Stat 477 Fall Q1 2020
Final Project

Introduction

The goal of this final project is to give you a chance to use the various data skills acquired during the course, to analyze a dataset and present a report.

The dataset (”project_477.csv”) has two possible y-variables and you can choose either of them to analyze.

They are:

1. Health index. To be treated as a continuous variable.
2. Majority of Democrat or Republican votes in the 2016 presidential election. Treated as a two-level categorical variable.

Background on the data

The data concerns health, economic and demographic characteristics of counties in the US. The unit of analysis is the ***county***. Though it is not necessary, if you are interested in the data source it is https://www.countyhealthrankings.org/

Many of the variables are reported on a percentage basis and measure the extent to which a variable is present in a county. For example, the column called Percent.Female records the percentage of females in each county. Most of the variables are self-explanatory, like “Life expectancy”, but a data dictionary is provided for your reference.

Deliverables

You will create a Jupyter notebook that contains all your work and you will save it as an HTML presentation using Reveal.js. The two deliverables are the Jupyter notebook and the HTML slide presentation.

This project is meant as a creative and open exercise. The goal is to tell an interesting story based on the data and accompany it with clear and relevant graphical output. One approach to this task is to imagine that you have a 15-minute presentation to put together that comprises about 20 slides. So you would want to present interesting findings that keep your audience engaged.

With this much data and this many variables, it is very possible to produce a large amount of output, but you should not overdo it! Do not create more than 30 slides.

Data analysis process

Process-wise the following steps may serve as a useful guide to your activities:

1. Select the y-variable.
2. Review the data dictionary.
3. Articulate a few key questions of interest regarding the y-variable, in terms of the available predictors.
4. Download, import and review the raw data.
5. Create some summary statistics by State to gain insight into differences between them.
6. Review your y-variable's association with some of the potential predictor variables (you don't have to look at every column, and it is fine to scope your questions so that they are refer to a subset of the columns.)
7. Create graphics that illustrate the associations.
8. Look at the relationship between some of the predictor variables. For example, the relationship between drinking and smoking, by State, is quite interesting. Using the "hue" argument in seaborn's graphics can be very helpful here.
9. Create a predictive model for the y-variable. Use a simple tree for this part. You do not have to do the cost complexity pruning to select the tree (however, you can if you want to). It is fine to create a tree simply by specifying the maximum number of leaf nodes and this may simplify the interpretation.
10. Use the variable importance metric to rank the variables' importance in the tree.
11. Create a graphical summary of the tree model.
12. Create narrative elements in the presentation that interpret and discuss the findings as you proceed. Never forget that the associations you find are not necessarily causal, so don't overstate your findings.
13. If you want to answer your questions (3) using a regression model ***as well*** as the tree model, that is acceptable too.
14. Make sure that there is both a "goals" slide that articulates your questions and a conclusion slide that summarizes your findings.
15. Make sure that the notebook runs without producing any errors.

Notebook creation

I would suggest creating the notebook in three steps:

1. Make the note book with just the Python code.
2. Once all the code is running, add the narrative elements.
3. Use the slide menu to create and then save the actual slide presentation. The slide menu allows you to hide cells that you don't want to show.

There is a "FP_starter.ipynb" notebook in the homework folder that you can use as a starting point if it is helpful.

Grading

When the project is graded, it will be assessed on the following components. This means that if you can check off the following items, you have done a reasonable job.

ELEMENT		Present/absent?
Were the guidelines followed?		
	20-30 slides.	
	Introduction slide(s).	
	Conclusion slide(s).	
	Coherent narrative linking the steps of the presentation.	
	Does the notebook run without any errors?	
Y-variable selected, described.		
Questions of interest articulated.		
Graphical and numerical summaries of the y-variable presented.		
Review of selected x-variables (numerical and graphical). You can scope your questions to involve only a subset of the x's.		
Exploration of the relationship between y and selected x-variables.		
Review of relationships between selected x-variables.		
Construction of appropriate regression or classification tree.		
Graphical summary of tree.		
Interpretation of tree, regarding the questions of initial interest.		
Identification of key terms in the tree (variable importance)		