Data Exploration and Preparation

So now let's see how we can explore and prepare our data, basically the heart disease dataset.

So some steps that need to be followed when you go through the process of data exploration and preparation is basically

First clean your data. That's the most important step.

- Make sure that your data does not have any wrong values or anomalies.
- Then you treat the missing values i.e. if there are missing values in the data, then you cannot build some models on this particular data. So you try to make sure that you either delete the missing values if they are very less in number or you can maybe replace the missing values with either a mean value or a median value.
- Similarly, we do not want any extreme values in our data, basically the outliers. So you can also treat your outliers. Again, if they are very less in number maybe you can remove them or you can maybe replace your outlier values with average values or the median value.

- You can also use the clustering technique to divide your data into different chunks and with each chunk you can either replace the missing values or the outlier values with the average values of the chunk. So that's a more accurate way of doing it. Then you can also create dummy variables if necessary.

Now after you clean your data, the next important step is to profile your data. So what do we mean by profiling the data? You want to understand the relationship between all the variables in your dataset and the target variable. Now what is a target variable? Target variable is that variable that you are trying to understand through your problem statement. So for example, in this case, we are trying to understand what are the characteristics of patients who have heart disease as opposed to the characteristics of patients who do not have the heart disease. So our target variable is basically the DV variable in the dataset which either says 1 or 0. Either the patient suffers from a heart disease or the patient does not suffer from a heart disease. Now you can use your data manipulations and visualizations that you have created to your aid. So we have

already done a bit of profiling of the data as you saw in your previous sessions.

Now once you have profiled your data and you have understood which variable impacts your target variable, then you can go ahead and decide what model you want to build. So based on our problem statement, we decided that we will be taking a look at two types of models and they are basically classification models. One is logistic regression. So you build a classification model to predict the likelihood of a patient having a heart disease and you can also build the same classification model using another algorithm called as a decision tree algorithm.

So now let's take a look at how we can profile our data and what kind of insights can we get from profiling our heart disease dataset.

So now you can see here that I have actually profiled the numeric variables separately from the categorical variables. Now what do we mean by profiling? As I told you, we are trying to see is there a difference in all the attributes in the particular dataset so that we can differentiate patients who suffer from heart disease from patients who do not suffer from heart disease. So basically I

am trying to find out in this case, for all the numeric variables in the dataset like age, then the resting blood pressure, then the cholesterol level, maximum heart rate achieved and the oldpeak variable.

What is the average value split by the DV? So this means that in this case, let's take an example here. I have split it and I have found out the average age of people who have a heart condition is 56 and the average age of people who do not have a heart condition is 52. So they are not very far apart. So in fact, 56 is around 7% greater than 52. So that's the difference that I have mentioned here. And you can see that not much difference across DV for age, blood pressure and cholesterol. So they are like 7%, 4% and 4%. So for the thalach variable I can say that patients who have a heart condition have very low values of maximum heart rate achieved as opposed to patients who do not have a heart condition which is highlighted in green. Similarly, you can see the difference here as well for the oldpeak values.

Now for categorical variables, what I have done is basically I have taken a frequency distribution. So I just want to see split by DV, what is the distribution of gender? Male and female. So this is the frequency

distribution. And I have just converted this table into a percentage of patients across gender. So now why am I doing this here? So I just want to see in general, is it true that a lot more patients who are females seem to be having a heart condition or is it males who seem to be having a heart condition. I just want to see those kinds of differences across categorical variables. So first thing is you observe the total. You can see that in general in this dataset, there are a lot more 1's than 0's in gender. That means that there a lot more females than males in this dataset- 206 females and 97 males. Now within the males, if you see 74% of the males do not have a heart condition. However, if you take a look at the females, you see that it's 45% and 50% - in the sense that you cannot see any variation or difference between them. So based on gender = female, you cannot categorize patients who have a heart condition versus patients who do not have a heart condition.

However, we could maybe generalize by looking at this information that there is a high percentage of males who do not suffer from the heart disease. This is something that we can see. And we can flag this indicator 0 and we can use this in our data,

maybe the model will pick up this particular variable as important.

Now what about chest pain type? Let's look at chest pain type. Again you see that chest pain type 4 is maximum 144 patients have chest pain type 4 followed by chest pain type 3 and 1 seems to be the least in the data. And looking at this, what do we infer? We see that there is very high chance that the person suffers from heart disease for chest pain type 4. So you can see 73% have the condition. Whereas for the rest of the scenarios whenever chest pain type is 1, 2 or 3, you can see that mostly the patients do not suffer from a heart condition. So again, there is a strong variation and we can use cp variable in our model as well. It's able to differentiate between 0 and 1.

Now what about the variable fbs? Again fbs 0 and 1, I see that there is not much variation across the DV. So this variable might not be significant.

Now what about exang variable? You can see that exang variable 1 and 0, yes, there is a variation split by 0 and 1. So we can use this variable as well. You can see 77% and 23%.

Now what about the slope variable in the dataset? You can see that the variations are

not significant for slope = 2 and 3. Maybe 2 slightly, 3 they are almost same. So using slope = 3, I cannot find out whether a patient has a heart condition or not.- because there is equal chance that there are equal number of patients in both the categories. However, when slope = 1, I see that mostly patients do not have a heart condition, 75% fall in that category. So this is all based on your absolute data. This is not prediction. Looking at the data, you are just trying to find the frequency distribution.

Now again for summary of heart condition, that is the thal variable. Again the variations between DV are significant. You can see that, except for? which is just a missing value. So there are just 2 variables here that relate to it. So we can ignore that. And also we can see that for summary of heart condition, thal =3 has the maximum number of patients. 166 rows are contributed to thal = 3.

Similarly, by taking a look at variable restecg, again we can see that the variations are significant for restecg 0. It's also significant for restecg =1. For 2, there is no difference across DV 0 and 1. There are patients in both the categories. So based on whether there is a variation or

not, you can shortlist variables that you want to use in the model. If there is not much difference between the DV's, then there is no use of using the model because then that variable will not help us separate patients who suffer from heart condition with patients who do not suffer from a heart condition. So that is the reason why we profile our data. So as soon as you have a dataset, when you start your data exploration, you can try and visualize this data as we saw in the previous session. Even through the use of histograms and Boxplots, we were able to separate patients who suffer from a heart disease with patients who do not suffer from a heart disease. Apart from that you can create a simple profiling like this and shortlist variables where there is a variation across DV and try and use them in your model.

Slide 4:

Now this marks the end of session for the data exploration and preparation for the running case study.