Slide-1

In this topic, we will talk about missing values.

Remember that in the data preparation stage, we usually have to clean data because as raw data is rarely ready to use.

In the last session, we had looked at one of the most common data cleaning issues which is dealing with outliers.

Now let's take a look what to do when we have missing values.

Slide-2

Now why should we worry when data that is missing, isn't it missing anyway?

Remember that missing values sometimes are a problem, specially if there is a pattern to the missing data. If your data is missing completely at random, there is no pattern then it is less of an issue than when there is data that is missing in a pattern.

An example could be many times when you ask customers to fill out a survey and let's say that there is a question about income. You will find very often that customers that have high levels of income choose not to answer this question in a larger proportion than customers that are at lower levels of income.

So if you look at your output data the missing values will not be at random, you will see that more customers that are high value, that are at high income tend not to answer that question.

So if there is a pattern to your missing data then you certainly have a very big problem.

But what happens when you have missing data?

Remember, if you have missing data specially not at random then you are introducing bias in your sample. In other words, the output of your analysis will not be representative, it will be biased.

So what should we do about the data that is missing?

One thing is we should certainly go back to the extraction process and see if there was some mistake in the extraction process because of which the data is missing.

Sometimes you may need to go back and redo your data collection to make sure that missing values are captured, but off course that may not be an option all the time.

So what do we really do about the missing data?

Again similar to outlier treatment, there are 3 options, ignore missing values, delete the missing value records or impute values for the missing data.

Slide-3

So the first step in looking at missing values is to do an assessment of your data set and see which of these variables have missing values and what is the proportion of missing data.

For example, if we were looking at the telecom data set, I would look at the number of observations and the number of missing for each variable, you can see that the subscriber id for example has no missing values but minused4 has high missing values 134, prom 5 has 370 missing. Now how much missing is a problem? Again it depends on the amount of data that is available to you. If you have lots of data then having 3 percent or 10 percent of your value is missing may or may not be a problem. However if you have very limited data then having even 2 percent missing values may be a problem.

In this particular example, 134 missing out of 12,500 may not be a big problem because we have 12,500 data points for each variable.

Is there a pattern to this missing data? What do we mean by pattern? When we look at the right most table, we are simply looking at overall missing values, but it may be more important to look at missing values by attrition, missing values by promotion that is… is there a difference in the percentage of missing values between customers who have left and customers who haven't left? That is what we need by looking at pattern of the missing data.

Slide-4

Now supposing we have identified some variables with missing values.

How do we actually deal with that in the data preparation stage?

The easiest thing is to ignore the missing values, the data is missing, I am not going to do anything about it. Of course it is the easiest thing to implement especially when you have lots of data.

The second thing that potentially you can do simply delete the values with missing data. Remember if you delete a value for that variable, it is essentially same as eliminating entire observation because you can't use the entire observation. So if you are deleting variable values that are missing remember there could be multiplicative impact because you are going to lose records where other variables also have values.

What happens when dependent variable is missing, when your target variable is missing?

Remember all data is not the same, some variables will be very important to your analysis and other variables may not be of primary importance may be only secondary importance.

So again if your critical variables have missing values, you may need to spend some effort in figuring out what to do with the missing data. Especially if your target variable has missing values, you may need to spend some time in figuring out what is the problem with the target variable, why are the values missing.

It is very important to solve target variable missing values separately from all the other independent variables.

Slide-5

Now if you don't want to delete the records with missing values or the variables with missing values. Another option is actually treat your missing data, in other words come up with suitable imputed values for your missing data. Imagine that you had usage data and let's say that for minused5 this particular value was missing.

Then I could come with a suitable imputation, one could be just a general average for minused5. But similar to how we looked at it in outlier treatment, it may be better to do a similar case substitution rather than a global case substitution. In other words, find other customers who are similar in their usage to this particular customer in the first 4 months and then on that subset of customers look at an average for minused5 and then basically use that average as a substitute for the missing value in minused5.

Again remember that effort that you put into your missing data treatment will depend on how much data is available to you and whether there is a pattern to your missing data. If you have lots of data, then a simple treatment for missing data may simply be just ignore or delete the missing values. If you have very limited data, then you may need to put in a lot of effort to impute values for the missing data.

Slide-6

A couple of things to remember when treating missing data.

1. Don't just replace all missing values with some constant, there has to be a logic to your imputation and remember when you replace missing values with one constant then you are reducing the variation in that variable.

There are many ways of imputing values into a variable, there could be single imputation or multiple imputations, single imputation is when for every missing value in a particular variable you are going to use the same substitute. Multiple imputation requires lot more effort, where for example you may run a regression to predict the value for the missing values for that variable using all the other variables as independent variables.

Off course multiple imputation is computationally is lot more intensive then a single imputation.

Again whether you use a multiple imputation or single imputation will depend on how large your missing data problem is.

Remember if you have lots of data, if you have data in the 100's or 1000's or millions of records then some missing data may not have a big impact on your analysis.

What if dependent has missing values? Should you impute dependent variable values?

Absolutely not, it is a very bad idea to impute dependent variable values, you may want to investigate your data extraction process and see if you can get the missing values back. But it is a bad idea to impute a value for dependent variable missing values.

Slide-7

Let's just make sure that we understand what we need by missing data. For example let's take a look at this table here. This is coming from the telecom data set that we have seen in the data exploration and the data preparation sections.

Minuse1, minuse2 are usage in minutes after the customer signs up for the service. So for example the first row has subscriber id is 19164958, this person has used 57 mins in the first month after signing up for the cell phone service, 21 mins in the second month, 40 mins in the third month and so on. This data is from long time ago from 2001 which is why the number of mins is so low, this is when cell phones were still being adopted early stage.

So remember every row is usage for one customer. Now supposing I were to ask you, how many missing values exist in this table. Most of us will count one, two, three, four and five and respond that there are 5 missing values. Some of us may also include the zero's in our calculation and say that there are 7 missing values, but neither of those answers is right.

Why is that, remember missing data is when you really don't have any idea what that value needs to be. In this particular data set this is usage data.

Remember that, if you have missing data for example for this customer 43061957 this could very well be missing simply because the customer has quit the service. If the customer has quit the service we shouldn't see any data in this 7th month, so if they have quit in the 6th month then ideally I should not expect to see any values in the 7th month or the 8th month, therefore is this missing data? No this is what we should expect.

However what about this cell, this customer has data for first 4 months and also for the 6th, 7th and 8th month, this is really missing; there is no reason why they should not be any data for the 5th month. If in fact customer had not used the service, we would expect to see a zero here, but certainly not a missing a value, so this particular cell has a missing value.

So if I ask you same question again how many missing values are there in this data set, really it is two this is missing and this is missing.

Remember that you must think about missing data logically before concluding the data is missing. Sometimes the data values are blank because they are supposed to be black.

Sometimes zero may be a value and not a missing value, so when we do a missing data check we must view some basic sanity measures to understand whether or not the data is missing.

Slide-8

So what we have looked at so far within the data preparation process is really data cleaning, data cleaning involves checking for outliers and checking for missing values and if we find outliers and missing values, how to treat them using imputation methods. Imputation for outliers and missing values, remember should be done only if you have very limited data and every data point is valuable.

Typically if you have lots and lots of data records set are in 100's or 1000's or in millions, it may be easier to simply ignore or delete both outliers and missing values.

Now let's take a look at next step with data preparation which is transforming data.

Slide-9

So when we talk about data transformation within data preparation, we are talking primarily dealing with

Qualitative variables

Categorical variables

Derived variables

Transformed variables

Slide-10

Let's start with qualitative variables, now in the data exploration session, we had looked at how to transform to qualitative variables. We had said that sometimes qualitative variables contain information that may not be usable directly in the model. In other words that qualitative values may need to be converted into quantitative data.

For example, supposing I have customer data set and have a gender variable, the gender variable has values, male and female. Now easiest way for me to use the gender information in my analysis is to simply recode the gender variables as male equals to 1 and female equals to 0. Similarly if I had customer type as high, medium and low, I would recode as 1, 2 and 3, so this is a very simple transformation for qualitative variables.

Sometimes the categories in the qualitative variables may be too many, in which case I may need to do a more meaningful substitution. For example grocery versus non grocery if I was looking at item purchased. Supposing I had a variable called profession and it had lot of different values, I might transform that into say blue collar or white collar.

Remember the substitution obviously needs to add value to the data and help in generating the answer for the problem being investigated.

Slide-11

Supposing I have some data here which of these is a qualitative variable, clearly the type is a qualitative variable because it has values like luxury, sedan, compact and so on.

If I wanted to use the type of the car in my data analysis, I would have to recode this variable as 1,2,3 where perhaps luxury is 1, sedan is 2, compact is 3, hatchback is 4 and so on.

Slide-12

In the next section, we will take a look at how to deal with categorical variables and derived variables.