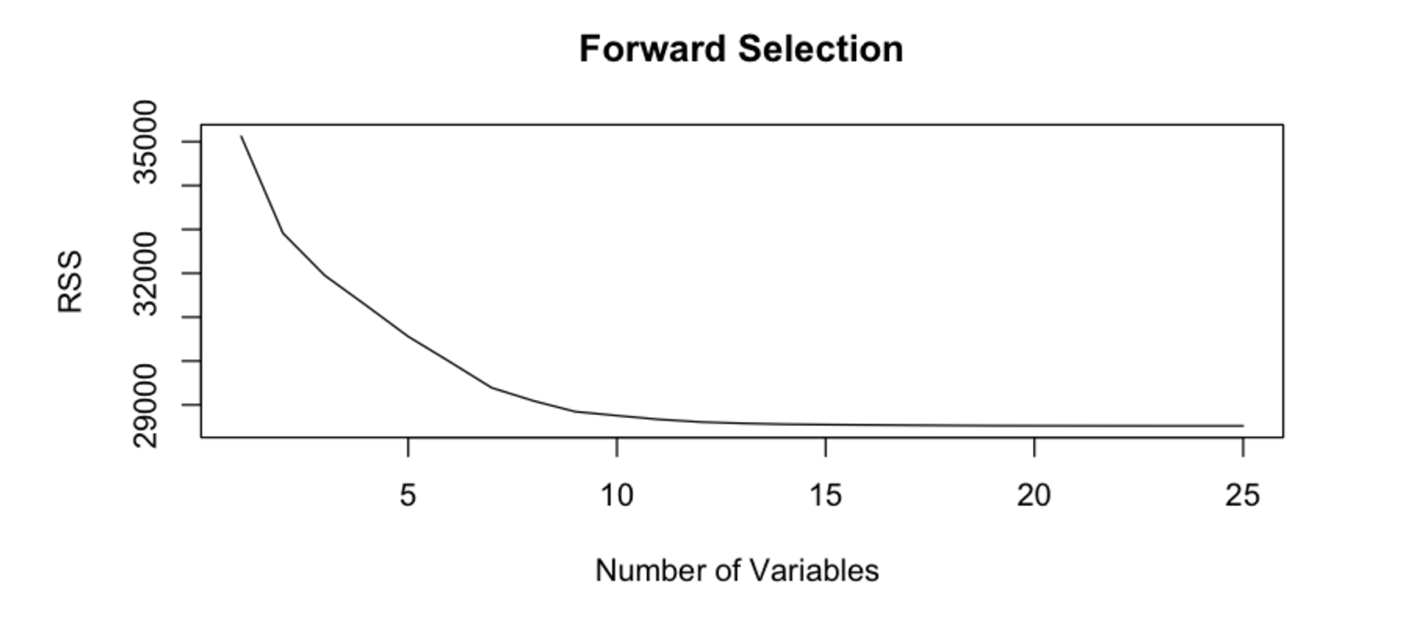
1. Pre-Process:
   1. Remove all dates, names, or location information.
   2. Dummy all categorical columns:
      1. occpy\_sts
      2. channel
      3. prop\_type
      4. loan\_purpose
   3. For details of cleansing, refer to “Prediction\_preprocess.R”.



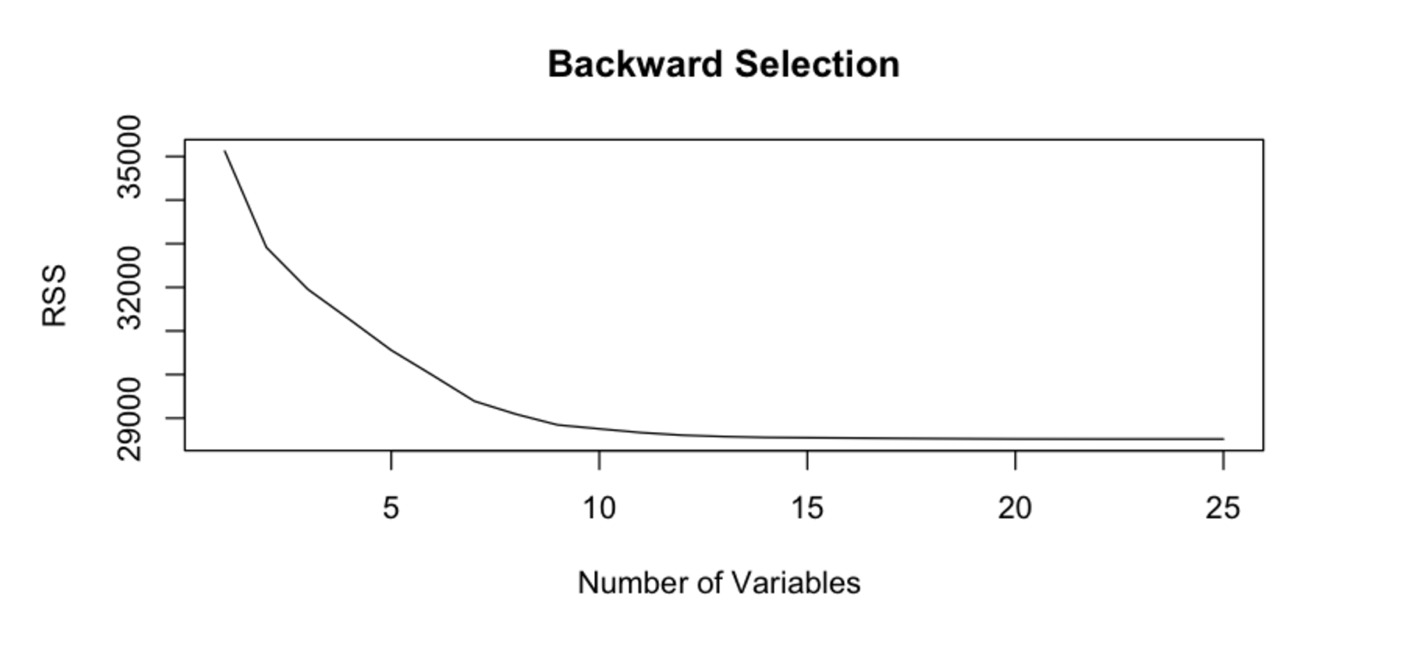
1. Regression
   1. Variable selections:
      1. Exhaustive:
         1. Choosing 10 variables



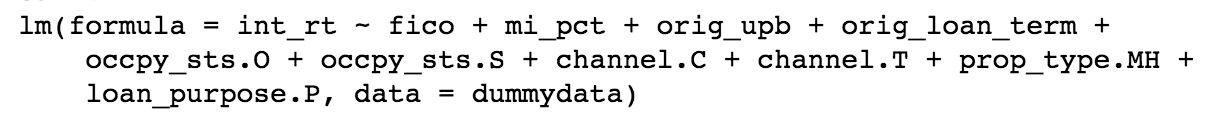
* + 1. Forward:
       1. Choosing 10 variables

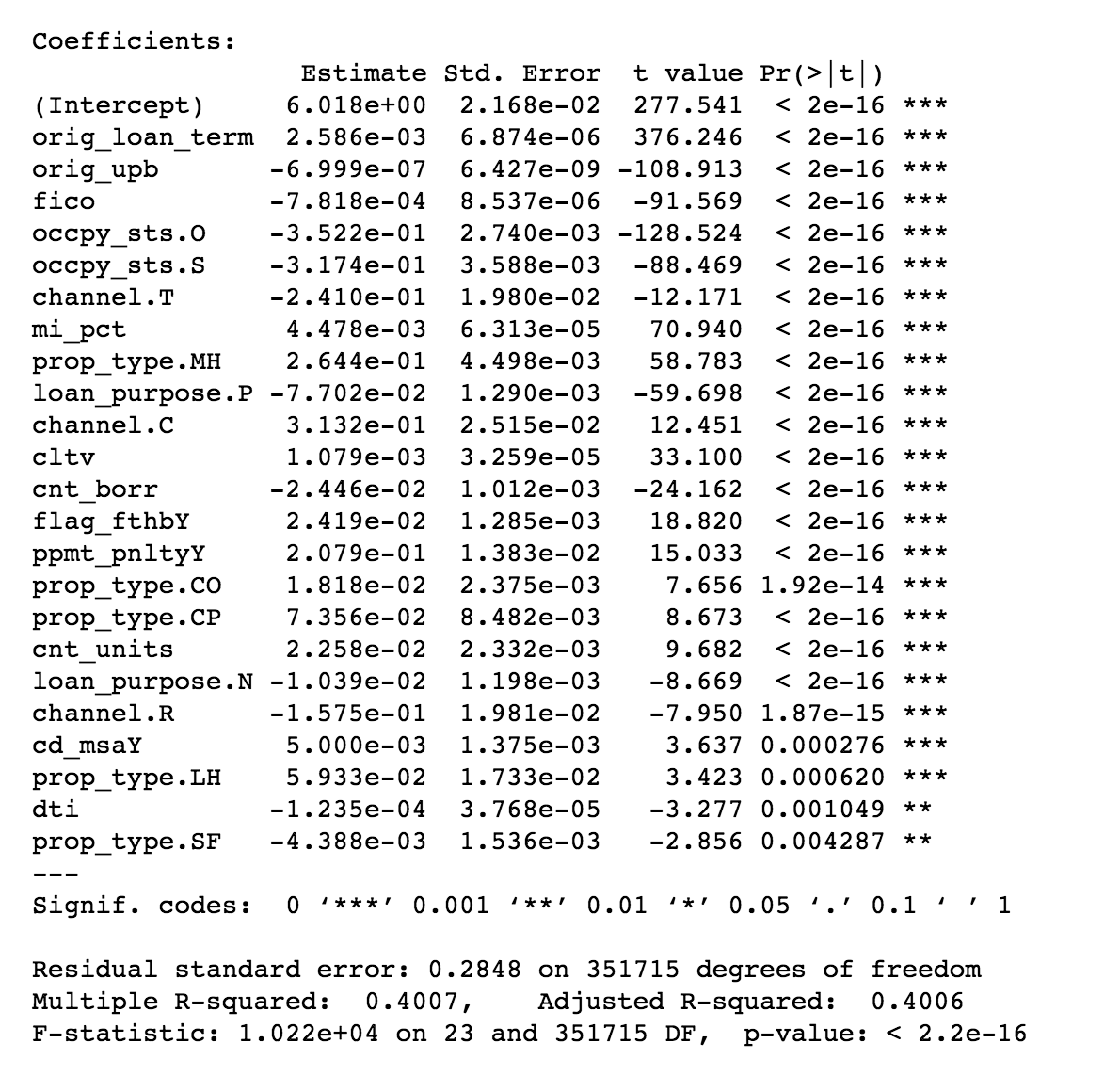


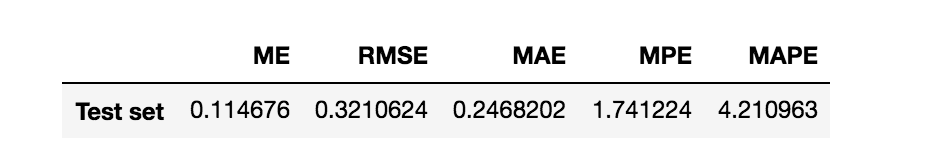
* + 1. Backward:
       1. Choosing 10 variables



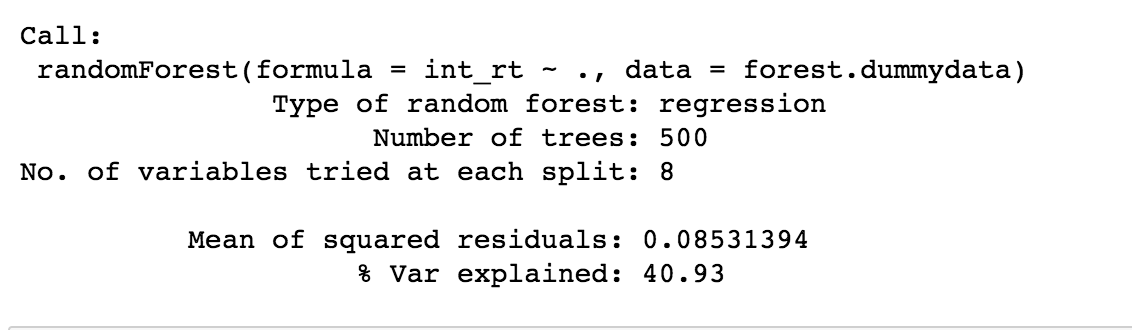
* + 1. Stepwise:
       1. Choosing 23 variables
  1. Best Mode:
     1. Using 10 variables:



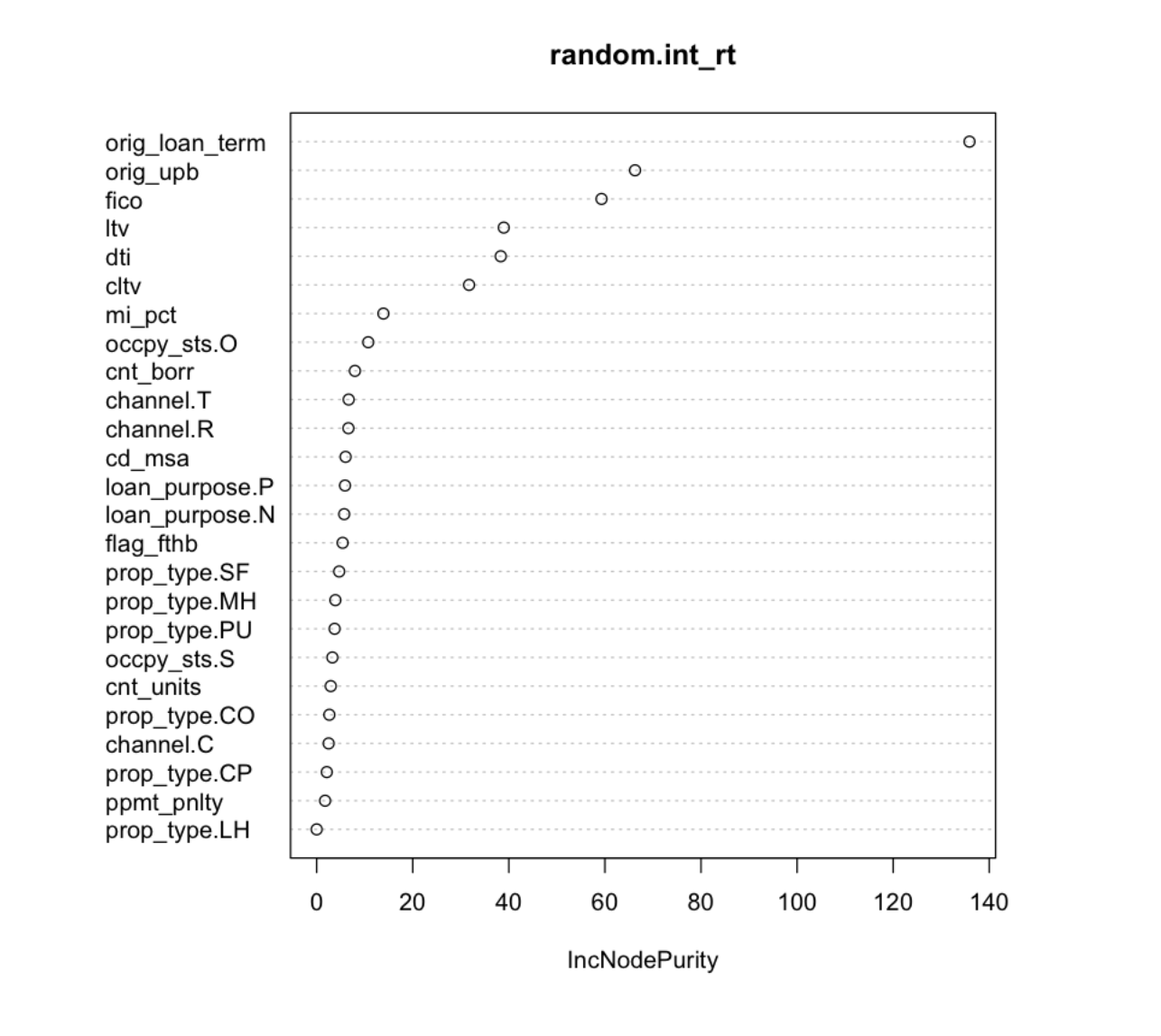




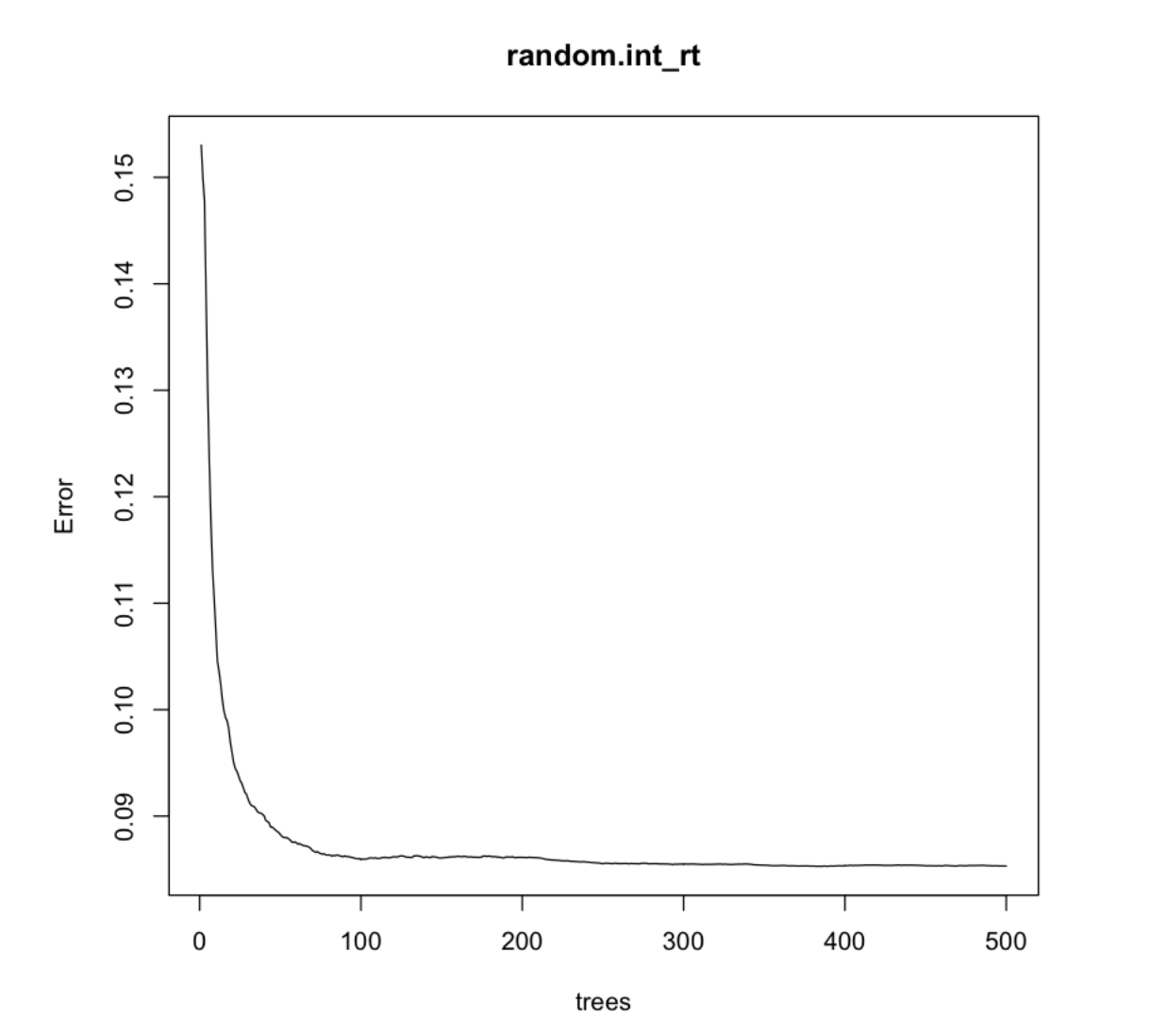
1. Random Forest
   1. Pre-process: Too slow, only use 1% data
   2. Result:



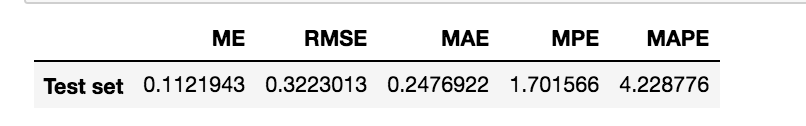
* + 1. Variable Importance:



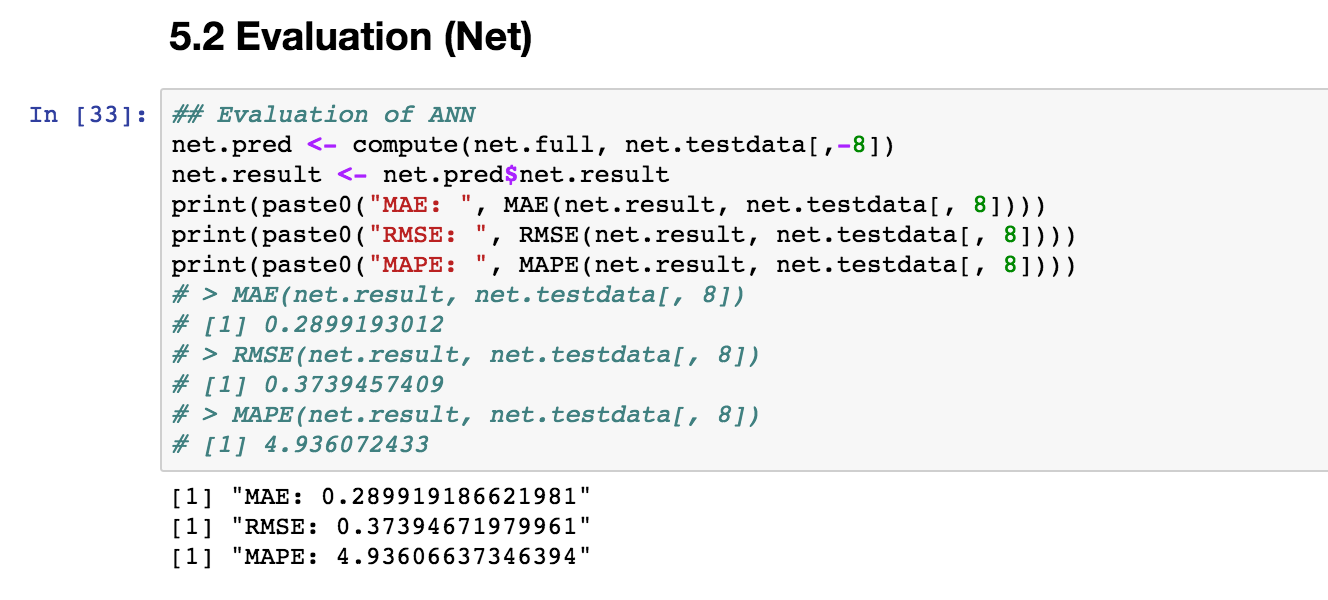
* + 1. Number of Trees:



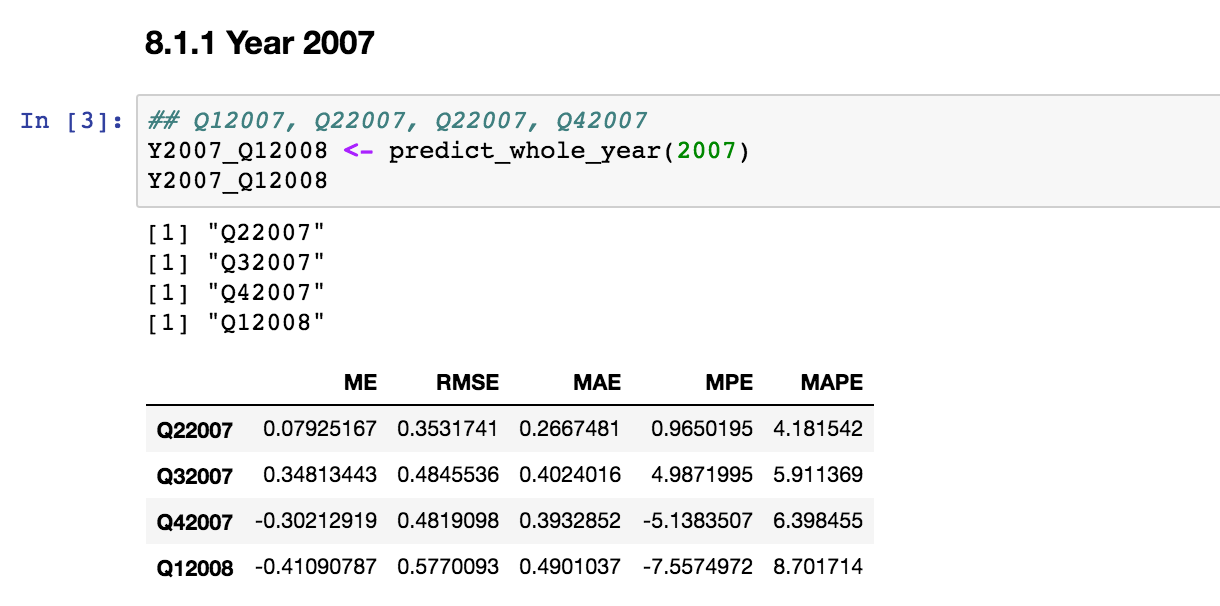
* + 1. Evaluation of Forest:



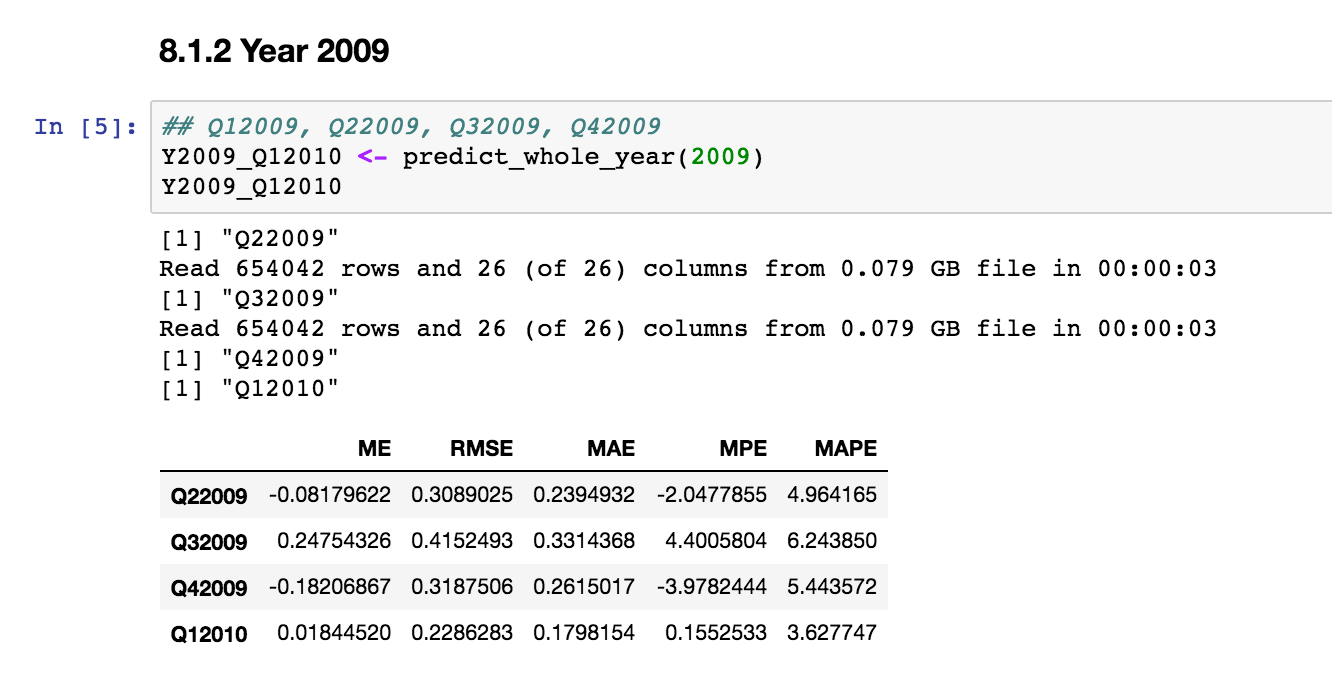
1. Neural Network
   1. Configuration:
      1. Variables: all
      2. 1 layer
      3. 2 nodes
      4. Threshold: 0.01
   2. Result:



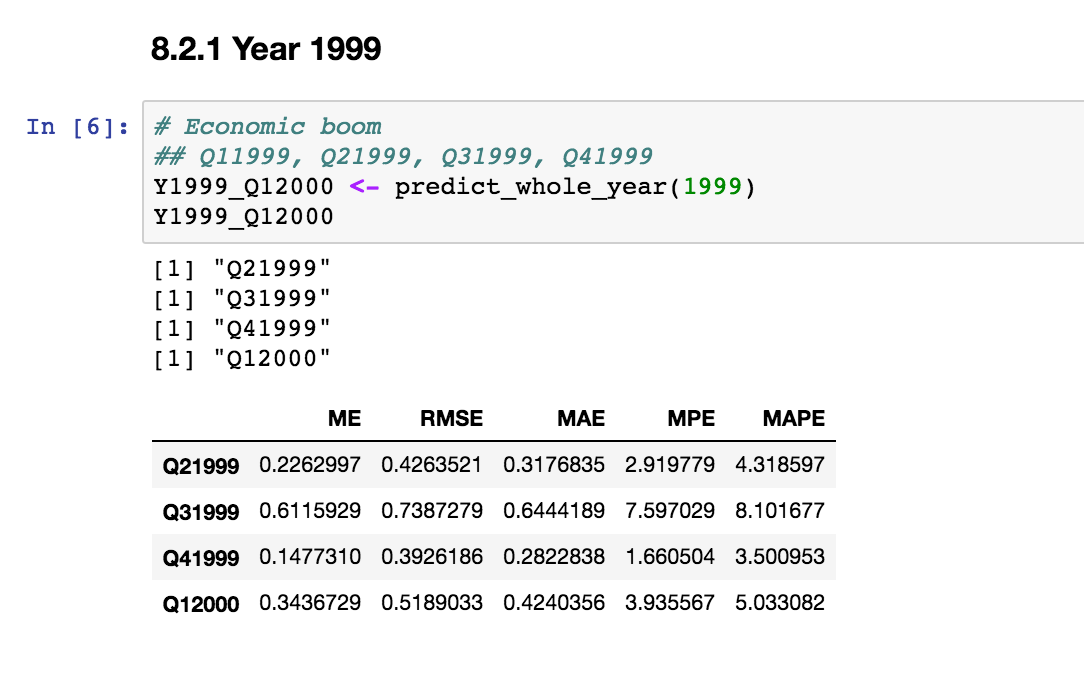
1. What-if Analysis:
   1. Financial crisis:
      1. 2007



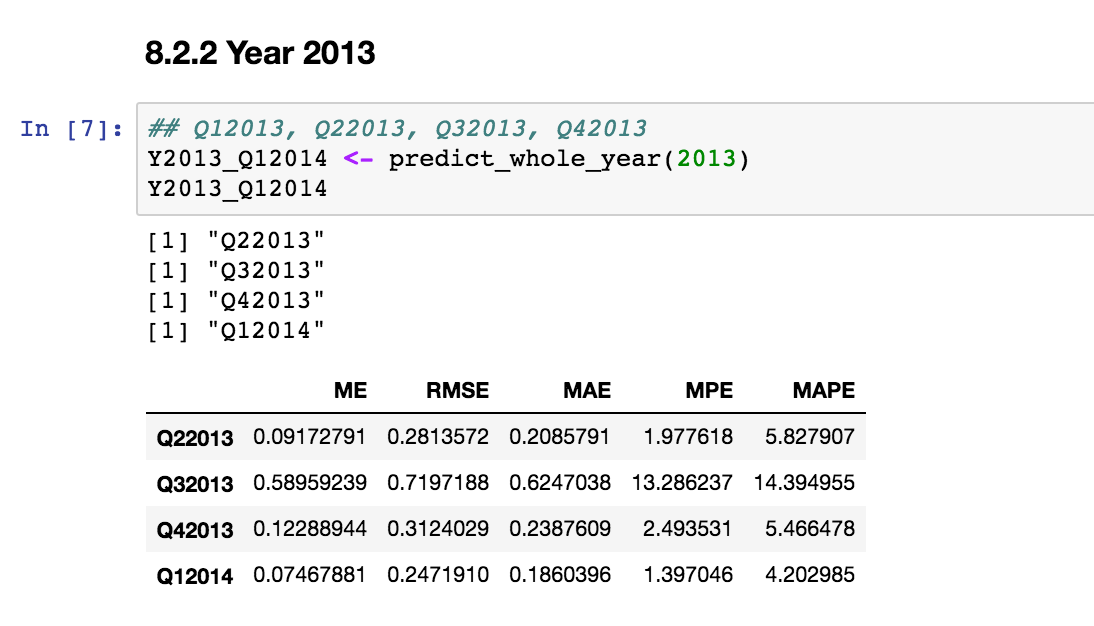
* + 1. 2009



* 1. Economic boom:
     1. 1999



* + 1. 2013

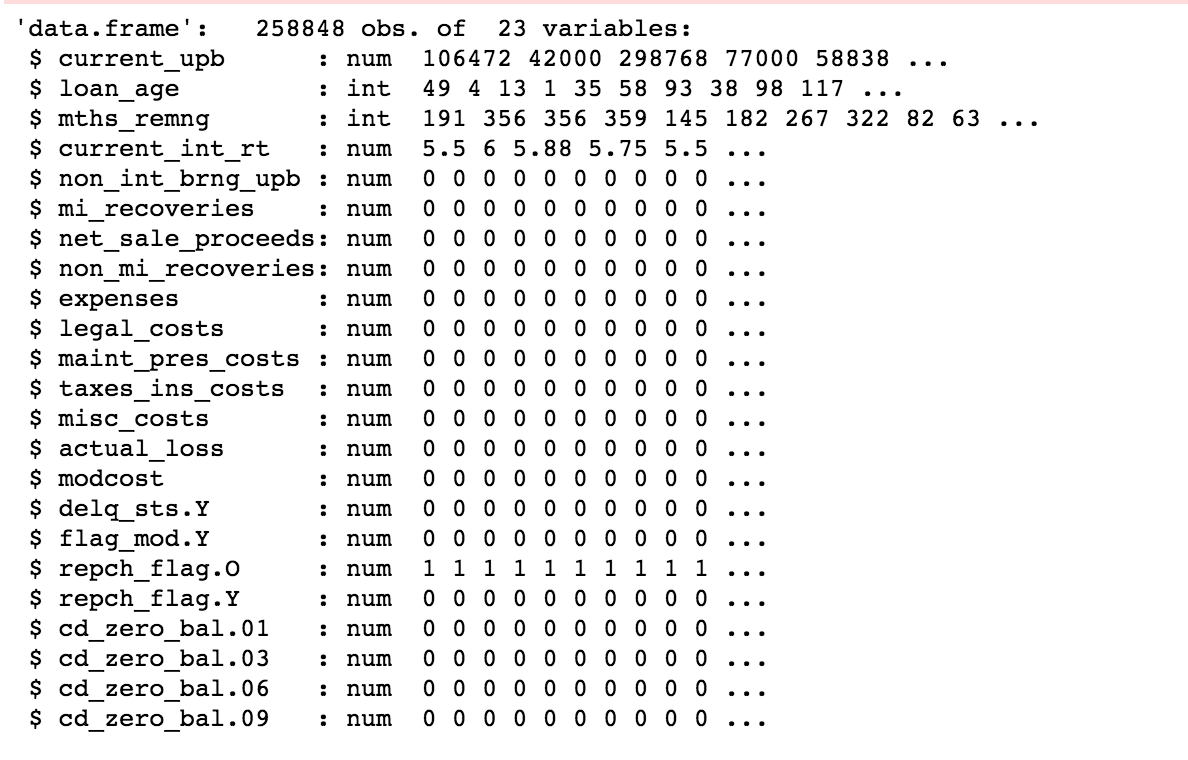


* 1. Analysis: The result is fairly good.
     1. Model:
        1. Regression
        2. 10 Variables: fico+mi\_pct+orig\_upb+orig\_loan\_term+occpy\_sts.O+occpy\_sts.S+channel.C+channel.T+prop\_type.MH+loan\_purpose.P
     2. a small RMSE
     3. an acceptable MAPE

**Part 2: Classification**

1. Pre-preocess:

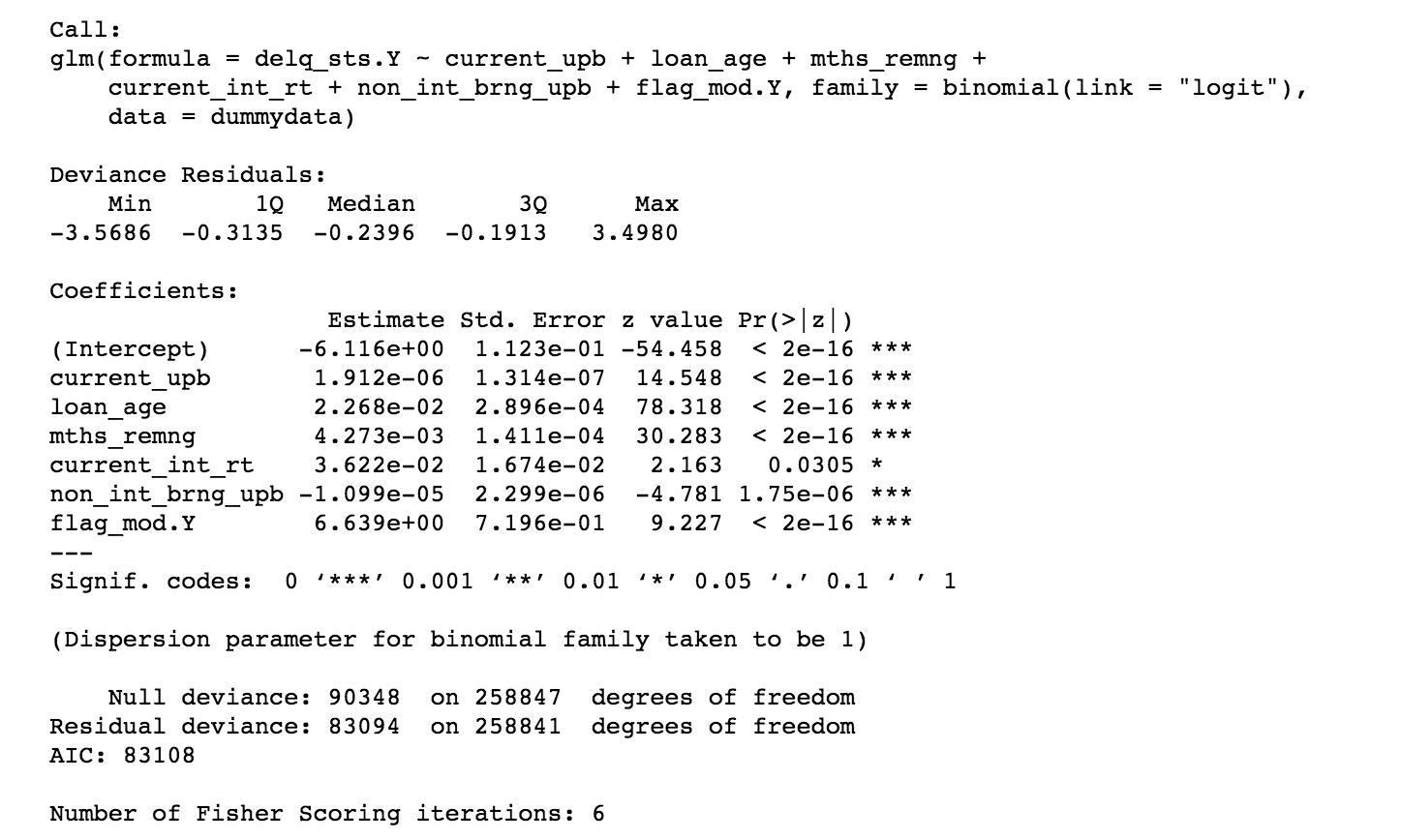
1. Remove all dates, names, or location information.
2. Dummy all categorical columns:
   1. delq\_sts
   2. flag\_mod
   3. repch\_flag
   4. cd\_zero\_bal
3. Only Picking 1% percent data
4. For details of cleansing, refer to “Classification\_preprocess.R”.



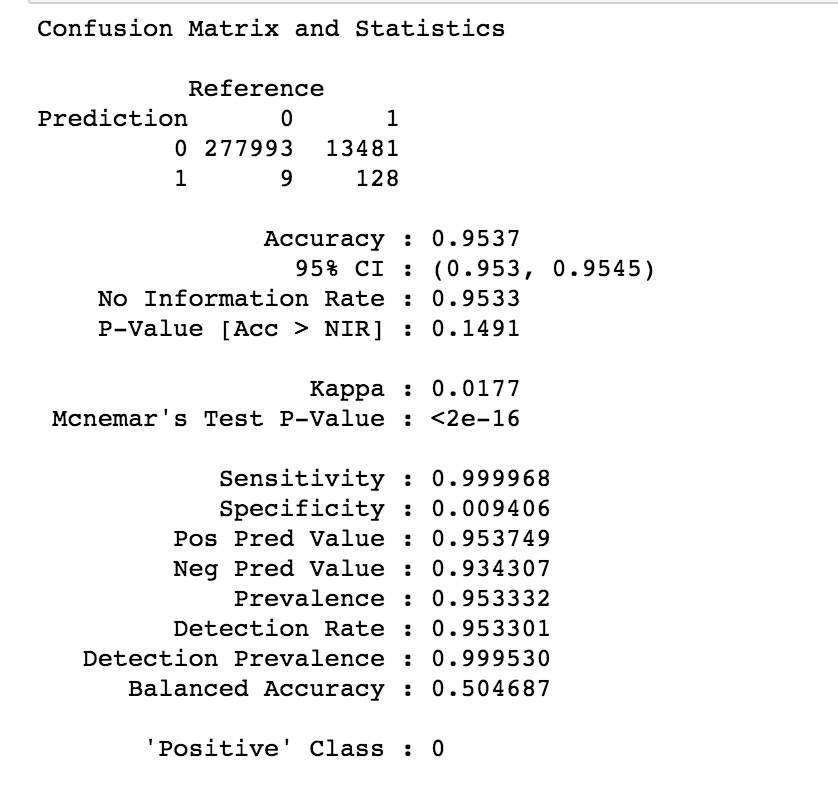
2. Logistic regression:

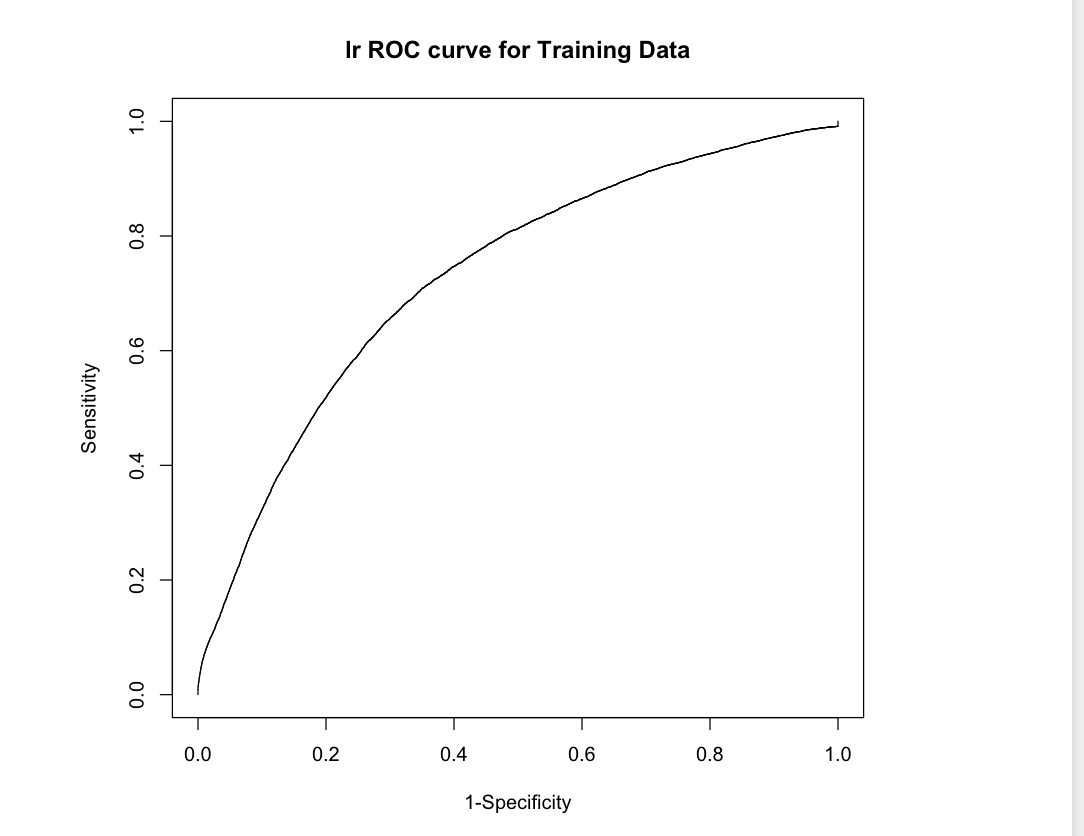
a. 6 Variables:

current\_upb + loan\_age + mths\_remng + current\_int\_rt + non\_int\_brng\_upb + flag\_mod.Y



b. Confusion Matrix:

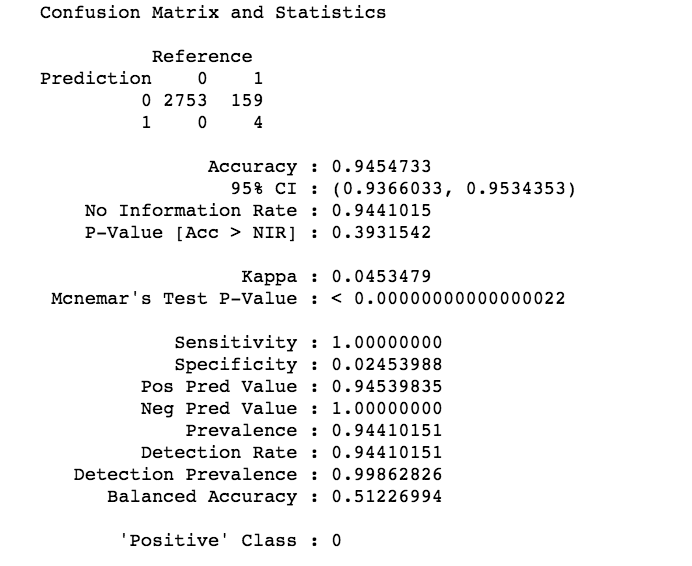


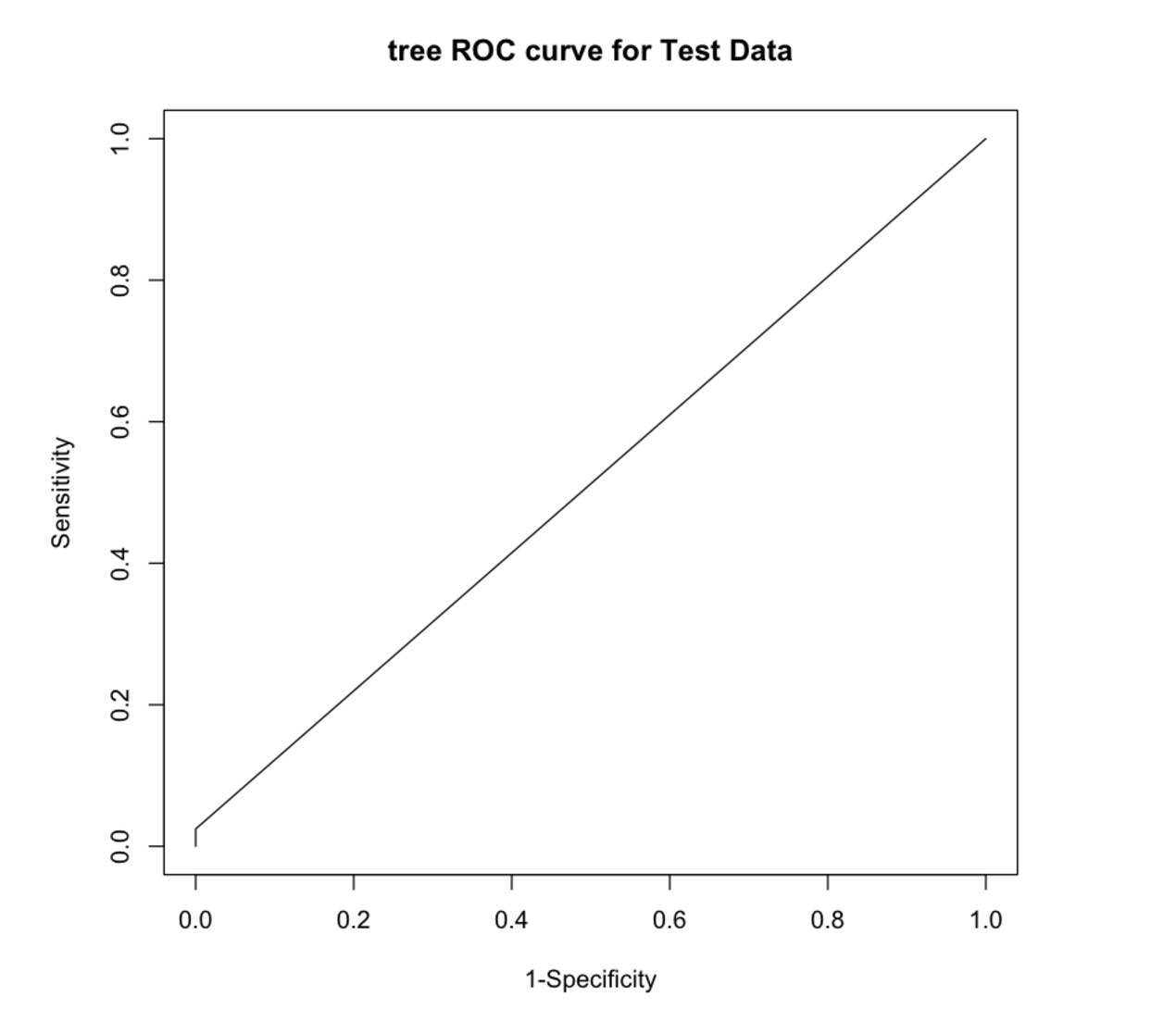
c. ROC curve:

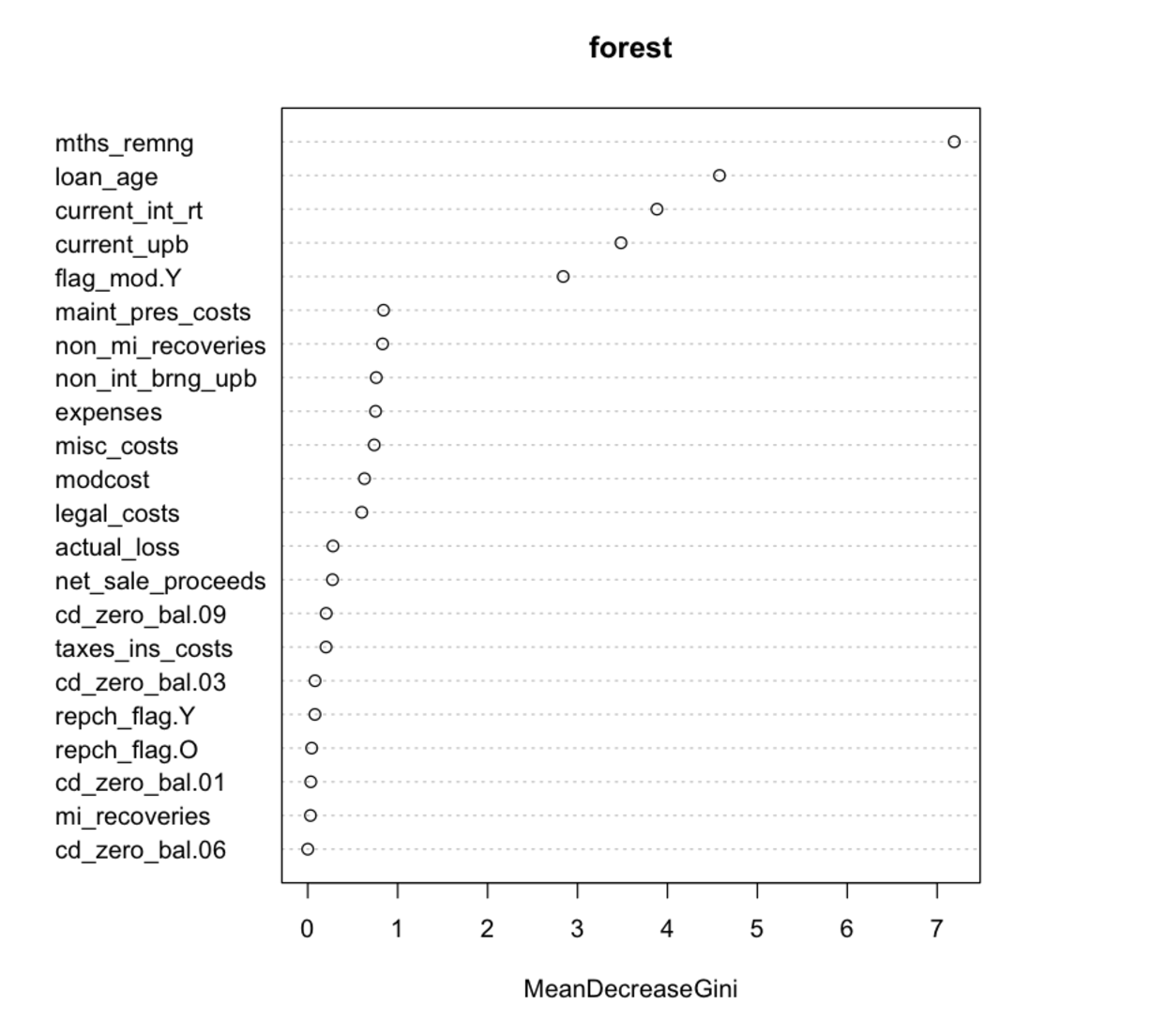
3. Random Forest:

a. Too slow, use 1% of sample

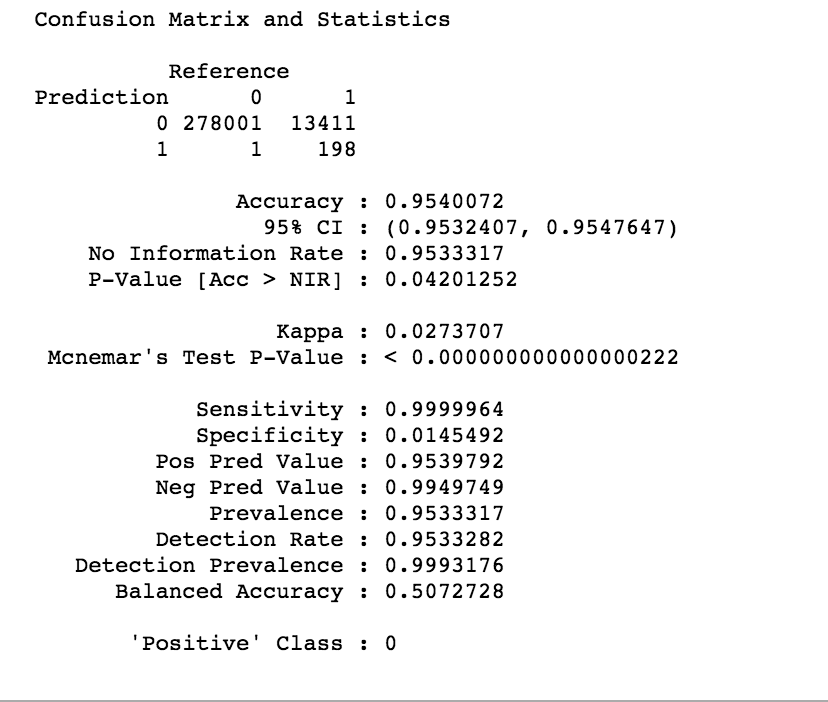
b. Confusion Matrix:



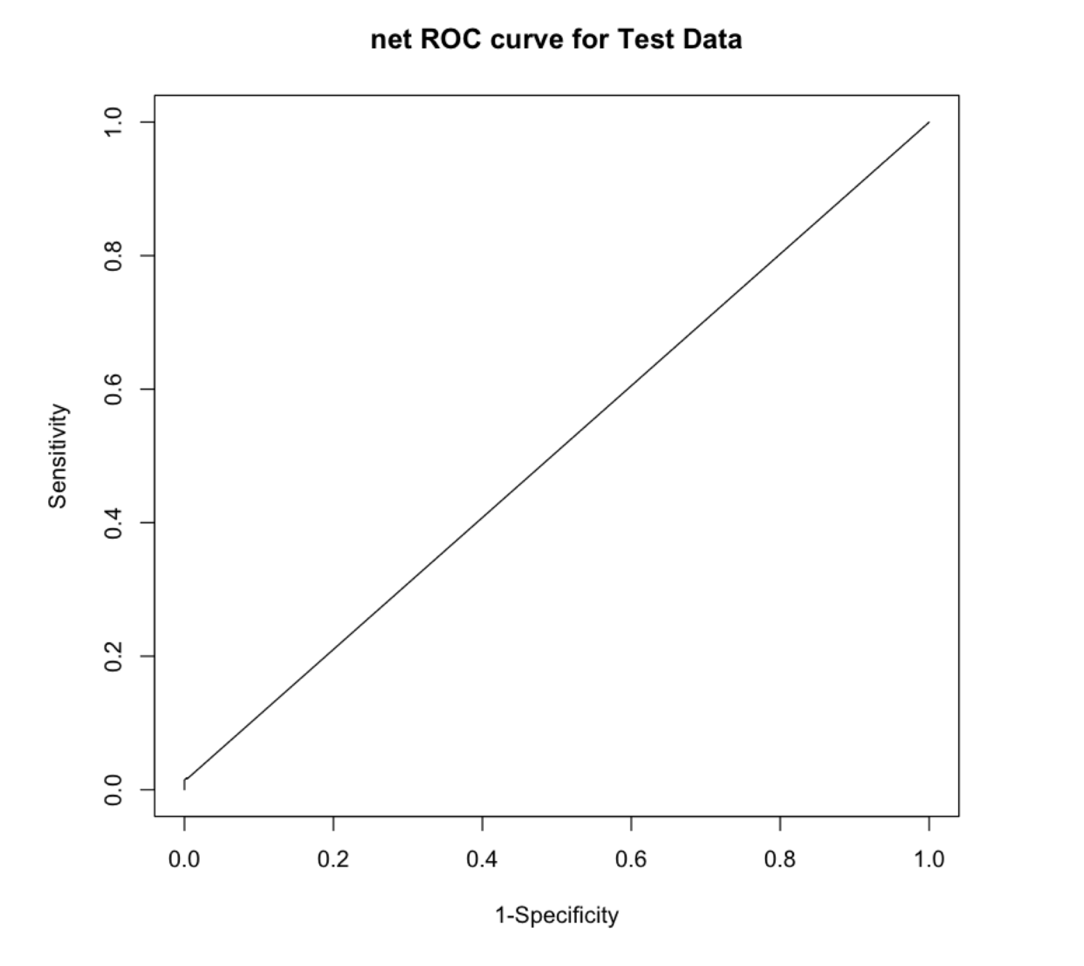
c. ROC:

b. Variable importance:

4. Neural Network:

a. Confusion Matrix:

b. ROC



5. Auto Classification for data from Q11999 to Q22016

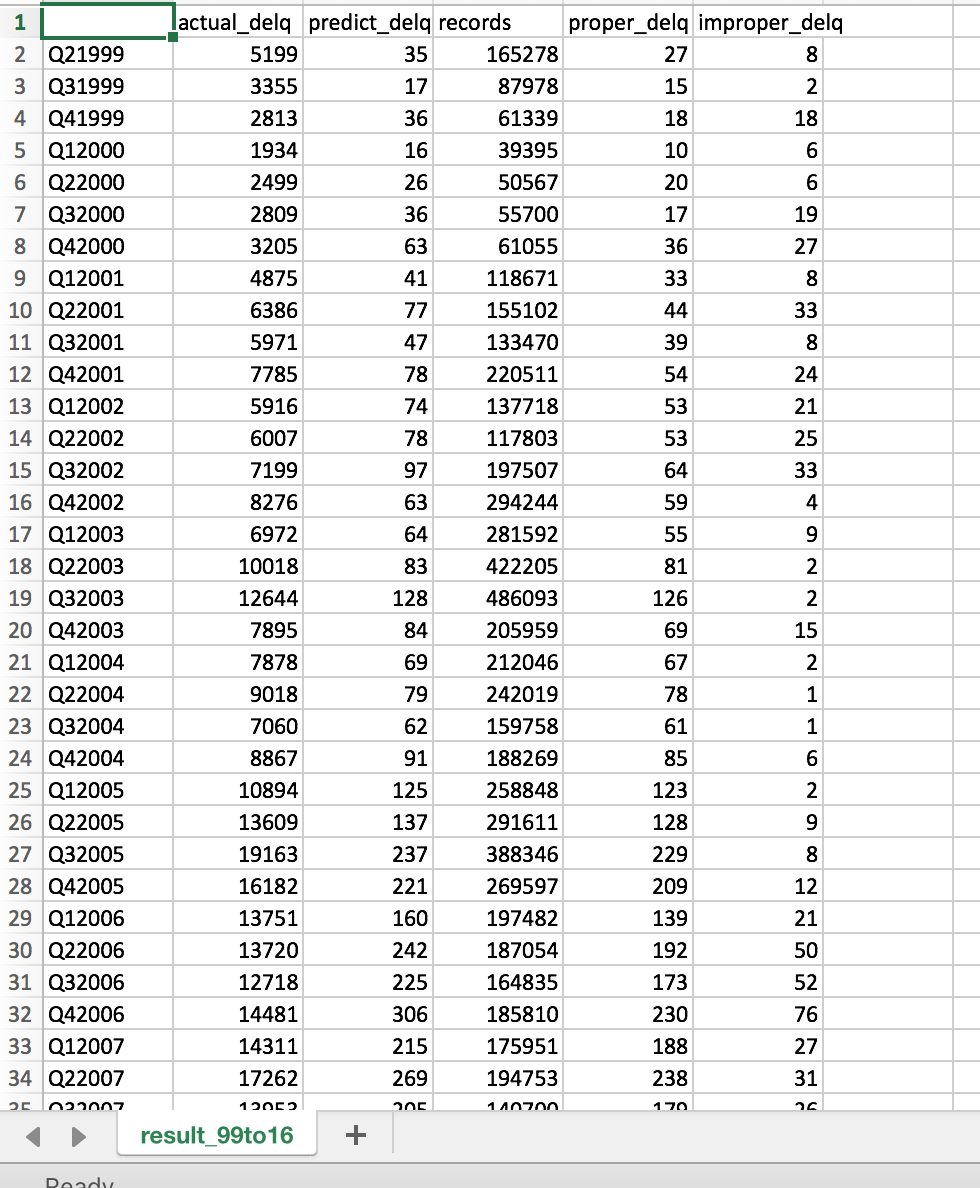
a. Best Model:

1. Algorithm: Logistic Regression

2. 6 Variables: current\_upb + loan\_age + mths\_remng + current\_int\_rt + non\_int\_brng\_upb + flag\_mod.Y

b. Result:

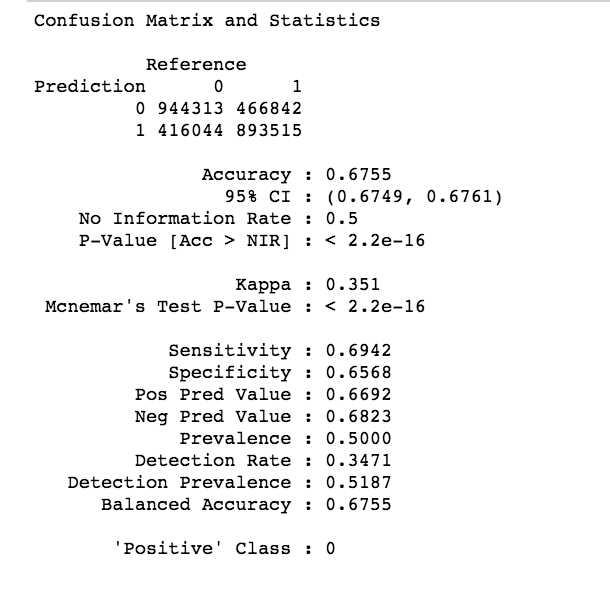
1. refer to “result\_99to16.csv”



c. Analysis:

1. All models are bad. I think it’s because that our data are unbalanced. Only less than 5% negative data (delinquent).

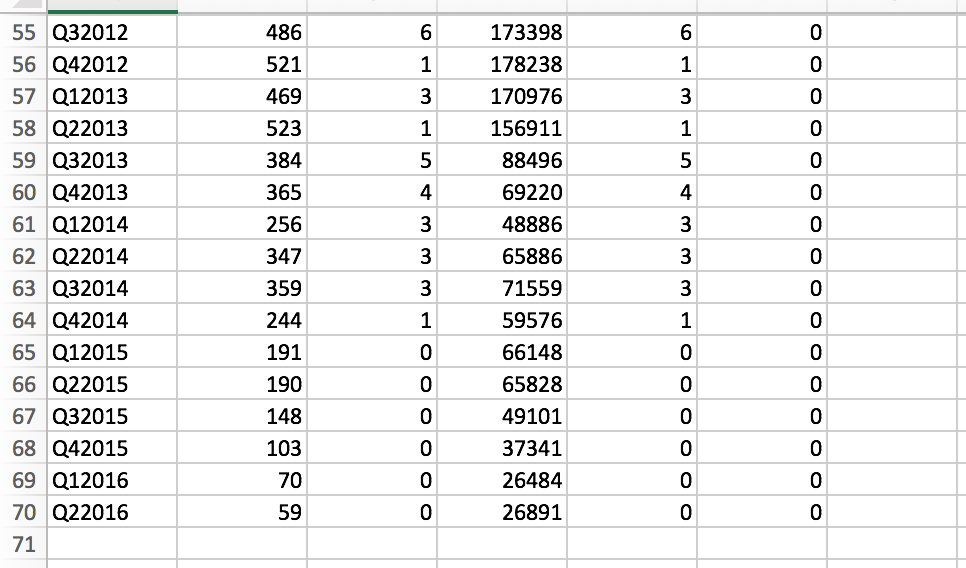
2. I don’t recommend using this model to predict. But I think we can pick a half-half sample from original dataset as the training dataset. That would improve the model: below is a half-half training dataset.



3. Metrics:

a. I use TPR and FPR to evaluate the model

b. If (TPR, FPR) is closer to (0, 1) or (1, 0), the model is better.



1. For Q12016:
   1. TPR: Sensitivity: 1
   2. FPR: 1 – Specificity: 1
   3. This model is really bad, I know.