Prelim Report based on 2008 BSA Impatient Claims dataset

Yongkai Qiu

2019/3/6

Contents

Introduction and Data summary

Introduction

This preliminary (pilot) study illustrates my ability to handle the complex heterogeneous Big Healthcare dataset. My dataset came from a public database provided by The Centers for Medicare & Medicaid Services (CMS). I've taken one of the dataset **CMS 2008 BSA Inpatient Claims PUF** to create this report. This dataset contains information on hospital claims for the inpatient services provided to a 5% sample of 2008 Medicare beneficiaries. It is a subset of the 2008 Medicare beneficiaries database and it contains 7 different variables and 588415 samples. I took this dataset for the report for two reasons: 1. it is a large dataset, which maybe similiar to the dataset I may work on in the future program. 2. It is a public data so it is easy to get access to. However as it is only a subset dataset of a bigger database. So It has some limits, such as it only contains 7 variables, which will definitely affect further analysis, like the accuracy of classification.

Data Summary

A brief introductions to the 7 variables:
1. Age (BENE_AGE_CAT_CD), the beneficiary's age, factor data, reported in six categories: (1) under 65, (2) 65 - 69, (3) 70 - 74, (4) 75-79, (5) 80-84, (6) 85 and above.
2. Gender (BENE_SEX_IDENT_CD), factor data, two categories: (1) male or (2) female.
3. Base DRG (IP_CLM_BASE_DRG_CD) or diagnosis set: factor data containing 311 different codes derived from MS-DRG codes. It identifies a basic diagnosis or a set of diagnoses. A base DRG code might

be comprised of up to three MS-DRG codes. For simplicity, in this report I see each code as a different diagnosis set.
4. ICD-9 primary procedure code (IP_CLM_ICD9_PRCDR_CD) or procedure: factor data containing 85 different codes from 0 to 99, each code represent a primary procedure of a claim. This variable has missing value which means no procedure has been conducted. For the analysis simplicity, I took this as code 100. Thus this variable has 86 different codes.
5. Length (IP_CLM_DAYS_CD),factor data represents the length of stay reported in four categories: (1) 1 day, (2) 2 - 4 days, (3) 5 - 7 days, and (4) 8 or more days.
6. Payment level (IP_DRG_QUINT_PMT_CD) factor data linked with variable 3, Base DRG. It has 5 categories from 1 to 5, which means for each different diagnosis set, all payments of each claims have broken into 5 approximate quintiles. So we can see this as the payment level of each diagnosis set.
7. Average payment amount (IP_DRG_QUINT_PMT_AVG) numeric data. Though this data is numeric, it actually represents the average payment in every quintile (i.e payment level) of different diagnosis set. So in this dataset, if given Base DRG code and the payment level, we can get a fixed number of average payment. Which also means variable 6 and 7 have very similiar meaning.
For more detailed information related to this dataset, you can refer **Data Dictionary and Codebook for the 2008 BSA Inpatient Claims PUF**.
All factor data shown in this report is based on its code for simplicity. You can refer to the Data Dictionary above to find the actual meaning of certain factor.

Project aims and main method

Through this project I wish to illustrate my capability of applying efficient statistical tools and approaches to handle complex large dataset. My aim is conduct classification and data visualization on this dataset. These two approaches are highly representative as they are useful in almost every data analysis research related to biomiedical information. And I will also conduct them with similiar but more advanced methods during the summer program.

Data preparation

Before data visualization and classification. It is always necessary to clean the dataset. Firstly I've deleted the ID of claims and Variable 7. Because variable 7 and 6 have almost the same information and Variable 6 contains factor data, which is easier to deal with. Then I also modified the missing value in Variable, different procedures into code 100, to represent no procedure is conducted.
Then I set up a function to choose random samples to create the training and testing set for future classification. I will randomly extract 80% samples from a given matrix containing claim information to set up the training set and the rest will act as the testing set.

Main approaches illustration

My analysis focusing on the Variable 3: diagnosis set. As this variable affect the meaning of Variable 6. Payment level is meanful only when the diagnosis set is given as same level in different diagnosis set actually represent different average payment. So firstly I will set up a function to extract all information of a specified diagnosis set and conduct analysis on the particular subset.

Data Visualizaion

After extracting a certain diagnosis set, I wish to show relevant information of that diagnosis set. Like how many different procedures are conducted on this symptom. And which is the most common procedures conducted. We also wish to know some ratio related to this. And Data Visualization is the quickest way to get those information. In order to show ratios. I will plot pie charts. I've created a function based on ggplot2 to plot pie charts. This function will return a list containing 4 pie charts for each diagnosis set.
1. the ratio of sex of the specified diagnosis

2. the ratio of different aged groups of the specified diagnosis. This can be used to search for symptoms aiming a certain aged groups.
3. the ratio of the length of hospital stays. This in a sense can reflect to the severity of a certain symptom and the overall cost level of a symptom (As we can think that longer hospital stay will lead to a higher cost of hospitalization expenses and may also suggest more complex and costly procedure) 4. the ratio of different procedures conducted on the specified diagnosis set. As we know a certain symptom may have only several certain procedures to be dealt with. By this visualization method we can actually see this assumption is reasonable and we can easily find the main procedure for each symptom.

Classification

We may be curious about the question that if given information related to the sex, age, symptom, main procedure conducted,and hospital stays for a certain beneficiary. Can we predict how much do he/she spend on this claim? In this dataset, as all relevant information are factored. We can apply Machine learning algorithms on the classification procedure to predict which payment level does this beneficiary belong. I've applied decision tree and random forest as the main predictive model. And I've tested the accuracy of each model based on the training and test set we set up.

Decision Tree

For decision tree, We need to prune the tree on a certain value of the complexity parameter. I've referred to the error-complexity parameter plot given by cross-validation to choose a reasonable CP value for the trimming procedure. After the trimming, a plot will be given for the certain decision tree we use. This plot shows the exact rules of classification based on this certain decision tree.

Random forest

Though the principle is much complex than decision tree, the process of random forest is much easier in this report. As we don't need to do the trimming.

Example

I will show two examples here on two random diagnosis set I choose. To show the classification and data visualization result.

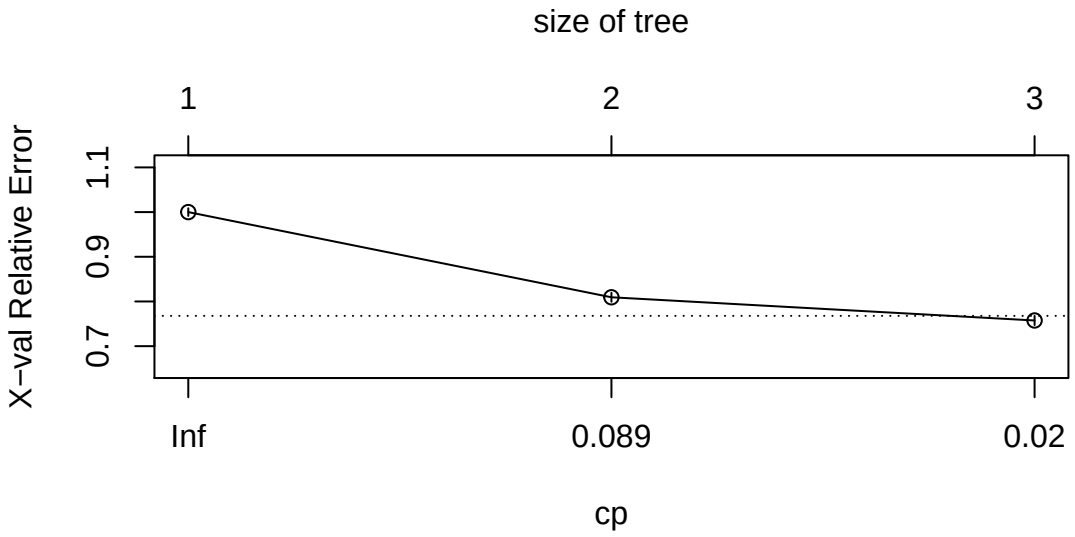
Example one

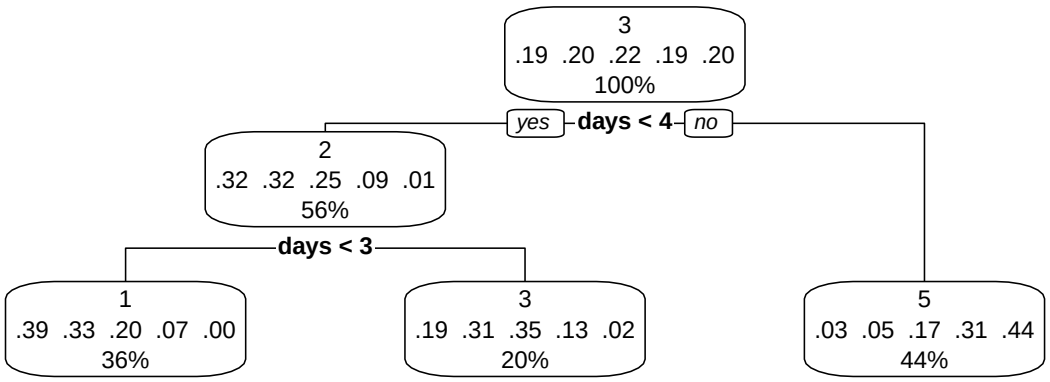
This example is based on the diagnosis 20 "Degenerative nervous system disorders"

Decision tree

Based on the plot, the value of complexity parameter we choose is 0.01

```
##           CP nsplit rel error    xerror        xstd
## 1 0.20346021      0 1.0000000 1.0000000 0.008627238
## 2 0.03910035      1 0.7965398 0.8093426 0.010106793
## 3 0.01000000      2 0.7574394 0.7574394 0.010308917
```






```
##       Predicted
## Actual   1   2   3   4   5
##      1 135   0  30   0  10
##      2 102   0  63   0  19
##      3  54   0  50   0  61
##      4  26   0  31   0 141
##      5   1   0   9   0 189
```

```
## [1] 0.4060803
```

```
## [1] 0.771987
```

By decision tree, the exact accuracy of the model is 0.4060803, the probability that we can control the error of predictive payment level and accurate payment level within 1 is 0.771987

Random forest

```
##
## Call:
```

```
##  randomForest(formula = as.factor(class) ~ ., data = train1, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 58.66%
## Confusion matrix:
##     1   2   3 4   5 class.error
## 1 407 171  83 1  47   0.4259520
## 2 293 242 129 4  75   0.6742934
## 3 172 184 161 8 267   0.7967172
## 4  71  59  65 7 496   0.9899713
## 5   4  11  14 6 705   0.0472973
```

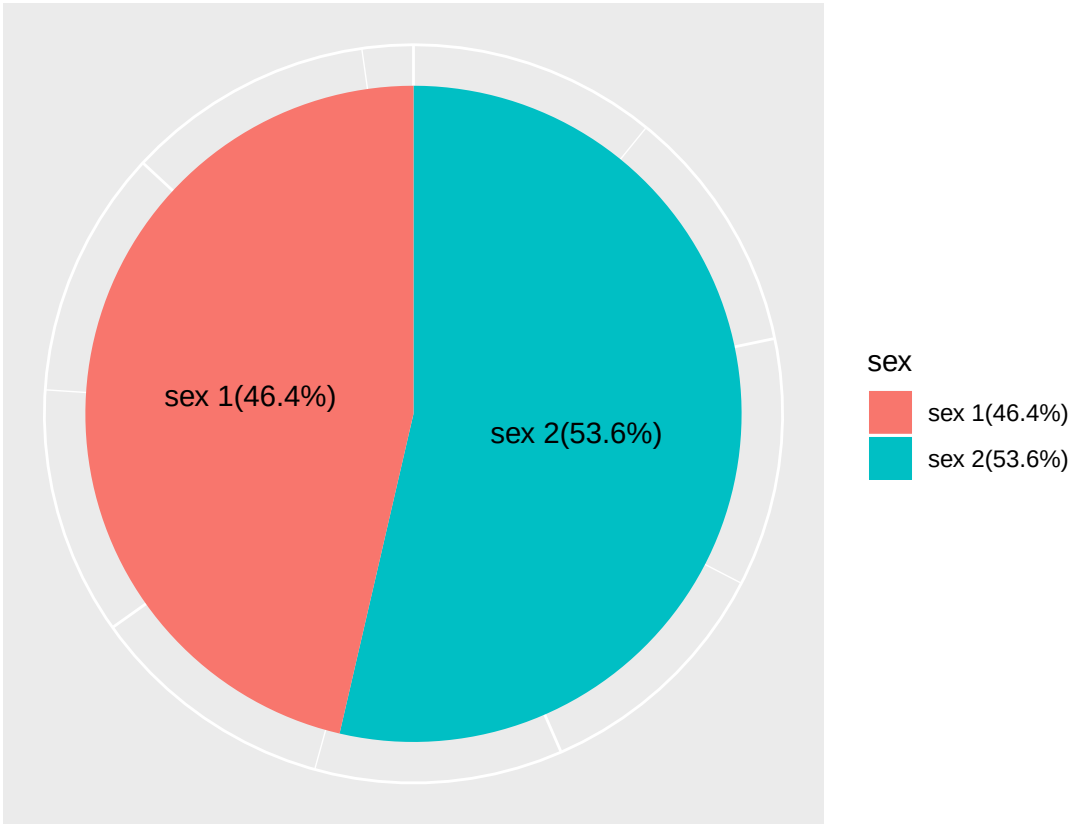
```
##       Predicted
## Actual   1   2   3   4   5
##      1 106  46  13   0  10
##      2  77  52  35   2  18
##      3  35  45  25   3  57
##      4  23  20  12   2 141
##      5   0   6   6   2 185
```

```
## [1] 0.4017372
```

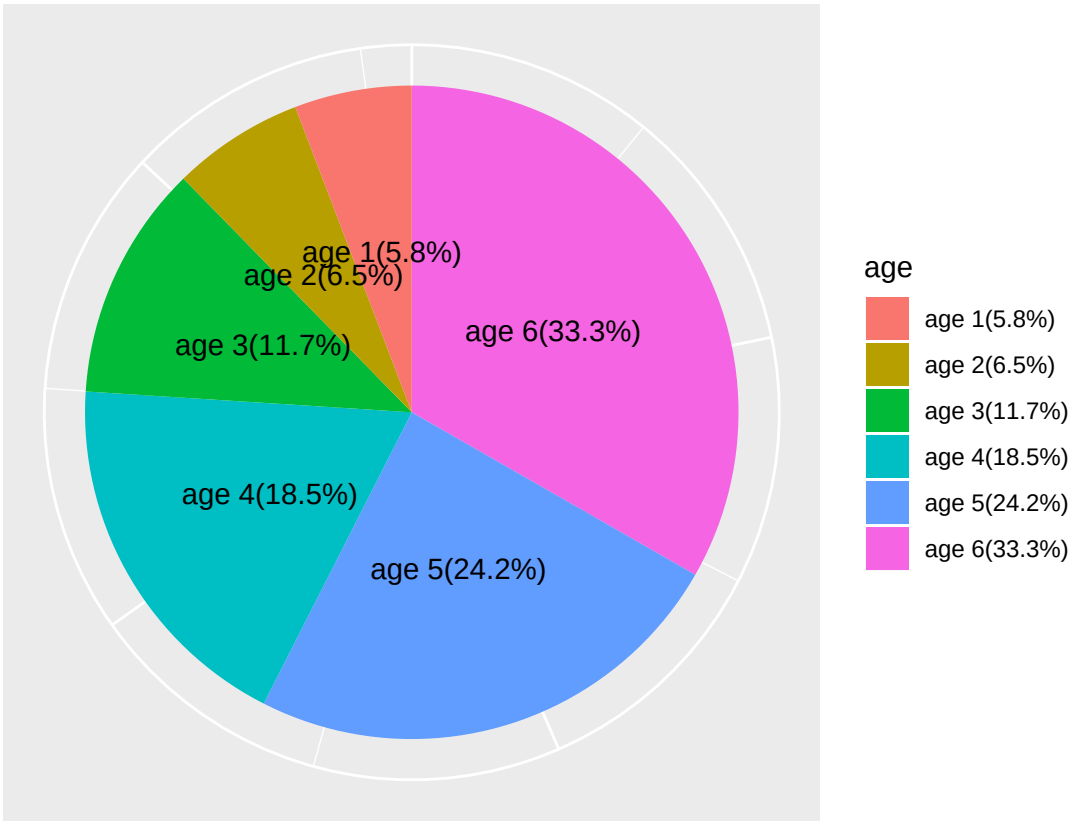
```
## [1] 0.7937025
```

by random forest, the exact accuracy of the model is 0.4017372, the probability that we can control the error of predictive payment level and accurate payment level within 1 is 0.7937025

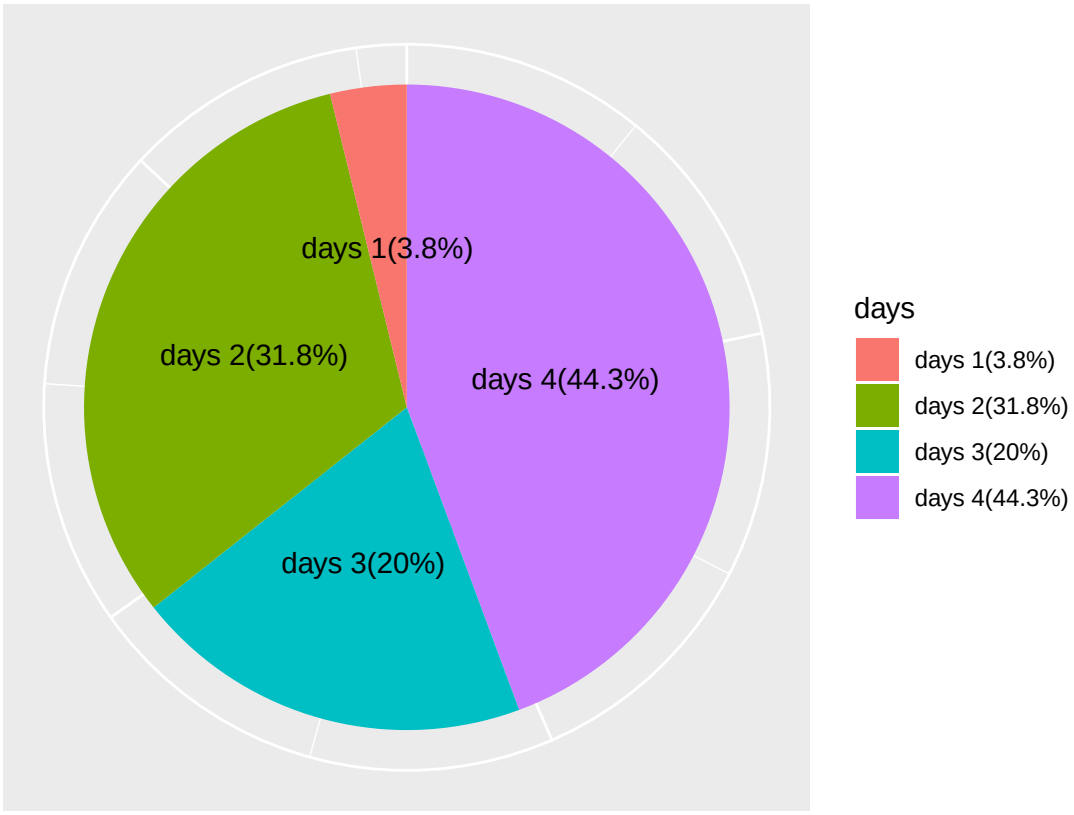
Pie charts



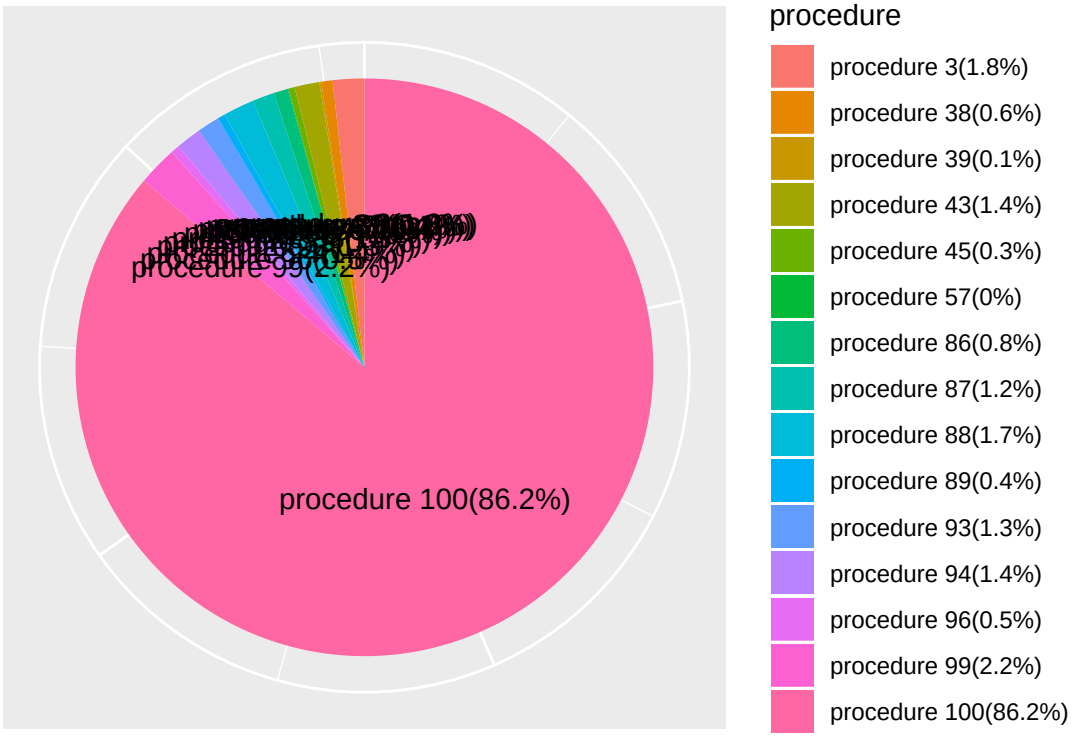
1. the ratio of patients sex in this particular diagnosis



2. the ratio of patients age groups in this diagnosis
   This disease/symptom occur easier on elderly groups.



3. the ratio of the length of hospital stays
   This disease/symptom may need more than one day to settle.



4. the ratio of possible procedures conducted on this diagnosis
   There are many procedures for dealing with this symptom for the reason that this symptom has a huge sample size. However, in most cases we take no procedure on this symptom.

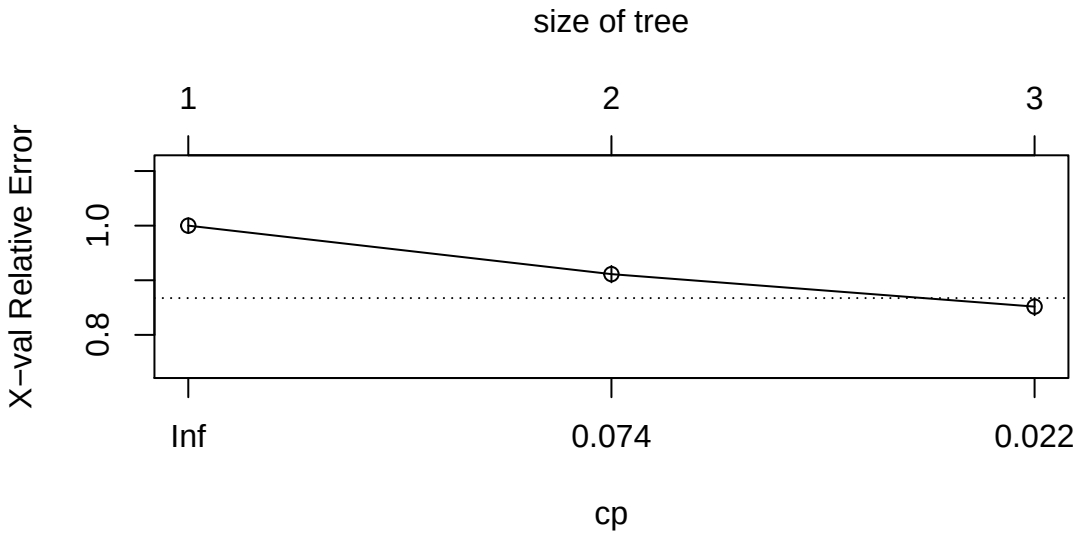
Example two

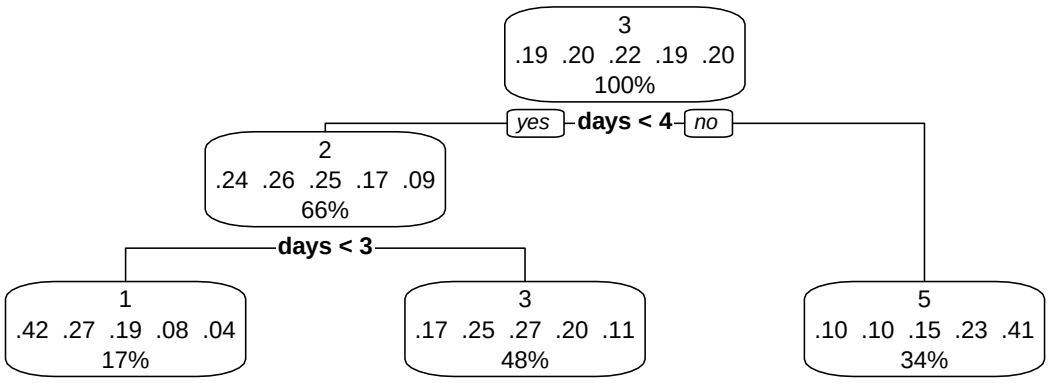
this example is based on the diagnosis 80 "Coronary bypass w/o cardiac cath"

Decision tree

Based on the plot, the value of complexity parameter we choose is 0.01

```
##           CP nsplit rel error    xerror       xstd
## 1 0.11465517      0 1.0000000 1.0000000 0.01368595
## 2 0.04827586      1 0.8853448 0.9112069 0.01500893
## 3 0.01000000      2 0.8370690 0.8517241 0.01564443
```






```
##       Predicted
## Actual  1  2  3  4  5
##      1 35  0 37  0 13
##      2 18  0 39  0 12
##      3 10  0 46  0 13
##      4  4  0 37  0 29
##      5  3  0 16  0 59
```

```
## [1] 0.3773585
```

```
## [1] 0.7088949
```

By decision tree, the exact accuracy of the model is 0.3773585, the probability that we can control the error of predictive payment level and accurate payment level within 1 is 0.7088949

Random forest

```
##
## Call:
```

```
##  randomForest(formula = as.factor(class) ~ ., data = train1, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 66.26%
## Confusion matrix:
##     1  2   3 4   5 class.error
## 1 100 37  92 2  53   0.6478873
## 2  68 40 138 2  53   0.8671096
## 3  47 43 153 0  79   0.5248447
## 4  24 19 120 0 118   1.0000000
## 5   8  9  68 2 207   0.2959184
```

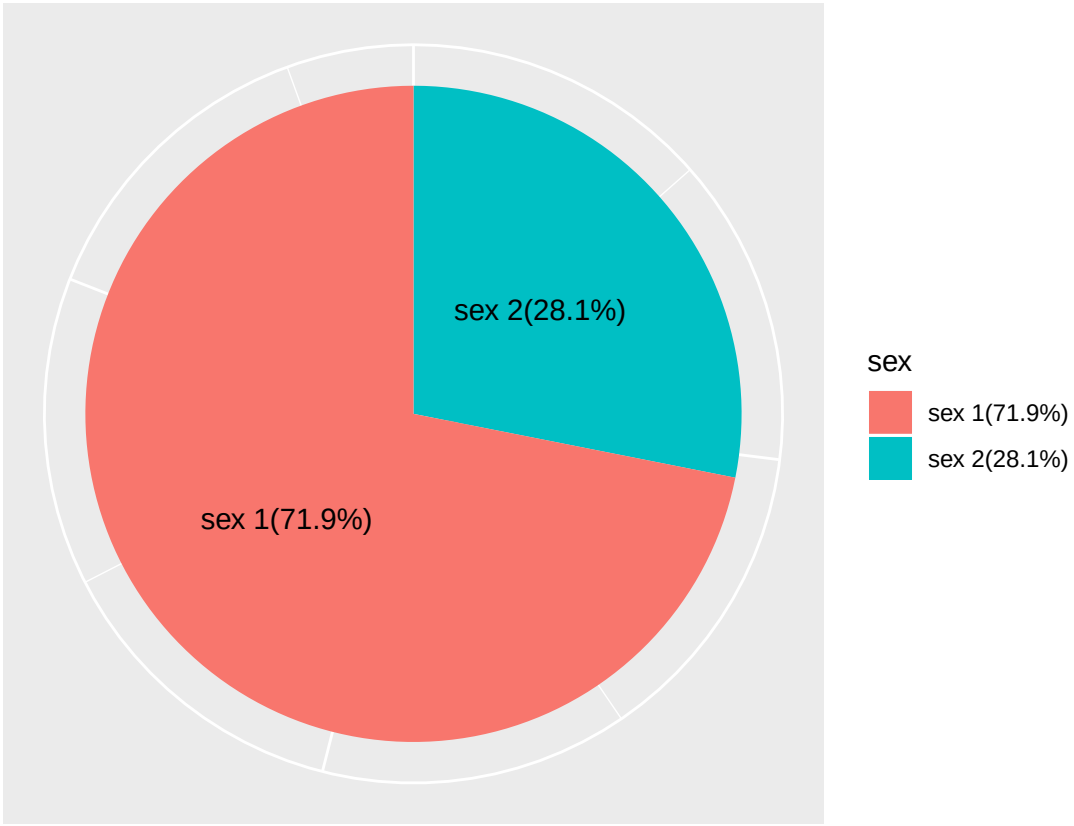
```
##       Predicted
## Actual  1  2  3  4  5
##      1 32  8 32  0 13
##      2 18  5 34  0 12
##      3 10  9 37  0 13
##      4  4  7 30  0 29
##      5  3  4 12  0 59
```

```
## [1] 0.3584906
```

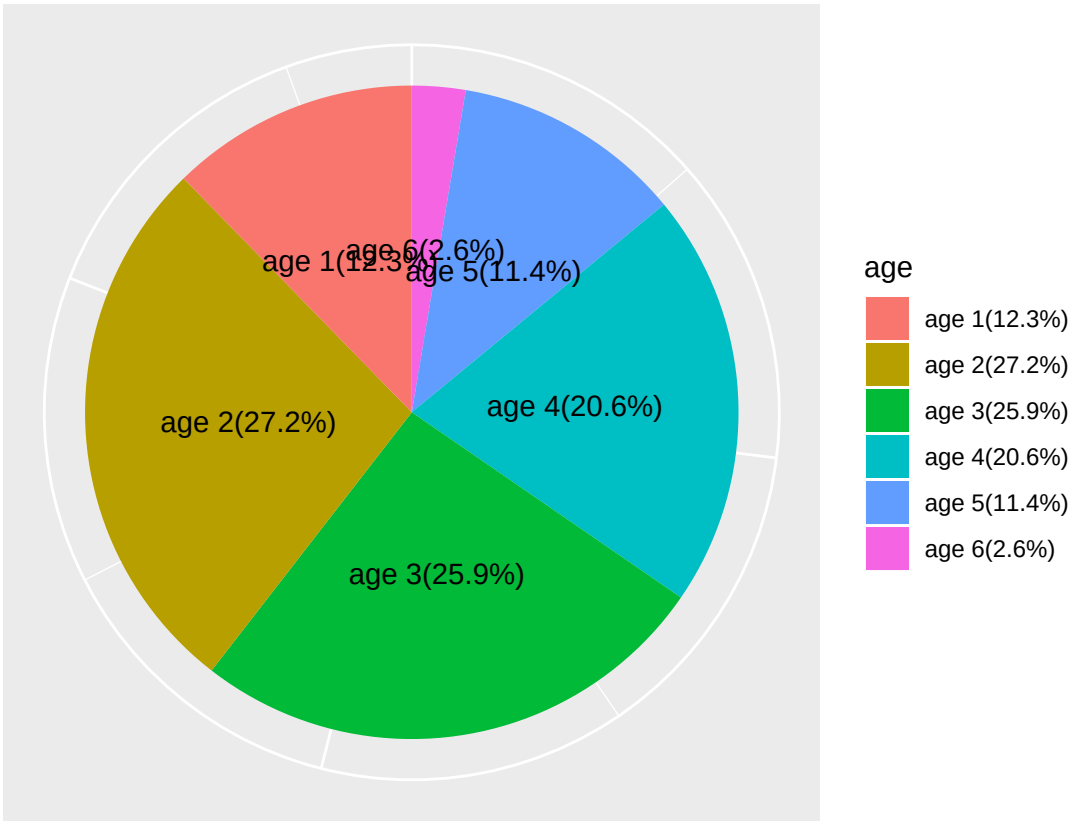
```
## [1] 0.703504
```

by random forest, the exact accuracy of the model is 0.4017372, the probability that we can control the error of predictive payment level and accurate payment level within 1 is 0.7937025

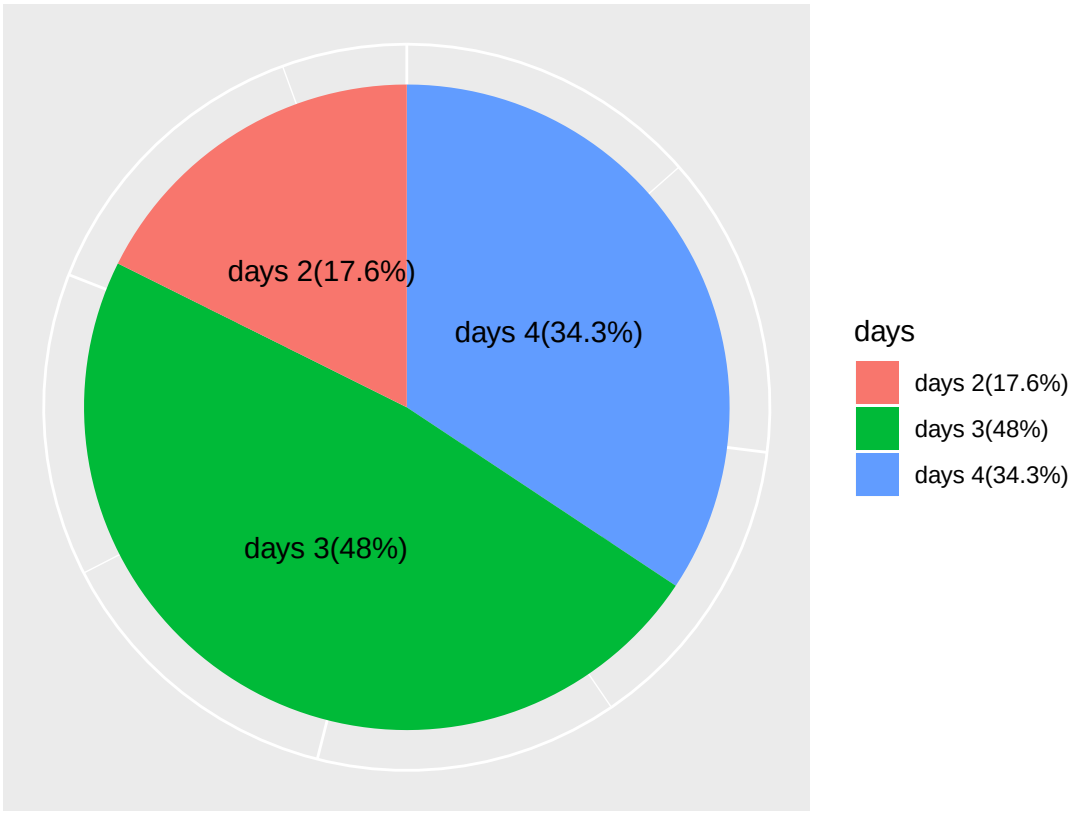
Pie charts



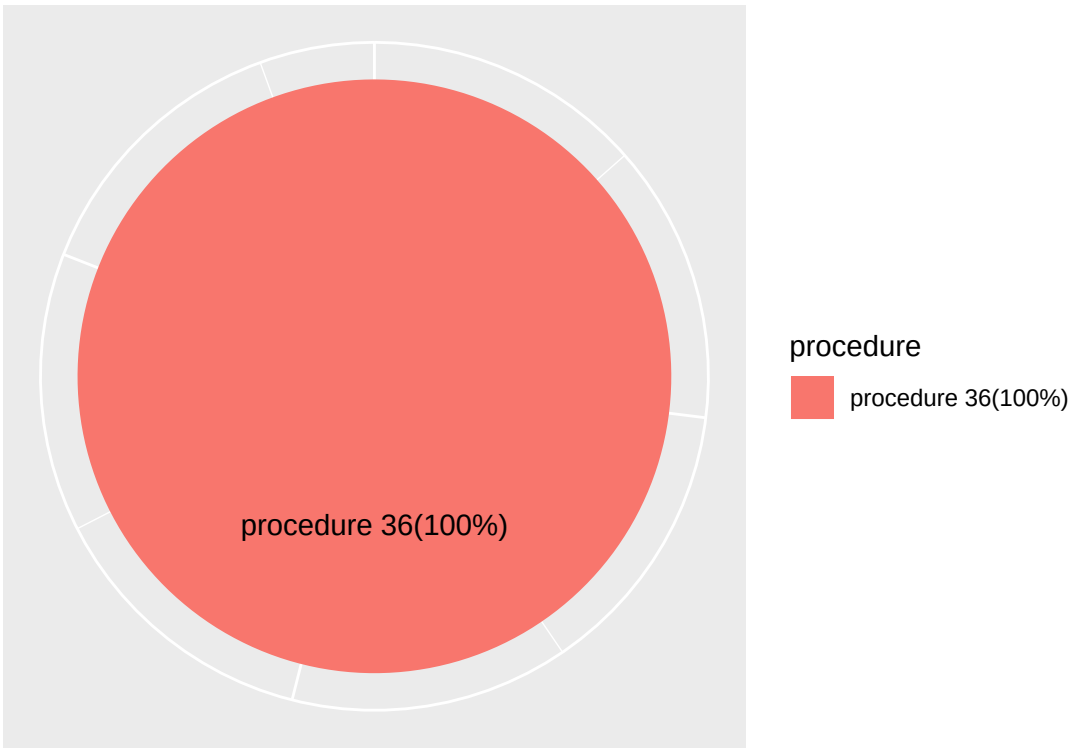
1. the ratio of patients sex in this particular diagnosis This symptom/disease highly concentrate on male.



2. the ratio of patients age groups in this diagnosis This symptom/disease concentrate on middle-aged group.



3. the ratio of the length of hospital stays It always takes more than one day to deal with this symptom and in most cases it needs 5-7 days



4. the ratio of possible procedures conducted on this diagnosis There is only procedure for this disease, procedure 36, *Ops on heart vessels*.

Summary

By the two examples above, we can see that based on this particular dataset. It is hard to find a good classification model based on Decision Tree and Random Forest. The exact accuracy of each model is around 0.4 and if we can tolerate some error so that we control the difference between predictive level and the accurate level within one, those two method can reach an accuracy level around 0.8. Still, they are not pretty good classification model. But this may due to the dataset itself. As the dataset is a subset of the database, when doing the classification, we are only using 4 variables. The predictive variables are so less and not all variables are so influential on the classification process. For example, the length of stays can only in a sense reflect to the level of cost. The real life situation can be much complex, like different hospitals may have different charging rates. Missing those information may highly affect the accuracy of the classification. In order to get a better classification model, we may need to refer to more datasets or even the whole database. To import more information related to the payment level.

More things that can be done

There are many more things we can do on this dataset even it is only a subset of a database. For examples, by the data visualization we can show that a certain diagnosis set may have a certain patients group (e.x. symptoms aiming erlderly people) and may reflect to a certain hospital stays(e.x. some symptoms may be worse and may require more time to cure). Based on this information, we may be able to rank the severity of different symptoms or even rank the predictive cost level of different symptoms when taking Variable 7, average payment into consideration. Similarly, we may also rank the cost of different procedures. But just like classification, we may not get a pretty accurate result merely based on such information. But still this may act as a feasible direction to dig in.

Coding and Rmd files

To save the page, I've uploaded the R code and Rmd file onto Github. You can **view them here** if needed.