DACSS 601: Data Science Fundamentals

Final Paper

Due Date: 11:59pm EST, Friday, August 13th, 2021

Assignment

For your final paper, you will present an original basic analysis of a research question. The goal of the project, though, is *not* the specific research finding, but rather the data pipeline that generates, ultimately, a visualization that contributes to our understanding of the underlying research question *and* an understanding of your decision-making throughout that pipeline. The focus, that is, should be on your process as a computational social scientist, rather than on the specific product produced. The paper should be 8-12 pages in length, written in R Markdown.

Pro tip: Each of the homework that you will do in this class is one part of the final paper. If you work on a single dataset in which you are interested throughout the course, your final paper will practically write itself (except for the writing, of course). I would recommend that is what you do.

The paper should include:

* *Introduction (1-2 pages)*: Define the research question or problem you seek to address. What animated your concern with this research question? What are you arguing?

Introduction

In this research project I want to understand what attributes of both the human body and attributes of the heart relate to heart diseases. My interest derives from the fact that I am currently under diagnosis for my own heart related disease and would like to become more versed in understanding what attributes that I myself should look into and moreover gain a better understanding of the terminology in the cardiology field. This will hopefully empower me to be more able to understand the terminology when discussing treatment options with my doctor as my diagnosis continues. I feel this understanding will help put myself more at ease as I go through this tough time and help me not feel overwhelmed by the overall jargon the doctors will be using when discussing any problems I may or may not have.

I am particularly going to look at the UCI Heart disease dataset because it correlates directly to real life patient data and is well structured data. In an attempt to see what attributes I should look at when trying to predict heart disease. I want to find out if heart disease is more related to attributes of the body such as age and gender or more so to attributes of the organ itself. As well as understand which features are more related to each other under the guise of heart disease. I am interested in mainly seeing how age, gender, blood pressure, resting electrocardiographic measurement seem to have a impact the overall chances of having a heart disease.

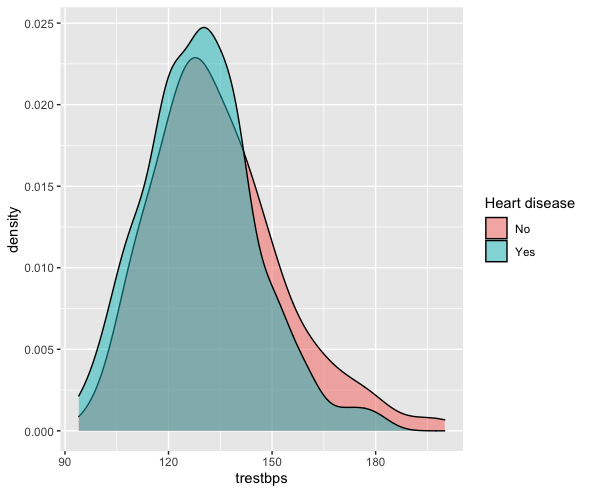
• *Data (2-3 pages)*: Describe your data source. Why is this an appropriate choice? What are the variables that you are specifically interested in? Describe those variables, both numerically and visually.

UCI Heart Dataset

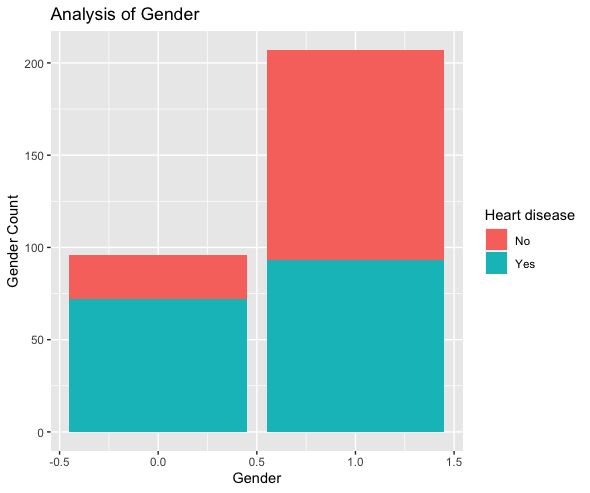
The original dataset is from the UCI which is a repository for datasets for developing and working with machine learning algorithms. However, the original distributors of the datasets were University Hospital, Zurich, Switzerland: William Steinbrunn, M.D, University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D, V.A.Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano,M.D., Ph.D. These are the following features of my dataset:

* age: The person’s age in years
* sex: The person’s sex (1 = male, 0 = female)
* cp: The chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
* trestbps: The person’s resting blood pressure (mm Hg on admission to the hospital)
* chol: The person’s cholesterol measurement in mg/dl
* fbs: The person’s fasting blood sugar (if > 120 mg/dl, 1 = true; 0 = false)
* restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes’ criteria)
* thalach: The person’s maximum heart rate achieved
* exang: Exercise induced angina (1 = yes; 0 = no)
* oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot)
* slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
* ca: The number of major vessels (0-3)
* thal: A blood disorder called thalassemia (1 = normal; 2 = fixed defect; 3 = reversible defect)
* target: Heart disease (0 = no, 1 = yes)

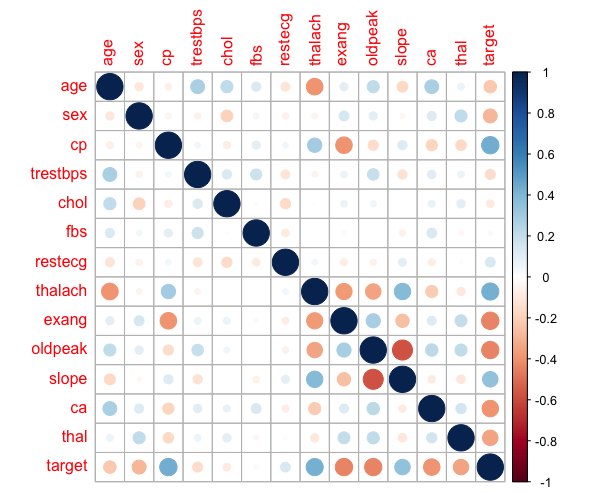
I want to see how people’s resting blood pressure and gender relate to heart disease. I would expect to see high variation in people who have an underlying heart condition highly related to their blood pressure. However, looking at the graph below we can see very little change between people with heart disease and those without in terms of high and low resting blood pressure. Thus the data shows that looking at a patient’s individual resting blood pressure is not an effective way of determining heart disease.



When thinking about gender I would expect more males to have greater heart disease due to the fact they are generally physically bigger individuals and therefore their heart strain must be greater and thus increasing the chances of disease. But upon looking at the data for gender when compared against heart disease, we see that our dataset has more females then males, nearly a 1 to 2 ratio, making our data imbalanced for this feature and as a result we can not use this is untrustworthy for decision making when related to heart disease.



Finally it is a good rule of thumb to always look at the correlation of each of your features with one another. To do this I simply generate a correlation plot map. We can see from the given correlation plot the our dependent variable is least related with fbs, chol, trestbps, restecg because they are between -0.2, and 0.2.



Before doing any statistical analysis, modeling, and more visualisations I will remove the following features due to high correlation or usefulness when determining heart disease: fbs, chol, trestbps, restecg, and sex.

• *Reflection:* Describe your experience with the process. What decisions in your pipeline are you most concerned about potentially influencing your findings? What were the most challenging and time-consuming aspects of the project? What do you wish you had been able to do? If you were to continue the project, what would your next steps be?

Luckily my data was very clean and I was not missing anything in any of my features, thus I had to do no data cleaning for my dataset. When looking at my data flow pipeline I was constantly thinking about what features may or may not make the most sense and impact in determining heart disease. Originally I thought gender and blood pressure would be impactful, but my dataset shows that these features do not make as much of an impact as I once thought. Looking at my decision tree tells us that the features of chest pain, number of major vessels, and the maximum heart rate achieved impacts the decision the most when making a determination on whether or not the patient has heart disease. But sadly we only got an 80% on this model for our testing data. This in itself is not terrible but it does not truly capture the full understanding that we would like to see to make a more significant claim. However, this does personally tell me that my own individual chest pain is nothing to scoff at as it might truly relate to a more underlying issue. If I were to start this project over again I would like to find a bigger dataset than only 300 patients.

The most time consuming part of this project was not necessarily the data itself but mainly learning how to program in R and likewise how to work with R studio. I was very confused about how to manage R sessions and thought there was an environment I had to point to for individual projects, but as it turns out each R session has zero packages and you install them into your own session from the code. My own concept of Python environments posed some problems and made me spend more time than needed. I also spent a few hours trying to get R to work in Jupyter Notebooks because of my familiarity with notebooks, but I kept getting strange errors that were crashing any R notebook I would try to spin up. To continue with, R markdown for notebooks took a few tries to properly run due to Knit complaining about cran repositories and lines of code needing to be encapsulated in uniquely defined markdown titles. However, the overall hardest part was simply learning and getting used to the general feel of the syntax. I had a difficult time getting used to the concept of piping and plotting differently as well as learning how to properly use the R markdown of Knit. I felt slowed down and hindered since I knew the same code in Python and constantly found myself typing code for Python by accident and wanting to call seaborn graphs. I hope to have more projects in the future in R to get more well versed in the syntax so it does not feel so foriegn to me and in turn I will add more to my own data scientist arsenal.

• *Conclusion (1-2 pages)*: Discuss your research question again briefly and state succinctly the preliminary findings. Also discuss the potential ramifications of action – or inaction – on this policy.

The purpose of this project was to find what features and likewise the values of those features seem to impact the determination of classifying heart disease the most. Looking at my statistical testing we do find that cholesterol actually is significant enough to consider it for further analysis and modeling which lines up with my general understanding of heart disease/problems. However, I was conflicted in removing the feature of resting blood pressure due to the fact my statistical test made it seem impactful. But when looking at the visualizations it did not appear that it would hold a significant impact. If given more time I would love to go back and experiment more with this feature. The model is telling us that the features of chest pain, the number of major vessels, and the maximum heart rate achieved impacts the decision the most when determining whether or not the patient has a heart disease. This makes sense as the number of major vessels changes the strain on the heart that could result in a disease as well as the maximum heart rate of the patient. With an 80% accuracy, there is still rom to grow. Moving forward, I feel as though more patient data would increase the complexity curve of the model making the given accuracy of 80% more impressive.

The findings from this study are crucial for the medical experts who work with cardiac patients. The information allows doctors to use certain factors to determine a patient’s future with heart disease. The more clear understanding we have of warning signs, the likelihood of survival increase. Preventative measures can take place before the patient is in crisis. This study is worth exploring further to allow cardiac patients to have a better quality of life.

*NOTE: If you are having trouble, let me know!*

1