Cancer Dataset Analysis Report

Xi Lu 15420170155243

mails: luxialan12@gmail.com

Introdution

In this paper, we analyze the cancer dataset. The whole report can be divided into five parts:

1. Data Analysis
2. Model Evaluation and Selection
3. Conclusion
4. Supplementary(Code)

We will introduction the details of every part in the data analysis trip. Some explorations may be of little help to get the final correct answer, but we hope those materials will help the reader to understand the data analysis procedure better. If you are not interested in the details, just jump to the conclusion part to get the final result.

1 Data Analysis

In this part, we analyze the cancer dataset. Concretely, the whole procedure can be divided into three part:

1. Data Description
2. Data Exploration

*Single Variable Exploration

*Variable Correlation Exploration

1.1 Data Description

A cancer dataset has been deidentified, recoded, and contained in “recoded data.txt”. Consider for example the first row:

```
[1,]   4.103   2.177   NA   0.808
```

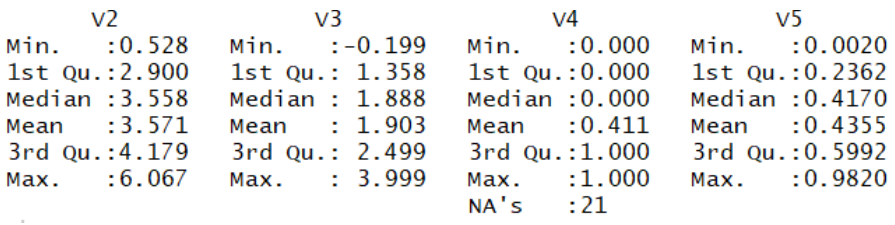
The name of the five variable is "V1" "V2" "V3" "V4" "V5". [1,] is apparently useless; the first variable (4.103) is the response variable and continuously distributed; the rest three (2.177 NA 0.808) are covariates and potentially associated with the response variable. Among the three covariates, the first and third are continuously distributed, and the second is binary and may contain missing values.

1.2 Data Exploration

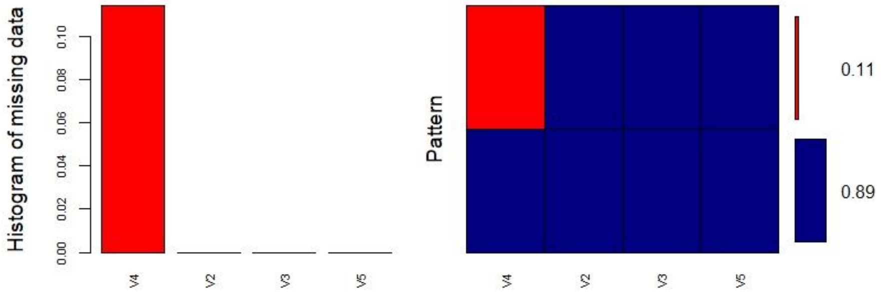
In this section, we try to dig out more information from the cancer dataset. We fist look at the statistics information of each variables and try to clean the data. Moreover, we visualize the statistical properties of the single variable and further explore the relationship between the variables in the dataset.

1.2.1 Single Variable Exploration

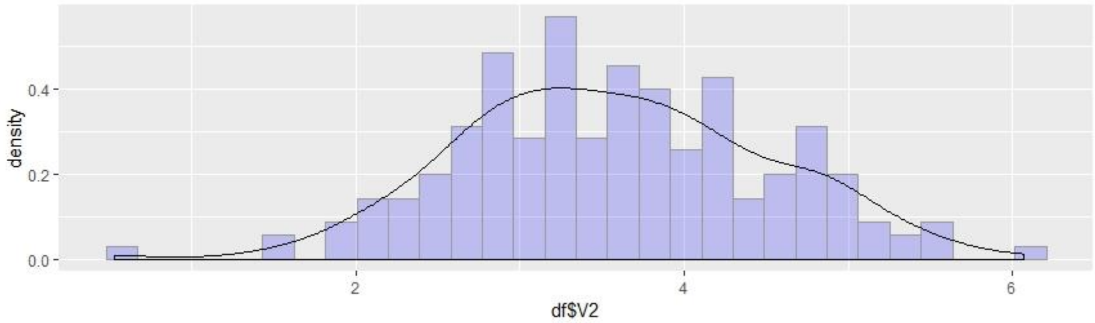
We first look at the statistics indicator of each variable.

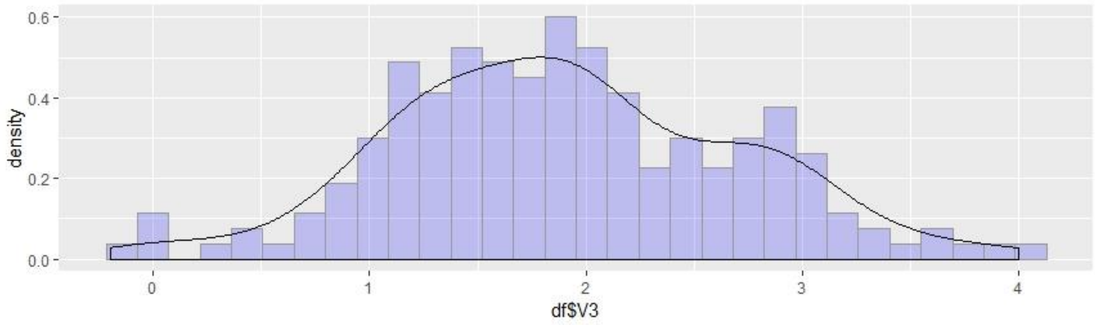


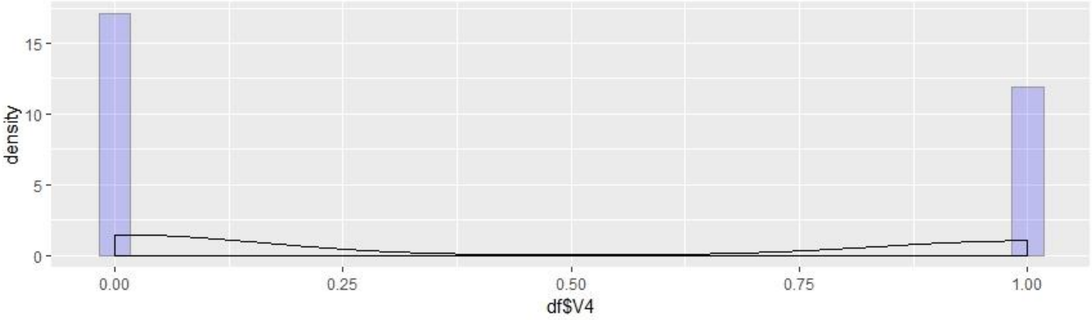
```
       V2                V3                 V4                V5
 Min.   :0.528   Min.   :-0.199   Min.   :0.000   Min.   :0.0020
 1st Qu.:2.900   1st Qu.: 1.358   1st Qu.:0.000   1st Qu.:0.2362
 Median :3.558   Median : 1.888   Median :0.000   Median :0.4170
 Mean   :3.571   Mean   : 1.903   Mean   :0.411   Mean   :0.4355
 3rd Qu.:4.179   3rd Qu.: 2.499   3rd Qu.:1.000   3rd Qu.:0.5992
 Max.   :6.067   Max.   : 3.999   Max.   :1.000   Max.   :0.9820
                                  NA's   :21
```

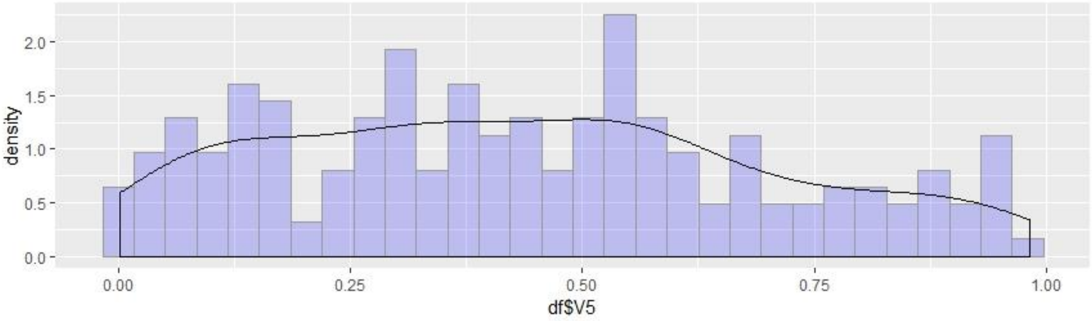


Obviously, there are some miss value in “V4”, the missing rate is 11%, which cannot be ignored directly. Since the value of “V4” is binary, we use the existing data of “V4” to estimate the parameter of binomial distribution, and sample from such distribution to fix the missing value. We look at the histogram with density curve to get a more direct impression on each variable.



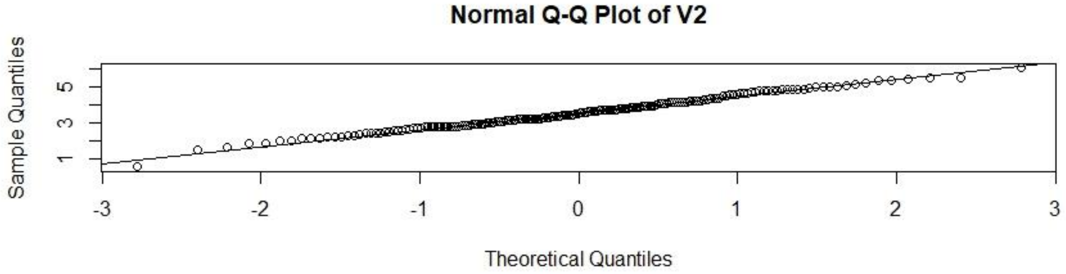


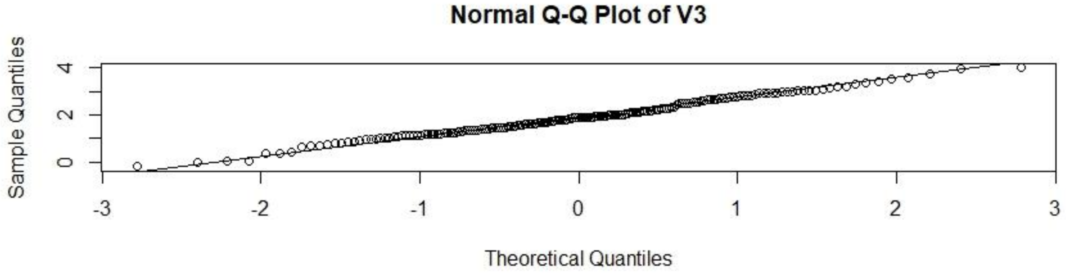


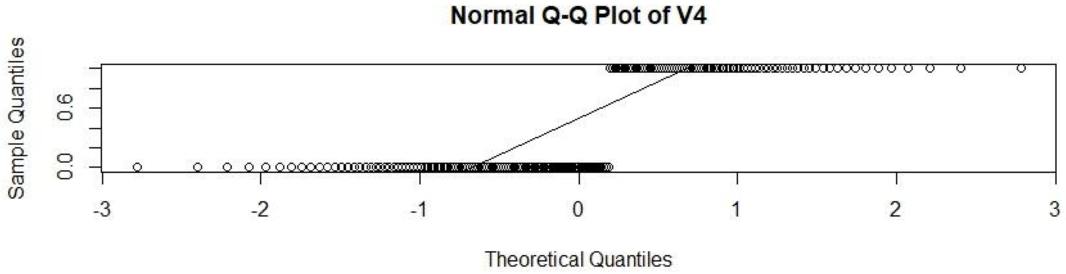


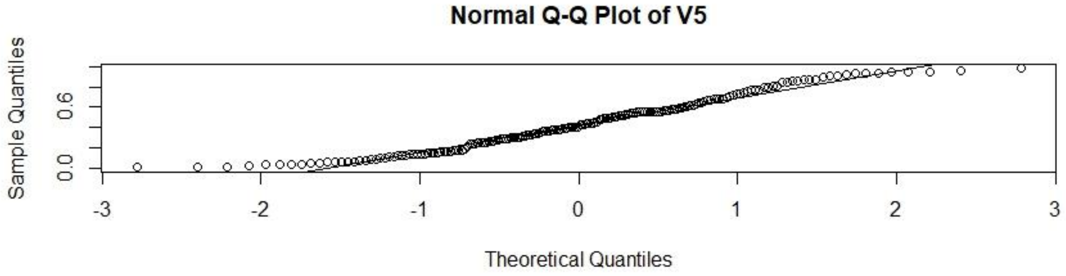
It seems that "V2" "V3" follows the normal distribution, while we are not sure about "V5". We further verify the normality of the variable by using Shapito-Wilk test and Q=Q plot.

```
        Shapiro-Wilk normality test

data:  df$V2
W = 0.99359, p-value = 0.6055

        Shapiro-Wilk normality test

data:  df$V3
W = 0.99215, p-value = 0.4225

        Shapiro-Wilk normality test

data:  df$V4
W = 0.6248, p-value < 2.2e-16

        Shapiro-Wilk normality test

data:  df$V5
W = 0.96522, p-value = 0.0001551
```





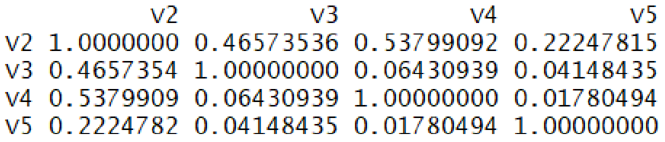




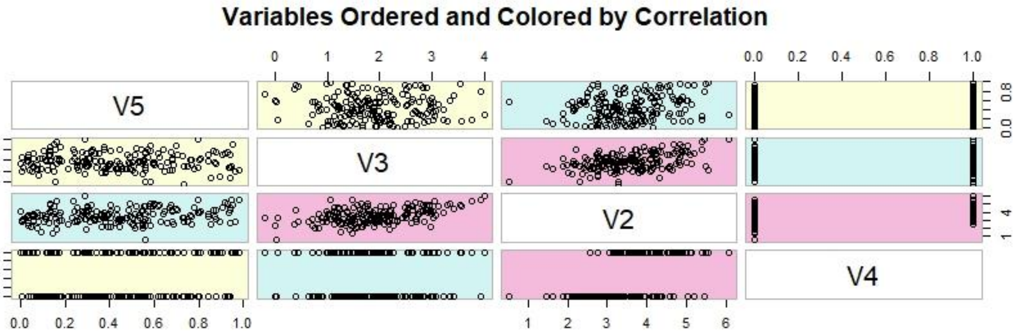
1.2.2 Variable Correlation Exploration

In this section, we try to get some insight about the relation between the variables in prepare to build a suitable model for the problem. First, we look at the correlation between the variables.

Correlation Matrix

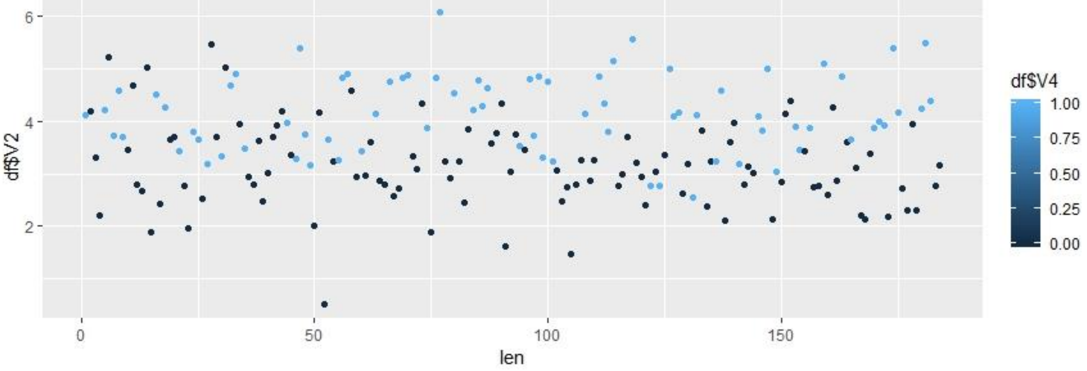


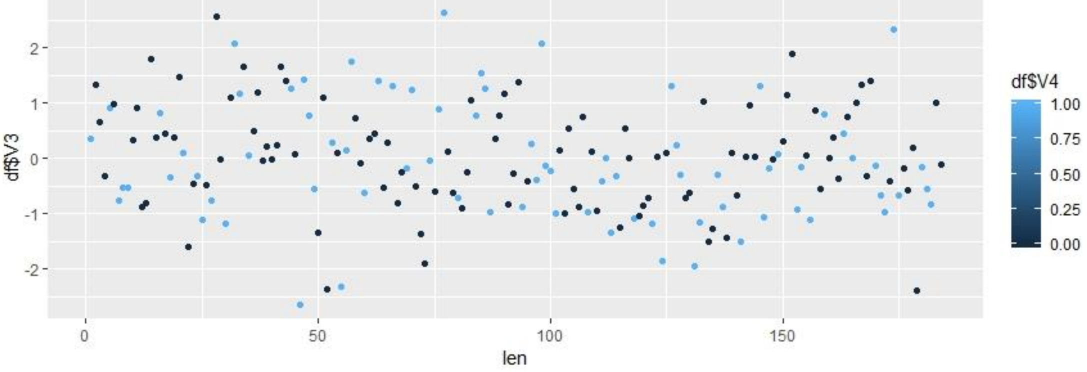
```
           V2         V3         V4         V5
V2 1.0000000 0.46573536 0.53799092 0.22247815
V3 0.4657354 1.00000000 0.06430939 0.04148435
V4 0.5379909 0.06430939 1.00000000 0.01780494
V5 0.2224782 0.04148435 0.01780494 1.00000000
```

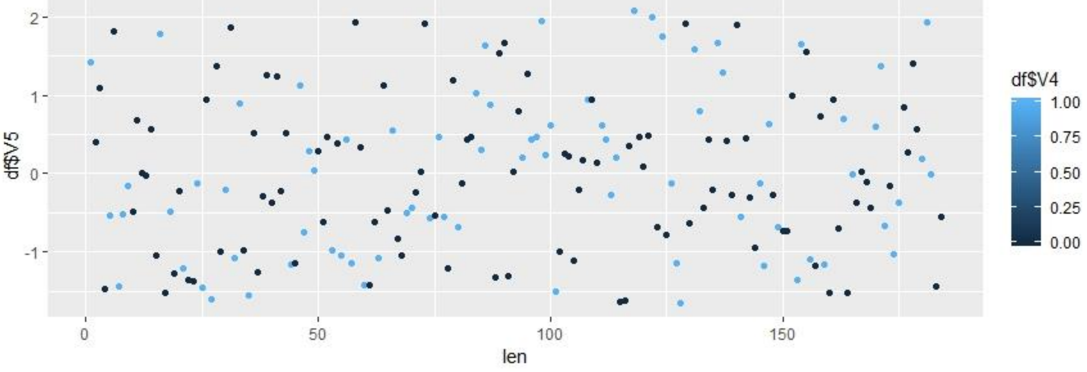


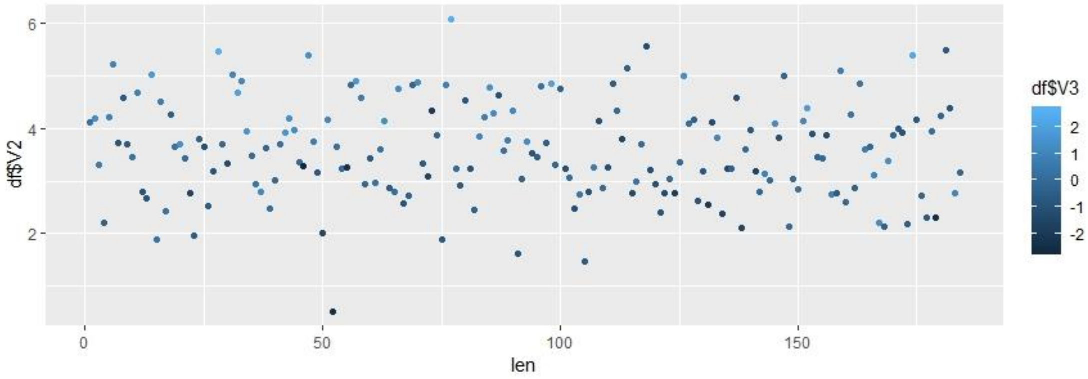
Variables Ordered and Colored by Correlation

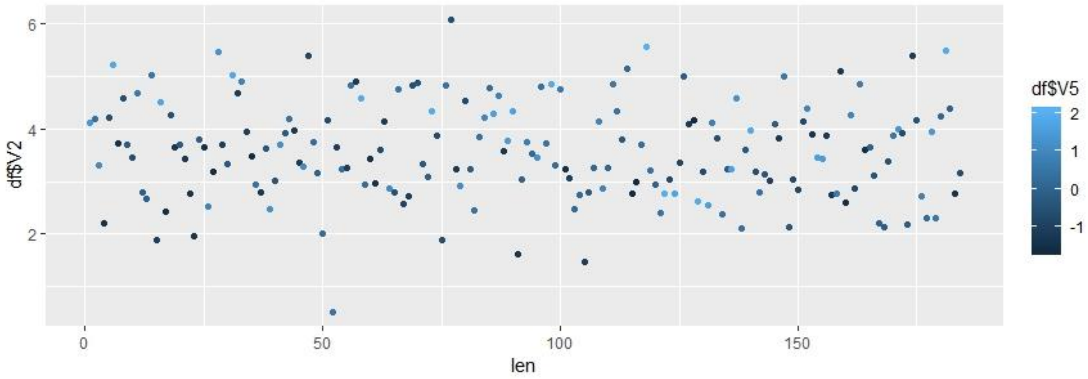
It seems that the “V4”、“V3” is more correlated with “V2”, while “V5” is less correlated with “V2”. Besides, “V3”、“V4”、“V5” is almost independent with each other. Since “V4” is a binary variable, which can be treated as a class label, we try to visualize such class information.







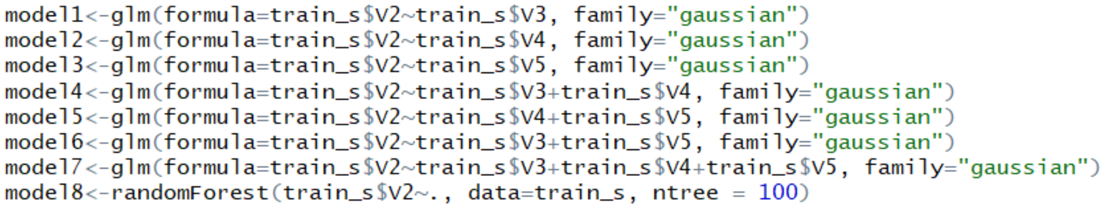




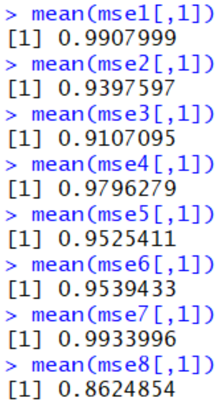
Amazing, it seems that “V2” can be linearly separated with the help of “V4”.While, “V3”、“V4” and “V5” do not seem to have linear relationship.

2 Model Evaluation and Selection

In this section, we consider 8 type of model and using 5-fold cross valuation to select the best model.



```
model1<-glm(formula=train_s$V2~train_s$V3, family="gaussian")
model2<-glm(formula=train_s$V2~train_s$V4, family="gaussian")
model3<-glm(formula=train_s$V2~train_s$V5, family="gaussian")
model4<-glm(formula=train_s$V2~train_s$V3+train_s$V4, family="gaussian")
model5<-glm(formula=train_s$V2~train_s$V4+train_s$V5, family="gaussian")
model6<-glm(formula=train_s$V2~train_s$V3+train_s$V5, family="gaussian")
model7<-glm(formula=train_s$V2~train_s$V3+train_s$V4+train_s$V5, family="gaussian")
model8<-randomForest(train_s$V2~., data=train_s, ntree = 100)
```



```
> mean(mse1[,1])
[1] 0.9907999
> mean(mse2[,1])
[1] 0.9397597
> mean(mse3[,1])
[1] 0.9107095
> mean(mse4[,1])
[1] 0.9796279
> mean(mse5[,1])
[1] 0.9525411
> mean(mse6[,1])
[1] 0.9539433
> mean(mse7[,1])
[1] 0.9933996
> mean(mse8[,1])
[1] 0.8624854
```

As a result, the best model is the randomforest has the minimax mse. So it is the best.

3 Conclusion

In this report, we investigate the cancer dataset. We first analyze the dataset, explore the statistics information of the dataset. And we use several generalized linear model and randomforest to fit the data. The k-fold cross validation shows that the randomforest has the best performance.