Slide 1

In the last session, we had looked at Data Preparation and specifically looked at missing value treatments as well as preparation of qualitative variables. In this session, we will take a look at Data Preparation for categorical variables and derived variables.

Slide 2

We have looked at categorical variables previously in the data exploration session. Categorical variables are essentially variables that have data in levels; categorical variables could have quantitative values or qualitative values. For example, I could have a categorical quantitative variable called satisfaction with the purchase process, the values in that variable are 1, 2, and 3; 1 being very satisfied, 2 being neutral and 3 being dissatisfied. So, this is a quantitative categorical variable, of course they could be qualitative categorical variables, for example gender, male/female, a location, a city, this would be qualitative categorical variables. Now typically how do we prepare the categorical data? Usually we may need to create in a special way categorical variables essentially what are called dummy variables.

Slide 3

What is a dummy variable? A dummy variable is also called an indicator variable, they have only 2 values, 0 and 1. 1 indicating the presence of an effect or a value. For example if I had a gender variable male, female then I may want to code the gender variable as 0 = male, 1 = female. Now, this is a simple example of a dummy variable with only 2 levels, but what if I had a dummy variable with 3 levels?

If you remember the car example that we had seen in the previous session, we had a car type variable as a qualitative variable and it had values of Sedan, Compact and Luxury. One of the ways we can use this data in an analysis we have to convert this into quantitative variable in which case you may think that the easiest way to do a conversion is to do a re-coding that we had looked at which is perhaps Sedan = 1, Compact = 2 and Luxury = 3. But very often that is not how we will convert this variable to a quantitative variable, we will in fact convert this in a very different way, instead of saying Sedan = 1, Compact = 2 and Luxury = 3, what we will probably end up doing and what we will probably end up using more often is dummy variables. I will create 3 different dummy variables from this car type, the first dummy let's say I call it a Sedan dummy will have values of 0 and 1, it will be 0 where the car type is not Sedan, 1 if the car type is Sedan, similarly I will create a second dummy variable and call it Compact dummy which will values of 0 where car type is not Compact and 1 if it is Compact and a 3rd dummy which is a Luxury dummy. To make it clear I have simply listed the output in this table here.

Slide 4

Imagine that we had car type for the first car as Sedan, then the Sedan dummy for that record will have a value of 1, the Compact dummy will have a value of 0 and the Luxury dummy will have a value of 0, because this car is a Sedan and it is not a Compact and not a Luxury. Similarly at the 5th row for a luxury car Sedan dummy variable value will have a value of 0, Compact dummy will have a value of 0 but Luxury dummy will have a value of 1. So, notice what we have done here, instead of having one single variable re-coded as quantitive variables with values of 1, 2, 3 we are creating 3 different variables.

Why do we do that?

This will become a lot more clearer when we start doing and creating some models but essentially if we use one variable think about it as getting one response or one impact across different car types. So, if I had one car type variable that had Sedan,

Compact and Luxury whatever analysis I do is going to have one response or one impact for the car type, whereas if I had 3 different dummy variables, Sedan dummy, Compact dummy and Luxury dummy then I have the luxury of getting different impacts based on the different types of the cars. Again if this is not very clear to you right now at the modeling process this will become a lot clearer.

Slide 5

There is also something called dummy variables and dummy variable trap, again something that we will discuss extensively when we start doing some modeling. One final thing about dummy variables, in the examples that I have used we have had two categories or 3 categories of a variable in which case we have created that many dummy variables. Now, if you have a categorical variable with n levels then ideally you will want to create n number of dummy variables. But what if we had categorical variable with hundreds of levels, do we really create dummies for all levels? As an example supposing we have a variable called item purchased, now item purchased in a grocery store could be anything you know there would be hundreds or thousands of items purchase possible responses of variables values that we will have. Do we want to create a dummies for all items purchased? Probably not, we would want to look at why we are running our analysis and what we want to do with that analysis and we may need to aggregate those levels to a meaningful level. For example I may say that I am just going to take the item purchased and aggregate that to grocery, non grocery, and household item, in which case I will create 3 different dummies one for grocery, one for non grocery and one for household instead of hundreds of dummies for every item purchased.

We will have to take a lot at what is that meaningful aggregation and whether there is enough data for each aggregation.

Slide 6

So, we have looked at how to prepare categorical variables typically convert them to quantitative and create dummy variables. Now let's take a look at what we do in terms of derived variable data preparation.

Now, derived variables is something that we have looked at data exploration, essentially derived variables are new variables in our data set that are created from existing variable. The simplest form of derived variables are those that involve very simple calculations, that is given a birth date can I derive the age, given height and weight can I derive the BMI, given usage can I bucket customers into low, medium and high. Why do we really need these derived variables?

Slide 7

Because sometimes the information that is captured in the raw data may not be usable directly or may not fit the actual business need. So we may need new variables because the usability of information is better in a derived variable or we may need some recognition of patterns at different levels of aggregation than what is applied in the raw data.

Slide 8

We have looked at derived variables in the previous section when we looked at categorical variables because we are really creating dummy variables from the categorical levels, but there are other examples of derived variables, for example lag variables or interaction variables.

Slide 9

What are lag variables? Sometimes the impact of a particular variable on outcome variable may have a time delay. A simple example could be supposing we are looking at the impact of inflation on sales, now it is quite possible that the impact of sales from the current period is driven more by the inflation in the

previous period than in the current period, so in order to capture that I have to create a time delayed inflation variable where I can say that sales in the current period is a function of inflation of the previous period. That is when I will create a lag variable.

Slide 10

It is a very simple concept, so supposing to give a business example we were looking at the impact of advertising on price, now again depending on the kind of advertising, some advertising may have a longer term impact, so I may say that the impact on current period sales is some function of current period advertising as well as previous period advertising. That is people have memory of the advertising that has been done in the past and the more advertising you have done in the past the more they are aware of your product. So the impact of advertising on sales is not short term, it is more a long term impact. So if I wanted to use that concept in an analysis I would need to create a lag of the advertising variable, in this particular example the lag of the advertising variable is simply shifting the advertising values one cell down.

Another common use of lag variables is when you run time series models and you look at auto correlation that is the volume of sales in period one has an impact on the volume of sales in the next period, this is sort of like base level. So lag variables are one example of derived variables.

Slide 11

Sometimes you may also need interaction variables, again where would interaction variables be needed? Very often in regression models we assume that the impact of independent variables on the dependent variables is additive, that there is a linear relationship. Sometimes that relationship is not linear; sometimes that relationship may be multiplicative.

Slide 12

We can explain that better with an example. Supposing that we were doing advertising and we were trying to do capture impact of advertising on sales. Now advertising can be done let's say using both TV and Radio, now when we run TV we expect a certain impact on sales, when we run Radio campaigns we expect certain other impact on sales, but there may be a potential impact of TV + Radio at the same time that is over and above the individual impacts of TV and Radio, in some sense a re-enforcement impact, the fact that you are running both TV and Radio at the same time may have an impact which is equivalent to saying 1+1 > 2.

Again when do we need interaction variables? When we suspect that we have an interaction impact there is an additional impact of having 2 variables run at the same time. Again some of this comes from some understanding of the business situation and variables and domain knowledge, some of it is trial and error. But, at the data preparation stage you need to think through this and figure out what variables will need to be prepared in order to use in the model.

Slide 13

Now let's look into final kind of data preparation which is data transformation.

Slide 14

What are transformed variable? Essentially when their underlying structure has been changed, examples of data transformation includes data normalization, log of data, taking logarithm of data values, squared cubed and other non linear transformation. Let's take a look at what each of these are and why we use them.

Slide 15

Let's start with data normalization. Sometimes data is normalized or scaled if there are variables with high variations in magnitude, for example supposing we have a couple of different variables in one data set, variable 1 has a variation with a minimum of 0.01 and a maximum value of 0.1, variation 2 has minimum of 400 and a maximum of 100,000, now these variables are on very, very different scales and sometimes it may make sense especially when we are calculating distance measures to bring all these variable values onto the same scale. You may not need to do this all the time but especially if you have many variables with very, very different levels it may be a good idea to normalize your data.

Slide 16

How do we normalize data?

There are two primary ways of data normalization; one is to use the minimum and the maximum criteria for normalization, the other is to calculate Z score of a normal score. So if I want to use the minimum and maximum variation the formula that I am going to use is listed here. The outcome

(variable = v min/max – min * (new_max – new_min) + new_min

So to explain this supposing that I have a variable with minimum of 30 and maximum of 340, let's say that the mean is 125 and the standard deviation is 21, if I want to use minimum maximum normalization I want to change that variable to a range of 0-1, so essentially I want to take this variable which has a minimum of 30 and maximum of 340 and I want to normalize that to a variable with a range of minimum of 0 and a maximum of 1. So how do I do that? Supposing that I had a value of 200, 200 using the minimum maximum normalization will now become

200-30/340-30*1-0+0 = 0.54387

New max is 1 and new minimum is 0, plus the new minium. In other words a value of 200 in a data series with a minimum of 30 and maximum of 340 now

becomes a value of 0.54387 in a data series with a minimum of 0 and a maximum of 1.

Remember in data normalization we are not doing anything to the shape of the data, we are simply scaling down the data.

Slide 17

I could use the same normalization technique but using a Z score normalization in which case I would simply take the value minus the mean divided by standard deviation

Value - Mean/SD

So for the same example if I wanted a value of 200 to be a normalized value that 200 will now become 3.57

200-125/21

So data normalization is done simply to get all the variables onto the same scale.

Slide 18

But there are other kinds of transformations as well that may be required in the data preparation stage, sometimes you may want to transform the data structure itself, for example a linear regression model will require that the independent variables are normally distributed. Now your underlying data may or may not be normally distributed but skewed distributions can be made a little less skewed by doing a log transformation.

Slide 19

For example the left side shows data before the transformation and the right side shows the same data after doing a logarithm, you can see that the data is lot more normal on the right side. So, sometimes data transformations or log transformation may be used to make the data a lot more usable for the particular model that we have in mind.

Slide 20

But there are other transformations that are possible as well.

Non-linear transformations.

Slide 21

The simplest way to explain a non-linear transformation is to look at this visualization.

Slide 22

Imagine that we had some situation where we had diminishing returns. What are diminishing returns? Imagine that you were looking at marketing as driver of sales; essentially the logic there is that for every dollar I spend on marketing, I expect a certain increase in my sales. But think about the marketing spend, should we expect that every additional dollar on marketing will give you the same return to sales? Not necessarily. Imagine a company that was just starting out with the company. You know marketing will return a lot of sales in the initial levels of marketing but as they do more and more marketing, for every additional dollar in marketing, they may not get same levels of sales return. In other words, there is a diminishing returns relationship between sales and marketing.

Slide 23

For a certain point, increase in marketing will lead to an increase in sales but after a certain point the rate of return of sales for every dollar additionally spent on marketing will reduce. Now if I had a relationship like this that I wanted to capture, let's say this is not a linear relationship, this is a non-linear relationship; one way that I could capture this information is to create a variable that would have a quadratic term. So to use the sales and marketing example, I would

say sales is a function of the marketing square rather than just marketing. Again the actual implementation of these higher order terms and how they impact your analysis will be better understood once we start creating some models. But again at the data preparation stage if we expect that there is diminishing returns based on our visualization of the data, then we may want to create some higher order terms as part of the data preparation stage.

Slide 24

Another important part of the data preparation is to do with continuous data. Most of us think if our data is numeric and continuous, there really doesn't need to be data preparation and most often that may be the case. However, sometimes even when we have continuous data, we may want to discretize the data, we may want to bin the data. Why would we want to do that? It may aid interpretation and it may improve actionability. For example, supposing we are trying to figure out what should be a credit limit for a customer? And one of the factors that we want to use to set credit limit is income. Now if you look at an income variable, that's a continuous variable. But if I wanted to implement a strategy based on income which sort of an income variable would be more useful to me, should I have an income variable that has values continuously from 20,000 to 15,0000, where I could have 20,000, 20,001, 20,500 and so on, or is it easier for me to build strategy on a variable that has categories of income, high, medium, low, wealthy, right? So in terms of actual implementation, the income could be better used as a categorical variable rather than a continuous variable, so what I would do in that case is actually take this continuous data and basically create bins of that data.

Slide 25

This process of binning data is also called discretization. There are two kinds of binning. You could do equal interval binning or equal frequency binning. Equal interval binning is when you take the

range of the data and you divide it into n equal intervals. If I wanted four groups, I would divide it into four equal intervals. Equal frequency binning is done when the data is divided into intervals that all have the same number of observations. Again, some judgment may be required on deciding how many intervals to use or what that frequency is? And sometimes it is a trial and error method and sometimes it may depend of the business requirement.

Slide 26

But again as part of data preparation, you may need to bin data, even if your data is already quantitative and continuous. For example, if we take the telecom data and I want to classify my users as light, medium and heavy. How would I actually bin the data? One way to do it would be to add up all the minutes used across all the months and look at percentiles and maybe we classify everyone less than the 25th percentile as light users, between 25 and 75 as medium users and greater than 75 percentile value as heavy users. Sometimes the classification may come from an existing business definition. The business may already have a classification for customers as light, medium and heavy in which case you may want to simply use that.

Slide 27

So what we have looked in this section is data transformation as a part of the data preparation process. The final step in data preparation is data reduction. Data reduction is not always applied. When would we want to reduce data? And what do we mean by reducing data? It doesn't mean getting rid of data but essentially what it means is, if you have many, many variables, high dimensions in your data, you may want to reduce the dimensions in your data because processing that data may be very time consuming. Or sometimes you may even run into a situation where the number of observations are lower than the number of variables. Why is there a problem

with high dimensions? It is not always a problem but sometimes you may have high correlations between your variables which will lead to what is called multi colinearity issues and the additional variables may provide very little additional information.

Slide 28

So, if you do decide to reduce data, remember only in the context of having many, many variables, lot more variables than you need there are multiple dimension reduction techniques that you could use. One is to simply look at correlations and drop correlated variables, it's a very simple approach but it may not always be justifiable, also you may need some judgment in deciding which variables to drop. There are more sophisticated dimension reduction techniques, for example, doing a principal component analysis or a factor analysis and the principle behind these dimension reduction technique is to identify components that are essentially combinations of those multiple variables, and using those components in your modeling process rather than the actual variables. Again remember data reduction is only applied if we have many, many independent variables and we believe there are correlation issues because of having so many independent variables or that there are underlying components or factors that are driving behavior rather than the variables themselves.

Slide 29

So to do a quick recap of data preparation, we looked at Data cleaning, Data transformation, and Data reduction, all with the intention of making sure that the data that we are going to use for our analysis is actually usable. One final thing that we will cover as part of the data preparation process is sampling.

Slide 30

Many times as part of the modeling process or the analysis process we will create or analyze a particular sample data set and we will validate our results on a second data set. So sometimes at the end of data preparation, you may take your master dataset and split it into two datasets. You will create a training dataset which is essentially the dataset on which your analysis or your model is run and you will also create a validation dataset, this is the data set on which your analysis is validated. Sometimes, people will create 3 dataset, training, validation and the third test dataset. How are these datasets created? Essentially, random sampling, you take your master dataset and split it say 60-40 or 70-30 using random sampling. You use the 60% for the training dataset and the 40% for the validation dataset.

Slide 31

Again when you are building your datasets or partitioning your data, you will want to make sure you have balance samples, that you have representative samples. So this is true for any analysis, when you are building your data for data testing and data validation, your test dataset and validation dataset should be representative of your underlying data, you need to make sure that you have a balance sample. So for example, if you want to access response to a direct marketing campaign, you need to make sure that your underlying data for both the test and the validation have both respondents and non-respondents, sometimes, your data on responses may be very, very low. For example, typical response rates to direct marketing campaign tends to be 2-3%. So if you had 100,000 observations only 2% of those observations will have responses, in which case you may need to overweight the responses to non-responses.

Slide 32

So instead of having 2% response rate, you will boost it artificially to 20%. How would you do that? If I have 1000 respondents, instead of taking 20 respondents or

2 respondents, I will take 10% of total non-respondents and 50% of total respondents, and I will overweight the respondents relative to the non-respondents. The reason we do that is we want to make sure that our analysis can correctly predict response and with very, very low response rates in the real data it may be hard for your analysis to correctly differentiate between response and non-response. If you are overweighting, you know, creating an artificial sample in some sense, there are some correction factors that we have to apply when we come up with the final result to make sure that our results are correctly interpreted. However, at the end of data preparation, sometimes the data partitioning and balance samples may then be the outcome of your data preparation process.

Slide 33

To do a quick recap, in this section we had looked at different ways of preparing our data. We had looked at data cleaning, data transformation and data reduction techniques. We had also looked at data sampling, usually splitting your dataset into a validation sample and a training sample in order to validate your data.