Give Me Some Credit
CSCI 7350 Final project

Ning Zhang

Xi Chen

Outline

- Introduction
- Background
- Methodology
- Experiments
- Analysis & Discussion
- Conclusion

Introduction: Problem

- The goal of this project is to build a credit scoring model by predicting the probability of credit default in the future.

Data

- The training data contains 15,000 instances.
  - 10 attributes
  - Class: Default or not

Ten attributes

- Total balance on credit cards and personal lines of credit except real estate an no installment debt like car loans dived by the sum of credit limits
- Age
- Number of times borrower has been 30-59 days past due but no worse in the last 2 years
- Monthly debt payments, alimony, living costs divided by monthly gross income

Ten attributes

- Monthly income
- Number of Open loans (installment like car loan or mortgage) and Lines of credit
- Number of times borrower has been 90 days or more past due.
- Number of mortgage and real estate loans including home equity lines of credit

Ten attributes

- Number of times borrower has been 60-89 days past due but no worse in the last 2 years
- Number of dependents in family excluding themselves

Goals of our research

- Most significant features
- Most effective model(s)

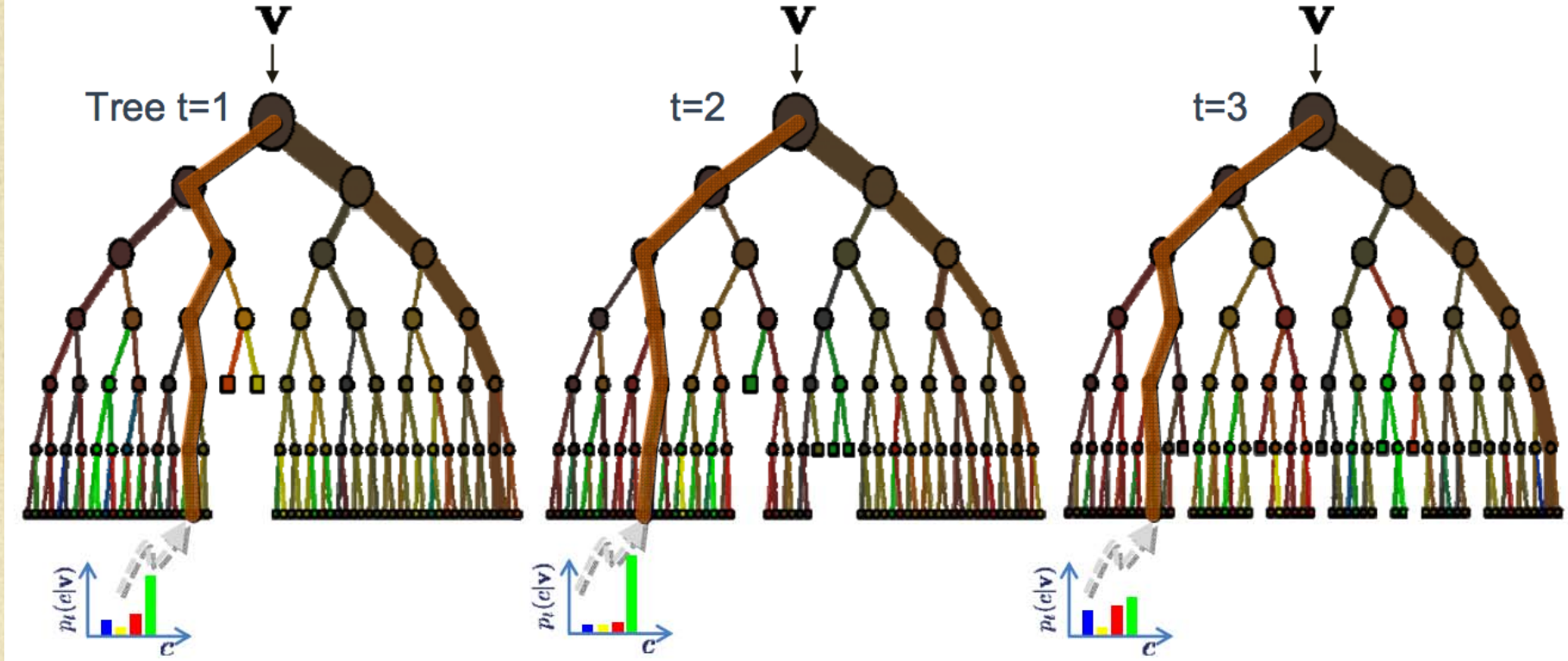
Background – First Place

- 2 Key Features:
  - total number of late days
  - difference between income and expense
- 5 Algorithms:
  - one random forest of classification trees
  - one random forest of regression trees
  - one classification tree boosting
  - one regression tree boosting
  - one neural network
- Result: 0.8695558 (AUC)

Methodology

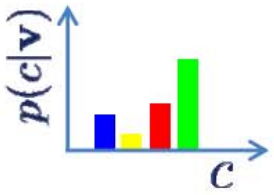
- Xgboost(boosting)
- Random forest
- Logistic regression
- KNN
- Neural Network

Random forest



The ensemble model

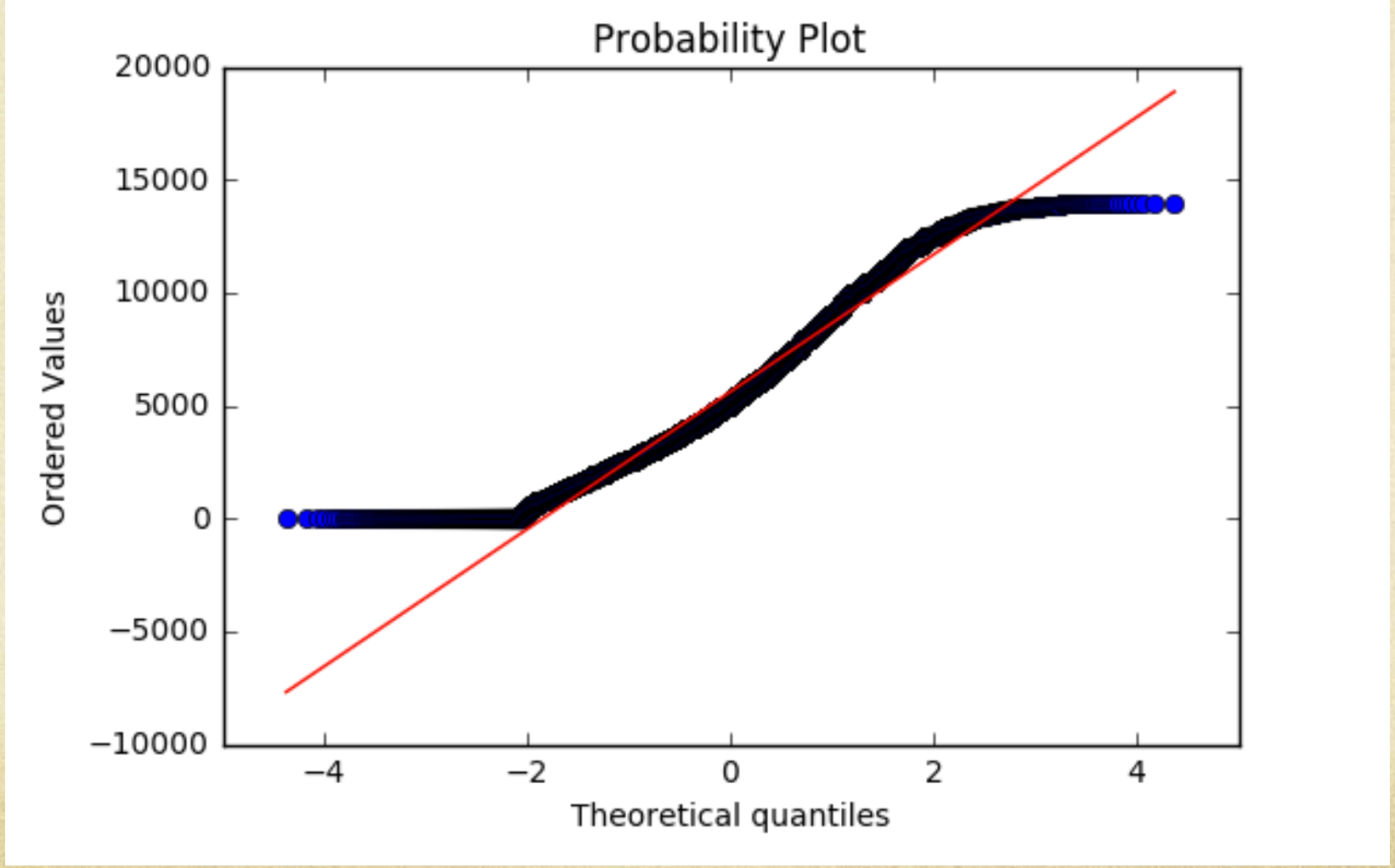
Forest output probability $p(c|\mathbf{v}) = \frac{1}{T}\sum_{t}^{T} p_t(c|\mathbf{v})$



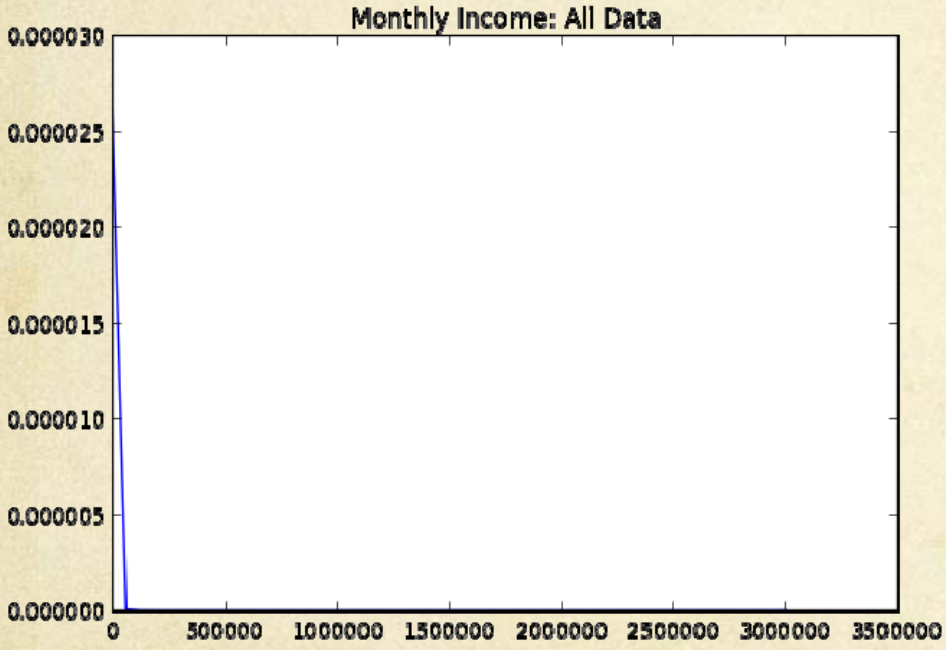
Steps

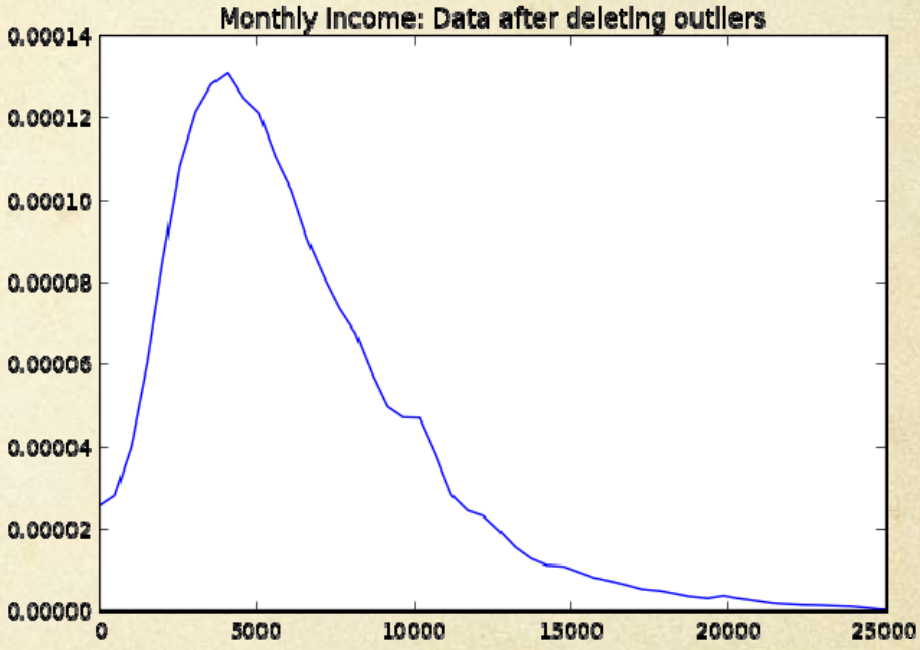
- Read data
- Preprocessing
  - Check the data distribution
  - Remove the outliers
  - Replace the NA value by the mean
- Use models
- Evaluate results

Data distribution

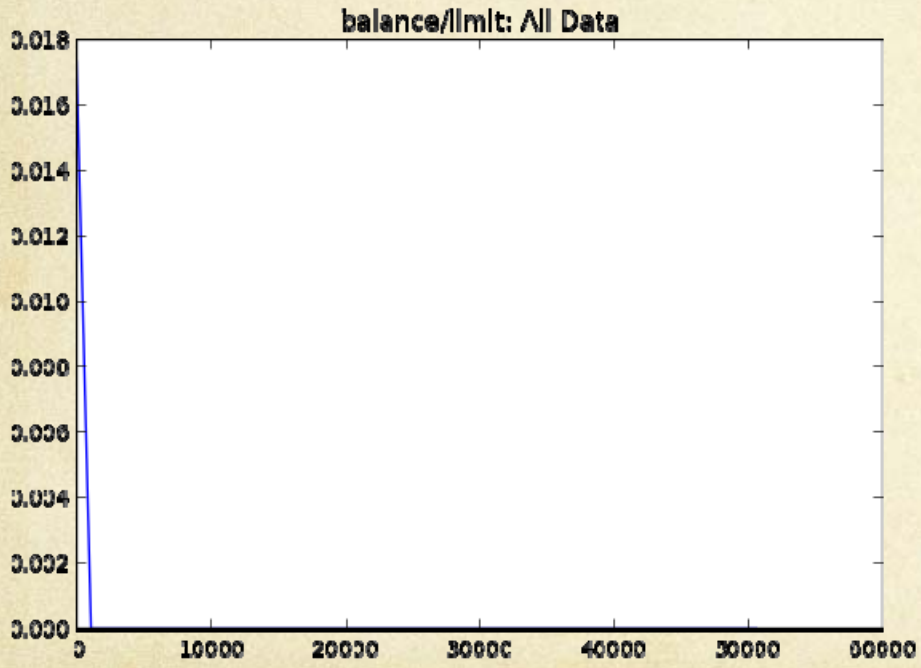


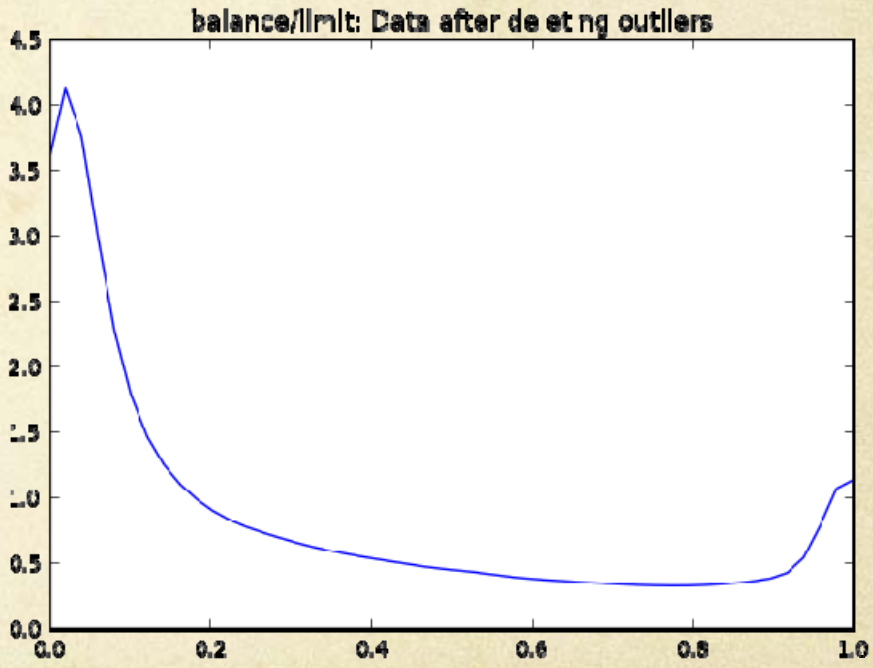
Remove outliers: monthly income



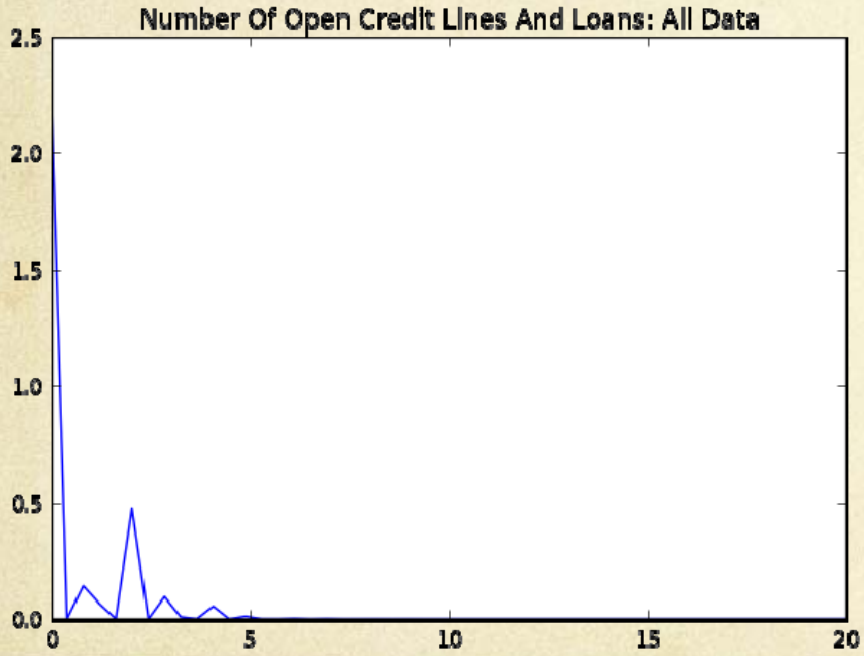


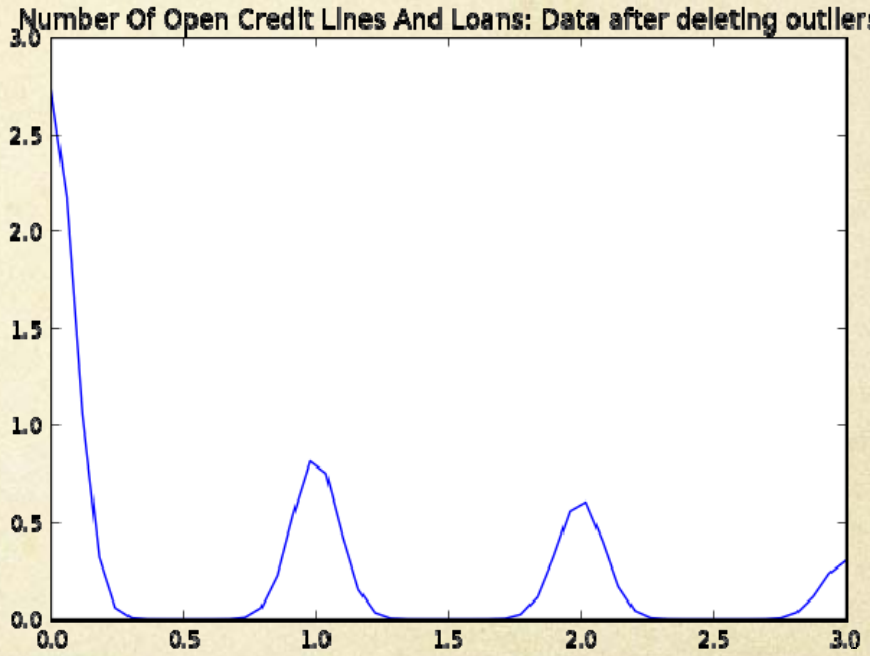
Remove outliers: balance/limit





Remove outliers: number of open credit lines and loans

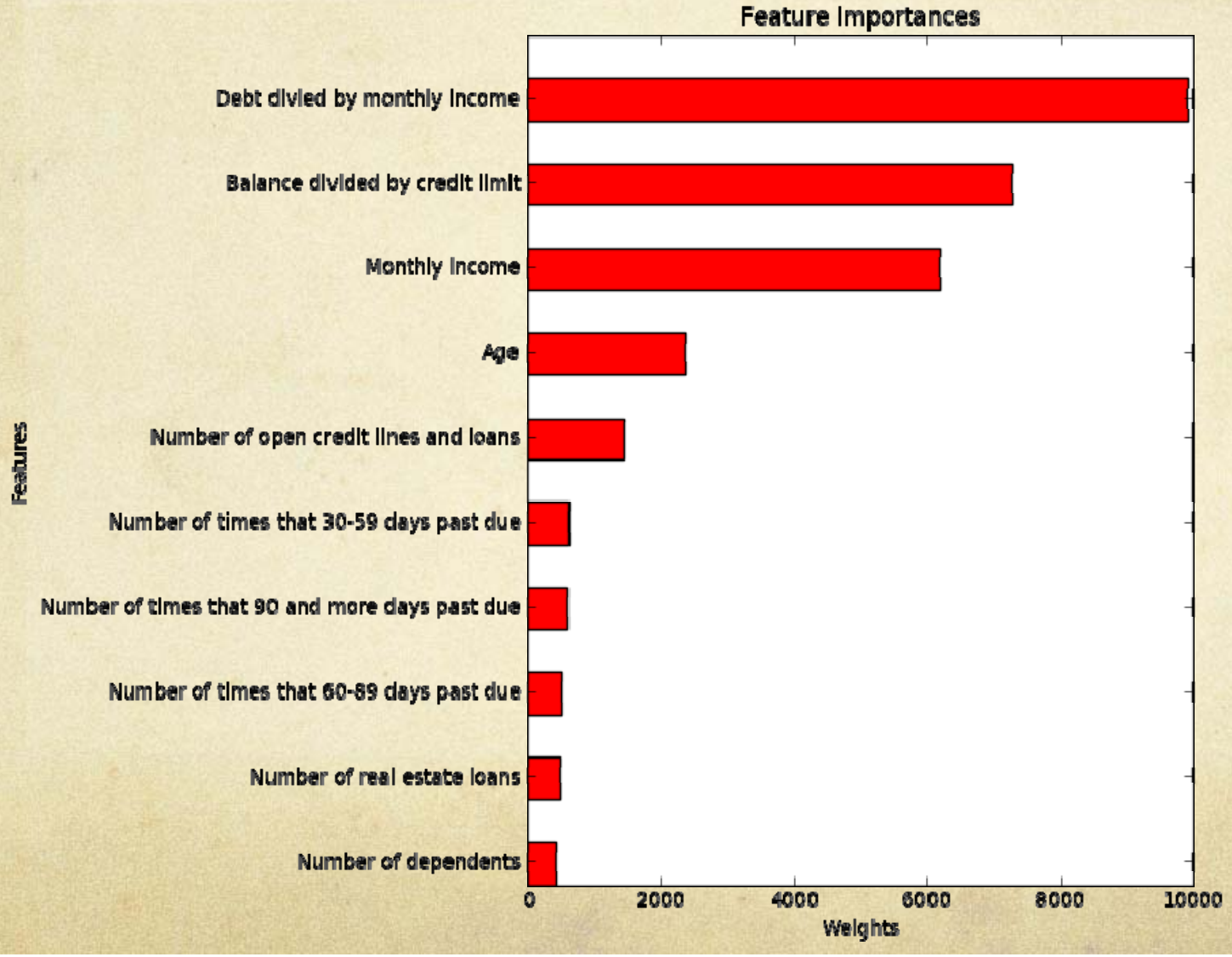




Results

Methods	Results
Xgboost (200)	0.86451 +/- 0.00564
Random forest tree (200)	0.84354 +/- 0.00739
Logistic regression	0.83315 +/- 0.03572
KNN (30)	0.62553 +/- 0.01060
Neural Network	0.50000 +/- 0.00001

The Importance of each attribute



Analysis & Discussion

- The most important features:
  - Debit divided by monthly income
  - Balance divided by credit limit
  - Monthly income
- The good performance of non-linear models (xgboost & random forest) are reasonable
- The good performance of linear model (logistic regression) is not expected

Analysis & Discussion

- Our result is pretty close to the first place winner.

Conclusion

- We applied several methods, and found boosting and bagging based methods are the most effective ones.
- Logistic regression can also achieve good results.
- The performance of neural network is highly relied on parameter tuning.
- Feature engineering is important.