Study on tree-based methods.

MATH 6380 project 2

Chenyang, DONG Tsz Cheung, LO Jiacheng, XIA

April 24, 2017

Outline

Table of Contents

Studying tree based methods...

Why did we choose tree-based methods?

Studying tree based methods...

Why did we choose tree-based methods?

- We went through several methods.

Studying tree based methods...

Why did we choose tree-based methods?

- We went through several methods.
- Tree-based methods are straight-forward and easy to implement.

Studying tree based methods...

Why did we choose tree-based methods?

- We went through several methods.
- Tree-based methods are straight-forward and easy to implement.
- There are yet many improvement methods.

Studying tree based methods...

Why did we choose tree-based methods?

- We went through several methods.
- Tree-based methods are straight-forward and easy to implement.
- There are yet many improvement methods.
- Studied the method on 3 datasets.

Table of Contents

The dataset

This dataset and the preprocessing are the same as project 1.

The dataset

This dataset and the preprocessing are the same as project 1.

Goal for this dataset

Do straightforward analysis and compare with Lasso (PJ 1).

What we found

In terms of MSE, simple regression tree(0.11) slightly worse than Lasso(0.06); bagging, random forest and boosting even better(0.04, 0.02).

Visualize the results

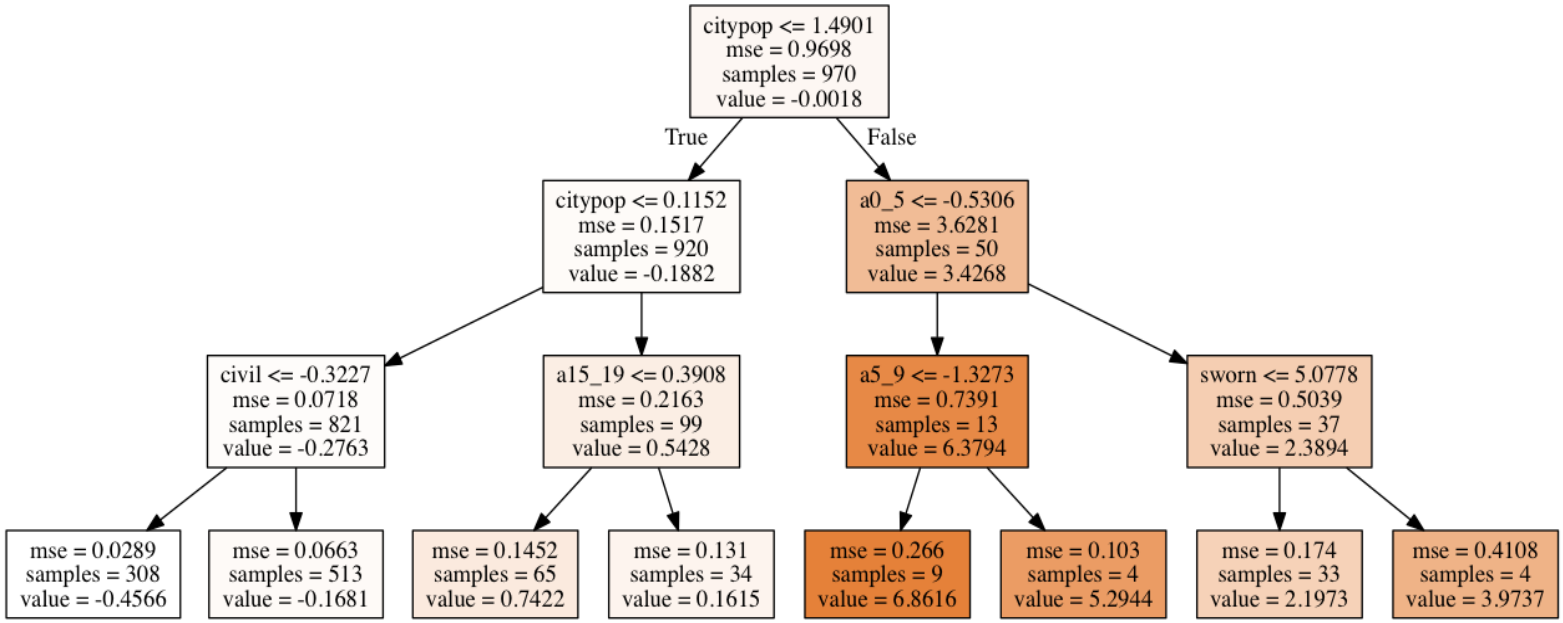


Figure: Regression tree on crime data

Boosting and random forests

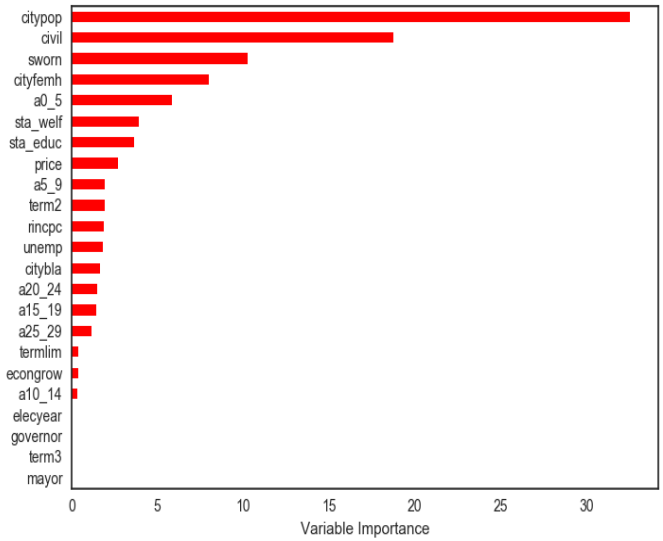


Figure: Importance from boosting

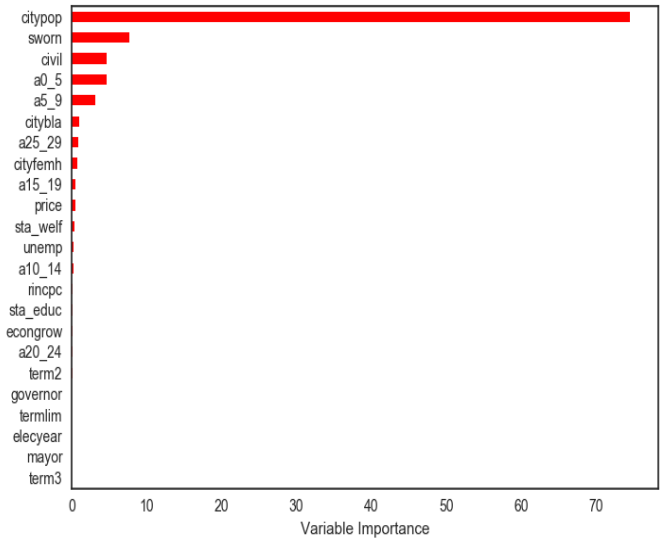


Figure: Importance from random forest

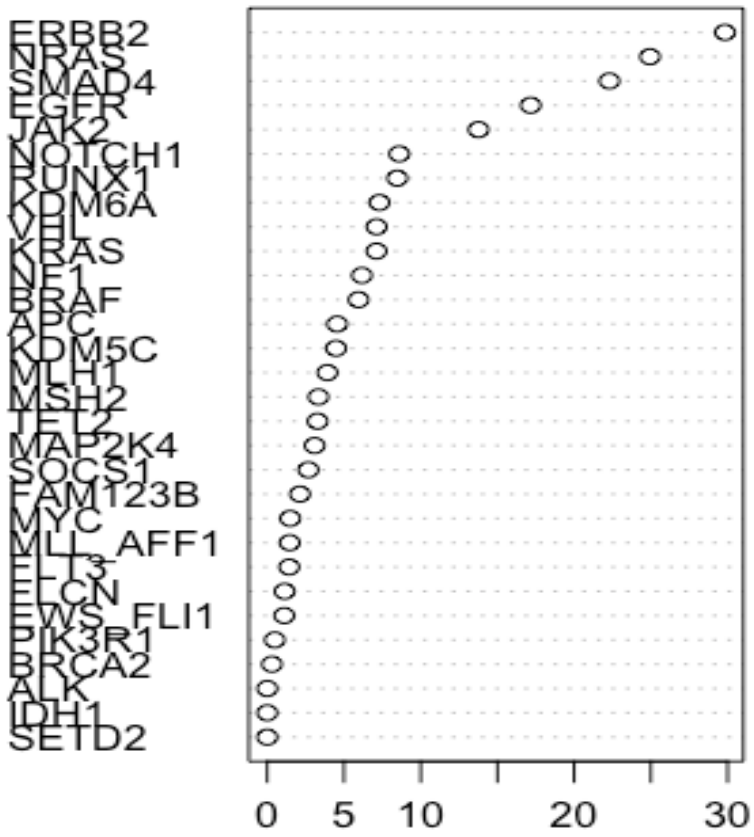
Table of Contents

Variable Importance

- The **R randomForest** package optionally produces 2 additional pieces of information. One is called **Variable Importance**, a measure of the importance of the predictor variables.
- The random forest algorithm estimates the importance of a variable by looking at how much prediction error increases when (OOB) data for that variable is permuted while all others are left unchanged.
- In this project, we define variables/predictors with $\%IncMSE > 0$ as important variables, then select them as our predictors in our final model.

Importanct Variables in BinaryDrug

The necessary calculations are carried out tree by tree as the random forest is constructed. Our experience has been that even though the variable importance measures may vary from run to run, the ranking of the importances is quite stable.



A potential variable selection method?

- We have tried *LASSO* in this dataset. Bad.
- Somebody introduced $p - value$ selection. Kaggle $MSE = 3.08$. Good.
- Can we do better using variable importance?

some V.I related models

- V.I + *MLR*, Kaggle *MSE* = 3.26057

some V.I related models

- V.I + *MLR*, Kaggle $MSE = 3.26057$
- V.I + *randomForest*, Kaggle $MSE = 3.23067$

some V.I related models

- V.I + *MLR*, Kaggle $MSE = 3.26057$
- V.I + *randomForest*, Kaggle $MSE = 3.23067$
- So sad.

Some Conclusions

We came up with some conclusions (inferences):

Some Conclusions

We came up with some conclusions (inferences):

1. dummy1

Some Conclusions

We came up with some conclusions (inferences):

1. dummy1
2. dummy2

Table of Contents