Final Project - step 2

Mithil Patel

2022-05-21

How to import and clean my data

There cardio dataset contains 4238 rows while heart dataset has 1025 rows. We shall shrink the cardio dataframe size to match the heart dataframe size.

male	age	education	currentSmoker	cigsPerDay
1025	0	1050	1025	1027
BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol
1037	1025	1025	1025	1037
sysBP	diaBP	BMI	heartRate	glucose
1025	1025	1028	1026	1115
target	sex	cp	trestbps	chol
0	1025	1025	1025	1025
fbs	restecg	thalach	exang	oldpeak
1025	1025	1025	1025	1025
slope	ca	thal		
1025	1025	1025		

The chart above shows the number of NA values in each column. Since both data frames do not have a specific column (i.e. 'id') in common, we shall utilize an in-built bind_rows function in r to combine data frame by the 'target' column, which determines whether or not an individual has a heart condition. However, there is a con when using bind_row to combine two different data frames half of your combined data frame contains na values. To handle na values present in our data frame, we will replace na values with column mean.

We have created a 28x28 correlation matrix to get a rough estimate as to which variables appear to be correlated with our 'target' variable. Any correlation value above 0.20 or below -.20 will be considered in our final data frame.

Note: The correlation matrix is not shown due to inconsistent matrix format

What does the final data set look like?

	sex	cp	thalach	oldpeak	slope	ca	target
2016	1	1	178	0.8	2	0	1
2017	1	0	160	1.4	2	2	0
2018	0	0	159	0.0	2	0	1
2019	1	0	143	0.1	1	4	0
2020	1	0	142	1.2	1	1	0
2021	1	1	173	0.0	2	0	1
2022	0	0	150	1.9	1	2	0

2023	1	0	113	1.4	1	1	0
2024	1	0	125	1.8	1	0	0
2025	1	0	163	0.2	1	2	0
2026	1	0	132	2.0	1	2	0
2027	1	0	178	0.0	2	0	1
2028	1	0	132	0.1	2	1	0
2029	1	0	147	0.1	2	3	1
2030	0	2	142	1.5	2	1	1
2031	0	0	130	2.0	1	1	0
2032	1	2	165	0.0	2	0	1
2033	1	3	162	1.9	1	0	1
2034	1	1	162	0.0	2	0	1
2035	1	0	181	0.0	2	0	0
2036	1	0	173	1.6	2	0	0
2037	1	1	170	0.0	2	0	1
2038	1	1	168	1.0	0	0	0
2039	1	0	140	4.4	0	3	0
2040	0	2	175	0.6	1	0	1
2041	1	0	131	2.2	1	3	0
2042	1	3	174	1.4	1	1	0
2043	1	0	95	2.0	1	2	0
2044	1	0	158	0.0	2	0	0
2045	1	0	143	0.1	2	0	1
2046	1	1	164	0.0	2	0	1
2047	1	0	141	2.8	1	1	0
2048	1	0	118	1.0	1	1	0
2049	0	0	159	0.0	2	0	1
2050	1	0	113	1.4	1	1	0

Questions for future steps

Initially going through to import and clean the dataset, there were some struggles such as combining the two datasets with similar columns (e.g. 'target' column) would end up multiplying the rows and columns together instead of merging them by columns. For instance, two datasets, one with 1000 rows and the other with 4000 rows, would produce a total of 4,000,000 rows. to tackle the issue, I used one of the R programming language functions called bind_rows which stacks the rows from the two datasets appropriately. If columns and rows didn't match the two datasets, then those entries were filled with 'N/A'. Considering that a large portion of the combined dataset was filled with NA values, I had to take the mean of the columns and replace the entries with the mean values.

What information is not self-evident?

One way to identify outliers in the data is by using a boxplot which helps to show the distribution of the numerical data in our combined set. A boxplot will be beneficial in displaying the skewness in our distribution as well as any outliers present in our data. We can identify outliers using IQR (interquartile range) to calculate an outlier boundary. Any data point that falls outside of the upper and lower boundary will be considered an outlier and, as a result, will be removed/filtered out to ensure our results is statistically significant. Similar calculation will be performed on exercise dataset to remove outliers.

What are different ways you could look at this data?

One very helpful tool that can further answer our questions is by creating a logistic model to determine which variables from our dataset are more significant than others. By analyzing the summary of our logistic model, we can determine the effect each variable will have on the dependent variable (heart disease) based on the p-values. Thus, we'll be able to identify which variables are strong risk factors and the leading causes of death. Before creating a logistic model, we can also perform a normal regression to check for multicollinearity and high VIF value to determine whether or not the distribution is normal.

How do you plan to slice and dice the data?

We can split the datasets in 80/20 ratio, that is, 80% of the dataset into the training set and 20% into the testing set. We can split the data to determine the accuracy of our model. Another idea I had to go about slicing and dicing the data would be instead of merging the data, I can individually evaluate the datasets in order to reduce the disregarded values of 'N/A' that I had to replace with the mean of the column, which has significantly impacted the distribution of our model. I plan on using the Logistic regression model on each dataset (total of three datasets) as mentioned earlier to find the predictors tied to increasing the risk of heart disease. This will help us to further answer the question we formulated for our project in the beginning and have more precision on the data that we've gathered.

How could you summarize your data to answer key questions?

Once we have gathered and transformed our data, decided which predictors are correlated with heart disease, split our dataset into training and testing sets, and perform logistic regression on our dataset, we can calculate a confusion matrix to determine the accuracy of our model. If the accuracy percentage is above 70%, we are confident that we have created a good model. If so, the predictors used in our model are likely to increase the risk of heart disease, and predictor with the lowest p-value (obtain from the summary) is the notable contributor to cardiovascular disease. We can conduct a similar analysis on the exercise dataset to evaluate the performance of how exercising can help reduce the related risk factors and improve a person's health and fitness. Gathering enough data will assist us to analyze traits/behaviors an individual must exhibit to lower the risk of getting heart disease. Additionally, it can be essential in influencing the person's life as it will promote an overall healthier lifestyle.

What types of plots and tables will help you to illustrate the findings to your questions?

1. Boxplot
2. Correlation matrix
3. Histogram
4. Scatterplot

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

As of today, I am not sure which machine learning techniques I can employ, but I can potentially research from Week 11 reading material to decide upon the algorithm I can utilize to create a heart disease prediction model. Perhaps I could look into using a decision tree or neural network to improve model accuracy using machine learning algorithms/techniques.

Questions for future steps.

Originally working on this project, there were many techniques I was confused about and did not know the best steps to take to improve the model I plan to create. Originally, I decided that a multiple linear regression model would be a great approach to the matter at hand, but I decided to take advantage of the logistic model considering that our output/dependent variable is binary (either an individual has heart disease or not). With the logistic model, since it takes the probability of an event occurring based on independent variables such as predictors, we can make use of them to show what risk factors are more likely to increase heart disease. I need to learn how the logistic regression model is created in R and how to interpret the results. Additionally, I need to learn if logistic regression makes any assumptions and check if I need to check for normal conditions.

I also need to learn how to develop a machine learning model and the necessary libraries. Above all, I need to determine which algorithm I can use for a binary outcome.