Final Report

Jie Gu

6/2/2019

Executive Summary

In STAT 301, I conducted a research on the dataset of a school survey on crime and safety. The candidate dependant variables are 22 kinds of crisis incidents data. I gave weighted to them after standardizing them then got on final dependant variable.

I split the entire data set into training set and testing set in a 9:1 ratio to varify the performance of various models. For the tuning parameters, I applied 4-fold cross validation.

I first built a linear regression model and got the MSE ~ 580. Then I built a Neural Network with 3 hidden layers, 500 nodes per layer and set epoch = 50. I got the MSE ~ 1. Then I built a Random Forests model with mtry = 12. I got the MSE ~ 420 . Therefore, Neural Network is the best model based on my dataset.

Finally, in order to find those most important variables, I built the Random Forests on the whole dataset and picked out the top-10 most important variables through `importance()` function.

Introduction

The dataset is a moderate-sized dataset, which is based on School Survey on Crime and Safety. It is provided by National Center for Education Statistics. The source of the dataset is: https://nces.ed.gov/edat/variableSelect.aspx?guide=&hnt=&srch=&rnd=197&agrmnt=1&sessionid=e772055d-b69a-40b8-9796-ef2c7903fb66

This dataset is about the security regulations and the number of violations in more than 2,000 schools. By studying this dataset, maybe I can know which measures can effectively reduce the occurrence of school accidents and thus protect students' safety. It's interesting and important.

Work Flow

Exploratory Data Analysis

This dataset is about the security regulations and the number of violations in more than 2,000 schools. Each school serves as an observation. The raw dataset has 414 columns, but half of the features are not statistically significant (such as "Collapsed STRATUM code", "Imputation flag" and "Jackknife replicate"), so I directly deleted those columns. Moreover, there are some features about the position of respondent, number of years respondent at the school, and the date questionnaire completed. I think they are the basic information of questionnaires and should have little effect on the response variables. Therefore, I didn't take them into consideration.

After preliminary selection, I leave 189 features. There are 2560 observations and no data is missing.

There are 152 factor variables and 38 numerical variables over all. The first variable is school ID. It is the unique school identifier.

Then the next 139 variables from c0110 to c0450 are various kind of security policies whether the school take or not. They are all categorical variables. Most of them only have levels 1 and 2, which represent "yes" and "no". There are 37 variables have levels 1, 2, and -1, which represent "yes", "no", and "legitimate skip".

The four variables from c0196 to c0202 are about parental participation, which have levels from 1 to 5, representing respectively “0-25%”, “25-50%”, “50-75%”, “75-100%”, and “the school doesn’t offer”.

The 13 variables from c0280 to c0304 are about effort limitation, which have levels 1, 2, and 3, representing “limits in major way”, “limits in minor way” and “does not limit”.

The 8 variables from c0374 to c0388 are about the frequency with which students make unethical behaviors such as race and gender. They have levels from 1 to 5, which represent the frequency from “happens daily” to “never happens”.

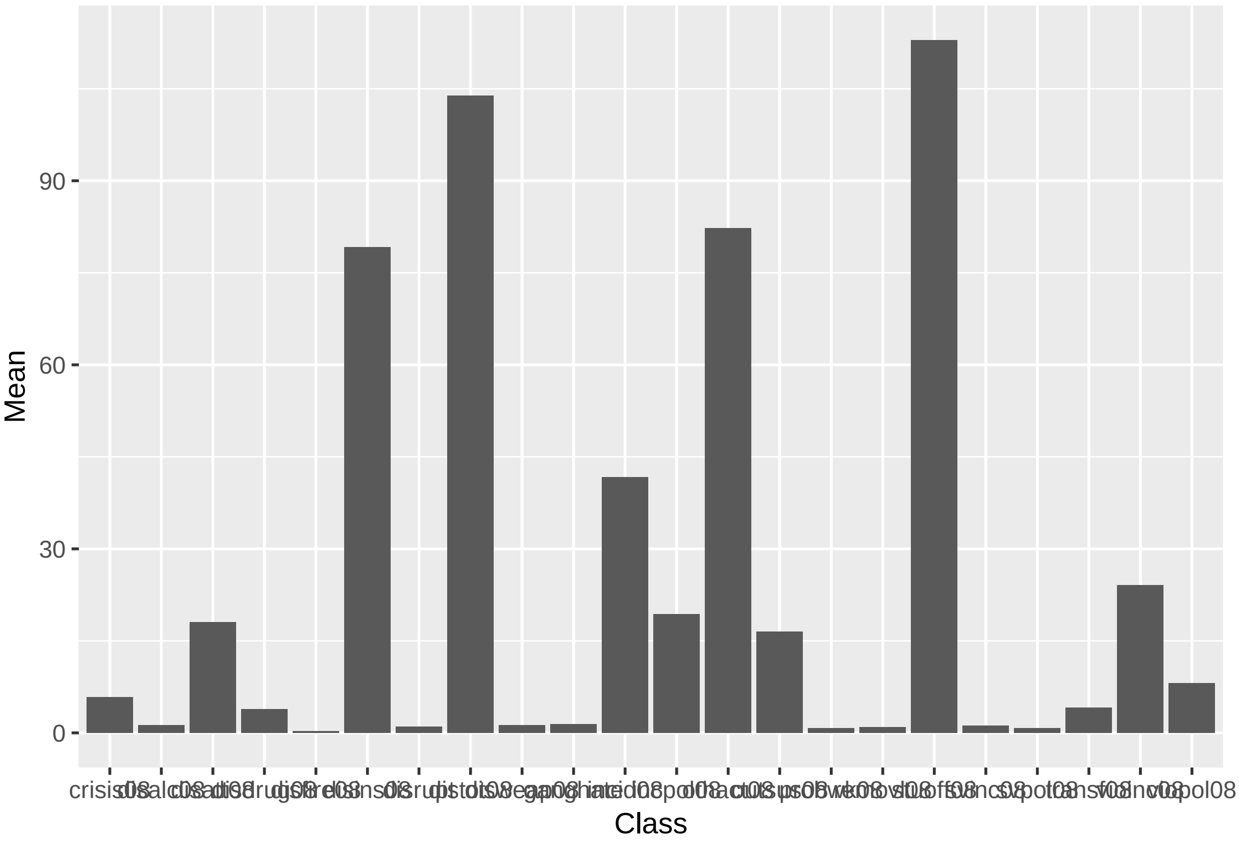
The 6 variables from c0540_r to c0558_r are about numbers of full-time or part-time faculty. Although they look like a numeric variable, the last level is “26 or more”, so they have to be viewed as factors.

Variables c0560 and c0562 are about crime level where students live and school located.

The last four categorical variables fr_catmn, fr_lvel, fr_size, and fr_urban are the basic information about the school, including “percentage of students like Black/African American”, “school grades offered”, “school size”, and “urbanicity”.

This dataset only has 38 numerical variables. There are 16 numerical variables from c0508 to c0572. They introduce the numerical information about the students and teachers.

There are 22 numerical variables from ‘crisis08’ to ‘disrupt’ about the number of all kinds of terrorist problem incidences recorded or reported. In my opinion, they can be viewed as the response variables. This is the bar graph of mean values of these 22 variables.



Data Pre-processing

But there is a problem, it will be too many to choose 22 variables as the dependent variables, which leads to a lot of repetition, and it will be difficult to analysis. Therefore, I selected out 12 most representative variables and gave them a weight respectively after standardizing. I distributed 20 questionnaires and then got a comprehensive severity ranking. The weight gave according to the ranking. The order of severity from high to low is like this:

12’ - GANGHATE - Total number of gang-related and hate crimes.

11’ - DISFIRE08 - Total number of disciplinary actions recorded for use or possession of a firearm or explosive device.

10’ - DISWEAP08 - Total number of disciplinary actions recorded for use or possession of a weapon other than a firearm or explosive device.

9’ - DISDRUG08 - Total number of disciplinary actions recorded for distribution, possession, or use of illegal drugs.

8’ - DISATT08 - Total number of disciplinary actions recorded for physical attacks or fights.

7’ - VIOINC08 - Total number of violent incidents recorded.

6’ - DISALC08 - Total number of disciplinary actions recorded for distribution, possession, or use of alcohol.

5’ - TRANSF08 - Total transfers to specialized schools for specified offenses.

4’ - OTHACT08 - Total ‘other actions’ for specified offenses.

3’ - REMOVL08 - Total removals with no continuing school services for specified offenses.

2’ - DISTOT08 - Total number of disciplinary actions recorded.

1’ - DISRUPT - Total number of disruptions.

My main job is to do regression. I split the entire data set into training set and testing set in a 9:1 ratio to varify the performance of various models.

Model Selection

Linear Rregression Model

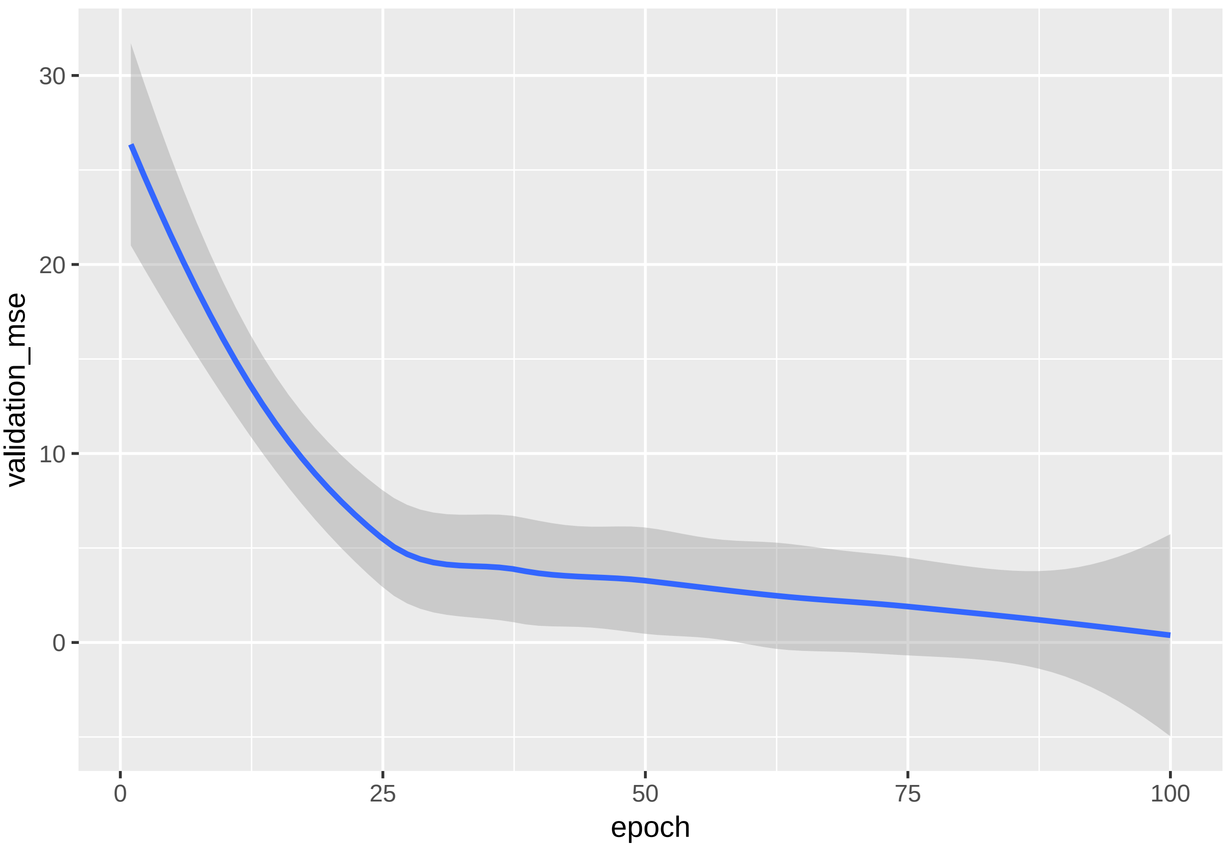
Firstly, I tried to build the simplest linear regression model. I standardized the numerical predictor variables and built the linear regression model of the weighted-sum output on the training set then compute the RMSE on the testing set.

```
## # A tibble: 1 x 2
##   Linear_regression   MSE
##   <list>            <dbl>
## 1 <S3: lm>           586.
```

Neural Network

Then I tried Neural Network.

Beacuse I have 167 independent variables and much more variables after one hot encoding, I decided to build a large neural network of 3 hidden layers with 500 nodes per layer to give it more flexibility. I split the training set into 4 folds to do k-fold cross validation. Meanwhile, I drew the plot of mse history to decide the number of epochs in case of overfitting.



I chose the epoch where the changing trend of validation mse gets flat, which is around 50. Then, I rebuilt the final neural network on the entire training set with the epoch number selected through cross-validation, and then verified the MSE on the testing set.

```
## # A tibble: 1 x 2
##   Neural_Network                              MSE
##   <list>                                    <dbl>
## 1 <S3: keras.engine.sequential.Sequential> 0.873
```

Obviously, Neural Network performs much better than linear regression.

Random Forests

I tried to use Decision Trees, because it has a function called `importance()`, that can tell me the importance of each variable. I built the Random Forests model and used 5-fold cross-validation on mtry from 1 to the number of variables, i.e. 167. But what's frustrating is that my laptop has been running continuously for more than 24 hours without finishing. Maybe it's because I have too many variables, I have to give up cross validation, and directly use the square root of dimention, i.e. 12. The MSE value of Random Forests model is like this.

```
## # A tibble: 1 x 2
##   RandomForests                 MSE
##   <list>                      <dbl>
## 1 <S3: randomForest.formula>  422.
```

Unsurprisingly, the MSE value of Random Forests is larger than Neural Network but smaller than linear regression model. I could control other tuning parameters like node size and number of trees, but I guess the MSE value is hard to become smaller than Neural Network and it's time-consuming to do cross-validation on those tuning parameters. Therefore, I didn't continue polishing this Random Forests model.

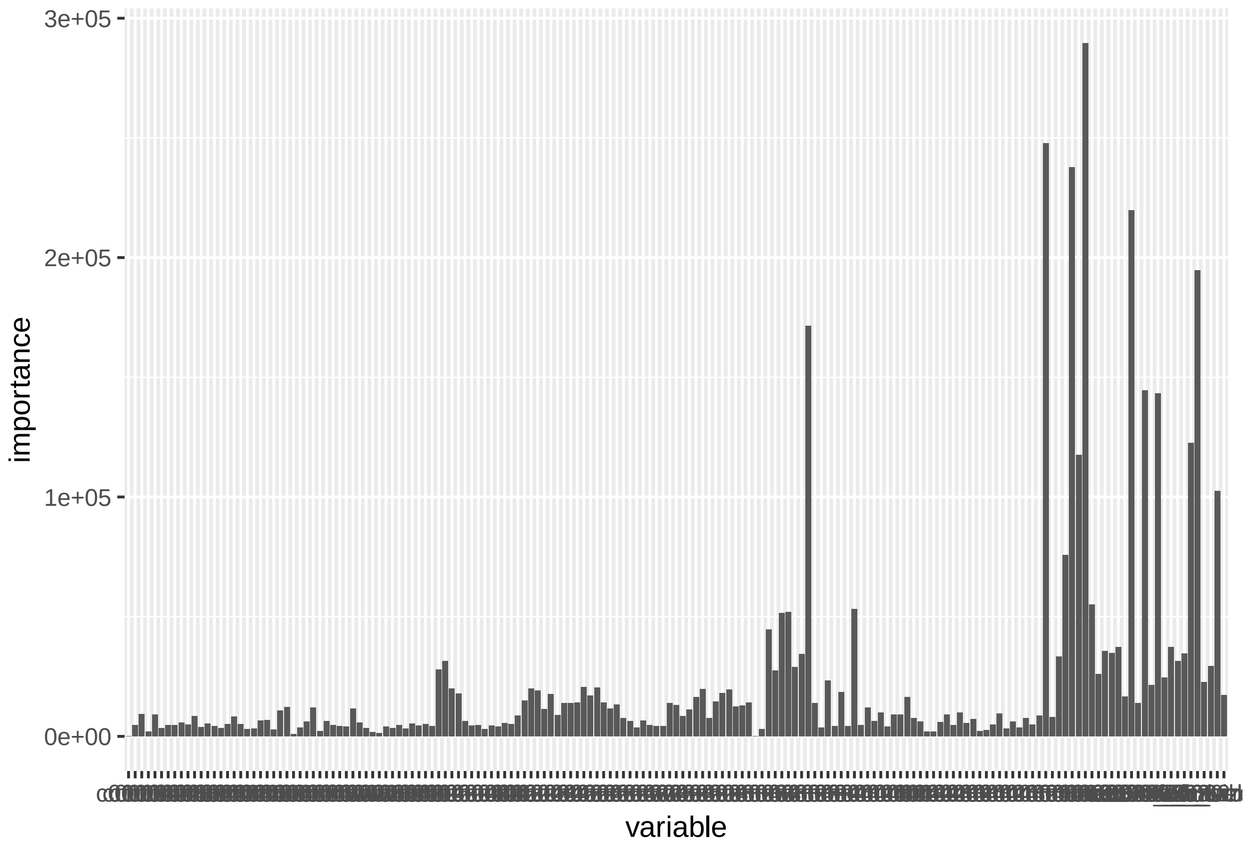
Variables Analysis

But it doesn't matter, because my main purpose is to use the decision trees to get the order of importance of the variables. I built the final Random Forests model on the whole dataset then got the rank of importance of all these 167 variables.

This is the top 10 important variables of all 167 variables.

```
## # A tibble: 10 x 2
##    variable importance
##    <fct>         <dbl>
##  1 c0520       289538.
##  2 c0508       247685.
##  3 c0516       237700.
##  4 c0540_r     219707.
##  5 c0572       194699.
##  6 c0386       171516.
##  7 c0544_r     144617.
##  8 c0556_r     143344.
##  9 c0570       122567.
## 10 c0518       117522.
```

I drew the bar plot of the importance of all these 167 variables. There are exactly 10 variables that have importance greater than 1e+05.



I standardized all the predictor variables before building the model, so these 10 variables can have certain practical significance. It is necessary to pay attention to these 10 aspects.

1. C0520 - # of transfers to specialized schools-total
2. C0508 - # of students involved in insubordination-total
3. C0516 - # of other actions for insubordination
4. C0540_R - # of paid full-time special ed teacher
5. C0572 - # of students transferred from school
6. C0386 - How often student gang activities
7. C0544_R - # of paid full-time special ed aides
8. C0556_R - # of paid full-time counselors
9. C0570 - # of students transferred to school
10. C0518 - # of removals with no service-total

Conclusion

In the draft report, I applied Principle Components Analysis to extract two outputs out of the 22 candidate dependent variables. But Jacob said the PCA can only exolain large proportion of variance. It's better to give them a weight according to people's ideas. Thanks for the great idea which not only reduced my workload, but also made my report more readable.

In the draft report, I didn't use cross validation to select mtry value for Random Forests because it took too much time. I compromised using the Bagging. However, Walker said I can directly choose mtry equals the square root of the variables dimension. What's more, he recommended me to enlarge my neural network to improve my regression accuracy. There two great advices helped me a lot.

To improve my model, I was wondering if I could combine several variables that describe similar attributes into a single variable and then use the integrated variables for regression. Because when analyzing the importance of each variable, for an individual variable, its importance can be small, but if add up with associated variables, the importance may become large. I tried this in the draft report, but I felt like that my combining method was too subjective. Plus the time complexity of my models were greatly reduced after adopting others' advices, I gave up on this in the final draft. I will try it if I encounter a relatively objective algorithm for combining variables in the future.