Fall 2021 - Math 421 - Midterm

Instruction

The midterm has two components: the Rmarkdown notebook (html) and the presentation. We will do the presentation in class. Post both the notebook and the presentation on your Github page.

1. The notebook

The notebook should be created using `rmarkdown` (like other assignments). The notebook should have a title. It should

The Presentation: Present your results in 5-10 minutes. To make the presentation using Rmarkdown, do the follows:

```
- In Rstudio -> File -> New File -> R markdown

- In the left panel, click to Presentation -> Click OK

- Now you have an Rmarkdown that can be knitted to be a html presentation
```

- You do not need to rerun all the codes for the presentation. For example, to show the model comparison, you just need to show the image of the model comparison instead of running all the models again.
- To inset an image in a slide, use `![](image.png)`
- To turn off message and warning of a code cell, use: `{r, message=FALSE, warning=FALSE}` for the cell.

What to present:

- Present Part 2 - Visualization
- Present Question Question 4, 5 and 6 in Part 3.
- Present any errors/challenges you run into and how you fix/overcome them.

Data:

The data for the mid-term project is the Rhode Island Department of Health Hospital Discharge Data. Each row of the data presents a patient.

Link: https://drive.google.com/open?id=15QNBf6YYKocK2nNIfpKDer58kQnCPNZJ

I. Data Wranggling

1. Download the data file `hdd0318cy.sas7bdat`.
2. Use `read_sas` in library `haven` to read the data.
3. Filter the data to have only patients of the year 2018 (`yod==2018`)
4. Select to work with only following variables:

```
                      "yod", "payfix","pay_ub92","age",
                      "sex","raceethn","provider","moa",
                      "yoa","mod","admtype", "asource" ,
                      "preopday" ,"los", "service" , "icu","ccu",
                      "dispub92", "payer"  ,"drg","trandb",
                      "randbg","randbs","orr", "anes","seq",
                      "lab","dtest", "ther","blood","phar",
                      "other","patcon","bwght","total","tot" ,
                      "ecodub92","b_wt","pt_state","diag_adm","ancilar" ,
                      "campus","er_fee","er_chrg","er_mode","obs_chrg",
                      "obs_hour","psycchrg","nicu_day"
```

Notice: You may want to save the current data to your computer for easy access later. To save the data file use `write_csv(df, 'midterm.csv')`, for example.

5. What are variables that have missing values?
6. Remove all variables with missing values
7. Refer to the data description in the file `HDD2015-18cy6-20-19.docx`, which variable recording the month of admission?, which variable recording the month of discharge?
8. Which month admitted the most number of patients? Which month admitted the most number of male patients?
9. Which month has the most number of teenage female patients?
10. Which provider has the most number of female patients in October?
11. Is female patients older than male patients, on average?
12. Calculate the average age of patients by months. Which month has the oldest patients on average age?
13. What is the name of the provider that has the highest total charge?
14. What is the name of the provider that has the least total charge for teenage male on average?
15. Calculate the length of stays by races. Which race has the longest length of stays on average?
16. On average, how much a 20 year-old male white get charged for staying 1 day?
17. Write a paragraph to summarize the section and give your comments on the results.

II. Data Visualization

Continue with the data from part I.

1. Provides at least 10 meaningful plots. Comments on the plots. All plots should have title, caption, appropriate labels on x and y-axis
2. Make an animation plot.
3. Write a paragraph to summarize the section and give your comments on the results.

III. Predictive Models

Continue with the data from part I. Use the follows as the target and input variables:

Target Variable: Create the target variable taking value of

- `low` if the total charge of a patient (`tot`) is smaller than the median of the total charge, and
- `high` otherwise.

Input Variables:

- "age","sex","raceethn","provider","moa","mod","admtype","campus", 'los'

---

1. Use `filter` function to filter out rows where `raceethn==''` or `admtype==''`. Make sure all the categorical variables are factor, numeric variables are numeric. Set Training : Testing Split = 10 : 90
2. Train a decision tree using `rpart`. Plot the decision tree. Plot the variable importance ranked by the tree.
3. Using caret for this question. Set `Training Control` to be: Use Cross-Validation of 5 folds across all models. Train & tune at least 3 different models (i.e. three different values for `method=` in the train function of caret). Plot the hyper-parameter tuning plots for each model.
4. Plot the comparison of the models in 3.
5. What is your final selection for the model? Test the accuracy of your final model on the test data.
6. Create another `target` variable (binary), decide the input variables and redo 1 to 5.
7. Write a paragraph to summarize the section and give your comments on the results.
