

Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:05
Lecture:13

**INTRODUCTION TO STATISTICAL LEARNING AND DATA
PRE-PROCESSING | Business Intelligence & Analytics**

In practice, how, what does that mean? In models, there is a problem called multi-dimensional collinearity. Multicollinearity is a problem when you have y is a function of $f(x_1, x_2, x_3, \dots, x_p)$. You have p number of variables. Now, you know you are working on a problem like the bizocity score. So, anything that comes in your mind you include as a variable just because data is available. So, then you have say 25 variables or 30 variables or 50 variables.

I have seen people doing or working on problems with large number of variables. And it is not very uncommon in machine learning or data science sort of context, where the context is different. We will talk about it later and explanatory versus predictive problems. But from a statistical standpoint, multi-collinearity is not a desirable condition.

What does multicollinearity means? If you have p variables, there are variables which are strongly correlated to each other. For example, x_2 and x_3 are strongly correlated. Suppose, x_2 is date of birth and x_3 stands for age. Of course, date of birth needs to be coded, but it can be an sort of ordinal data. Now, you look at this.

These two variables are included in a model. As soon as you see this, you feel well, this does not look good. Why does not look good? Because one contains the other or they are strongly correlated. They are strongly correlated to each other. There is a strong correlation between them.

Or in other words, the information that is there in age, is already there in date of birth. You do not have to include both age and date of birth in a model to depict how old is a individual. And so, what is the reason why you should not do it? Because one particular aspect of an object is over represented. Suppose other aspects that you are trying to include is say, for example, income or it could be education. So, you are actually building the profile of individual or it could be region.

You have included different demographic variables to profile a customer. And then you have date of birth and age. Both talk about the same thing. So, what happens is when you build a model like a multiple regression model, each of this will get a coefficient. And what happens? Because one dimension is over represented, it affects the representation of other variables.

Or in other words, the age of the individual gets over represented in the model, which leads to under representation of other models, which is not desirable when you are building an explanatory model. We will talk about it a little more as we go, why we call or we highlight this problem in the context of explanatory models. So, how do we address multi collinearity? So, essentially the idea is if two variables are strongly correlated, have only one of them. You can do away with age, have date of birth alone or do away with date of birth have age alone, drop one variable. That is the principle.

Now, how do you do that? You can actually set up a correlation matrix, a diagonal matrix. And you have x_1, x_2, x_3 here, x_1, x_2, x_3, x_p here. And then you have all the correlations coming here. So, you can see which are the pairs of variables which are strongly correlated and drop one of them if they are strongly correlated. Set a criteria.

For example, if Pearson coefficient of correlation is greater than 0.6 or 0.7, drop one of them or 0.5. That is a judgment.

This is one way of addressing the problem. Look at the correlation coefficient. But that does not address the problem completely. Because it is only looking at pairs. It is also possible that the information is in x_1 , is contained in x_2, x_3, x_4, x_p .

In that case, what do you do? The correlations does not help because it is only looking at pairs. So, then you need to model x_1 or you need to model each independent variable, independent variables meaning x_1 to x_p in terms of rest of the independent variables and have some measure of how x_1 explains rest of the variables. And that explanatory power of one variable through other variables is available in a measure called R^2 called coefficient of determination. I will explain this a bit later when I try to demonstrate you how a simple regression model works in order to build understanding about explanatory and predictive models. But in this, at this point, just be informed that R^2 , like the correlation coefficient is a measure of the explanatory power of a model or explanation of outcome variable in terms of rest of the variables.

So, therefore, R^2 is something that varies from 0 to 100. 100 meaning the x_1 is fully explained by x_2 to x_p . 0 meaning there is no correlation between x_1 and rest of the

variables. So, you can actually again use R^2 as a basis for determining whether a variable should be included or not. There is a specific measure called variance inflation factor or VIF, which is $1 / (1 - R^2)$. And generally if VIF is greater than 2, if VIF is greater than 2, then delete that variable. Delete which variable? It is x_1 's VIF that is determined. The idea being or the principle being the information in x_1 is already contained in x_2 to x_p because there is a high R^2 . So, x_1 is explained by x_2 to x_p , in other words.

And therefore, you do not need x_1 , when x_2 to x_p are there. So, that is another way of addressing multicollinearity. So, you saw there is correlation matrix that is step 1 and then VIF as the basis for including or excluding variables in a model. So, this is the third principle as we have seen in deciding on variables. We have seen four principles.

We include if a variable is relevant, if a variable is useful in controlling or explaining the outcome. Third is well, do that but do not have too many. And fourth is see and ensure that data is available for the variable that you include. In some literature, this is also called auto correlation, especially in econometrics. Now, we move on to the next step.

So, we looked at step 1 and step 2 in the analytics process and what is step 3. Collect data and formulate model. Well, we are okay now. We have thought about what we are going to do. We have decided our variables.

We have ensured that variables are the right variables and data is available for the variables that we have selected to build the model. Now, what is stopping you now? So, go ahead and collect data and get the data and formulate the model, formulate the model. It does not even say fit the model. Formulation or specify the model and also of course, subsequently you have to see how the model works.

That is the step 4. So, collect data and formulate model is the next important step in the process of analytics. So, let me actually elaborate to you what does that step really mean. Here is an example of data collection and using data for model building. Here is an example of data that is collected and ready to be used for model building. And this data, I am charting here as an example of how you need to look at data before you use it for modelling.

So, this data corresponds to a fashion retailer in northern India, specifically in Faridabad and the retailer has shared with us data for, pilot data for a few years which pertains to sales of a particular store as I said. It is store sales, not only sales, but also discount. Discount is in red and this is sales. This is discount. And you can see that this is for the year 2008 and this starts with January and ends in, ends in December.



So, it is for the full year month wise aggregated data. Monthly aggregated data over one year. And you can see the different charts, the two charts of different scales. So, the scale of sales is in the left axis, sale of discount is in the right axis. And discount is in percentage and sales is in absolute units or in Indian rupees.

Now with this plot, if you look closely at this plot, what do you think? So, this is something I show students for them to think and respond. And often times they look at the data for some time and the typical responses I get which you also may be thinking now is, well sales and discount seem to be correlated sometimes and sometimes there is no correlation. They are trying to increase discount here, but still sales is decreasing. And this is fashion retail, should they be increasing discount so much. Well, in fashion if you have discounted sales, you know the affluent class of customers may not even go or look at that because they do not want such kind of products as fashion. So, all these sort of you know, business knowledge will come when they look at this data.

And also someone would towards the end look at the year. Sir, this is 2008 and you know that is a year there was a financial melt down. So, there you know you had a global financial crisis during the year 2008, that explains, you get into explanation that explains why they are giving discount and sales is not coming up and so on and so forth. And all this information is not unfortunately there in this chart. You are getting into a lot of hypothetical situations.

What is there in the chart? What is there for an analytical mind or an analyst who is going to use data to model, interpret and apply should pay attention to here. That is something often times people rarely pay attention, but some people do. They notice, well sales touch to zero here, but discount is still on. Sales touch to zero in the month of March, but of course it was declining and of course after the month of March, it is also going up. But we know that it is monthly aggregate. So, therefore for some period of time which covers a whole month, the store store sales is zero.

There was no sales in the store. As soon as this is highlighted, immediately the answers are well there was a strike, there was labour problem, there was short supply, there was financial crisis. You start giving a lot of reasons as to why there was no sales. It was all your conjectures or all your imaginations. These are not real reasons. Actually if you ask, is this a problem? Yes, this is a problem.


This should be a problem and analyst should actually spot when you look at data. And what is this problem? There is no sales. And instead of you deciding as a analyst whether the counter was down or there was some problem etc etc, you better ask. If in doubt ask, do not assume.

If in doubt ask, do not assume. So, since this was a pilot project we did, we contacted the CIO and showed to the CIO, you do not have any sales in the month of March, is that true? He said it is not true. Then what happened? Then he looked at the database and said well, we had a problem with data collection during that period. What happens is that at the end of the day the POS, the point of sale systems transfer data to the data centre, you know as a batch process. Every day at the end of the day. So, it does not transfer real time.

The data is transferred at the end of the day. And for one month there was a problem, there was a technical snag because of which data was not transferred to the database. Now what do you say? Was it a business problem? Ok, that there was no sales. So, that you should include that in the model. You know the general principle is that all patterns in the data should be captured to build a model. Otherwise you are not properly informing or training a model etc.

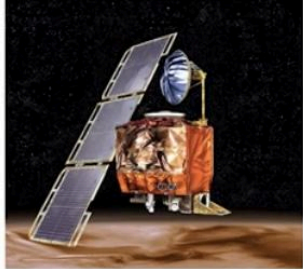
We will talk about it later. But here this is not something that actually happened to business but this is a technical issue. Therefore, the title of the slide is Inspect Data. Or this is, in data analysis we call it missing data problem. There is actually missing data in the data that you are going to analyse. There was the, actual phenomenon was happening which is about sales but we do not have data about it.

Now we have a situation when if you simply say it is zero, it is not correct. The model you could build but this model is not useful or the model is not learning properly because you do not have the right data. So, in real life before you go on to build models, when you collect data, there is a problem called data quality. Data quality is an issue in analytics or data analytics. And I will underline this because this is an important factor.

 INTRODUCTION TO STATISTICAL LEARNING AND DATA PRE
Data Quality

▶ **Good data characterized by (Han *et al.*, 2012):**

- ▶ Accuracy
- ▶ Completeness
- ▶ Consistency
- ▶ Timeliness
- ▶ Believability
- ▶ Interpretability



September 23, 1999: The Mars Climate Orbiter approached Mars 170km too close to the surface; atmospheric forces are believed to have destroyed the spacecraft.

Proximate Cause:

- Ground software used English units, while onboard software worked in metric. The discrepancy caused

The quality of the data determines quality of model. If you build a model using garbage, garbage is the output. We all know that. And therefore it is important to spend enough time on data preparation. And that is when you look at data and its characteristics and look at data quality, as an aspect of data. And your textbook talks about six attributes or six characteristics of data quality, which includes accuracy of data, whether the data is captured accurately or not. Then completeness of data, that is where the problem of missing data comes. You have a lot of missing data and therefore that is something that needs to be addressed before you use the data. And consistency. What is consistency? Consistency is, there are different aspects to consistency.

One aspect is the units of measurement should be in the same units. Data should be measured in the same units in all the records. For example, suppose you are working on

a customer loyalty program and you are trying to segment customers based on customer value, customer recency, frequency etc. But your customers are dispersed throughout the globe, you know in different regions of the globe. Suppose you want to build a model that covers the whole globe, that may not be the case often.

But suppose it cover multiple countries. Then your data is collected from multiple countries. So in some countries it may be dollars, some could be Indian rupees, in some places it may be rupees but a different rupee. So there are different currency units that are used in different countries. And there is a need to standardize it, otherwise all these measures mean different.

It is not consistent data. So in this slide there is an illustration of a problem which happened in 1999 because of the confusion of unit of measurement. This is a classical problem often used in data analytics. That is one of the NASA projects. A climatic orbiter actually crashed after reaching close to the Mars orbit and the reason after investigation was that there was a confusion in the units of measurement in distance measure. For example, distance was measured using feet and meter and the data got mixed up, you know in raw terms, you know some kind of a problem which has to do with consistency.

And timeliness, data should be collected using certain timeline. This is something I am going to explain later, what does timeliness means. There is something called cross-sectional data. Data, there is also time series data. In time series data, data should be collected at regular intervals. That regularity is important. There should be a constant frequency and if it is not followed the data is not useful. In cross-sectional data, data should be collected at the same time from multiple points and if you do not collect data at the same time then it is not useful. So depending on the nature of model and nature of problem, one should pay attention to timeliness and should check or inspect data to ensure what is the time at which the data was collected.

So therefore the key point I am trying to stress here is when you are shared with data and the data has lot of column, columns, columns, columns, columns, each column has a title, A B C D and you see a lot of data. Instead of looking at the data, first look at what are the column titles, what do they talk about, do you understand what it means.

Oftentimes the ABCDs will be coded titles and you do not understand what is the meaning of that variable itself. Always ask for metadata or description or data dictionary etc. which accompany databases where a proper description of each column is available. Ask for that column data or column metadata details. Do not accept a data set if it is coded and you do not understand what that code A B C D mean.

Get the descriptors and then look at the data and ensure that data is in the same unit, data is timely. Also look at, ask what time the data was collected. Have that specific information that defines the context of the data. Instead of simply accepting the data that is given, if you are working on a cross-sectional model, ensure that data is collected at the same time. You have to ask that question. Only then you will know. If you do not ask question and assume that it is cross-sectional data and it is not, you are actually doing a disservice to organization in building very misleading models which can misinform decision makers.

The other aspect is believability of the model or the data. When you look at data itself, sometimes you may develop doubt whether it is correct or not. For example, in the last slide, we saw that there is no sales in one month.

Can you believe it? Was the store closed for one month? Generally not. So there is some problem to believe this data. So you have doubt. You start asking questions. So believability may arise out of several aspects of the data. When you look at the data, look at the hygiene, are they in the same units, are they collected at the same time or if it is time series data, is there a regularity in it and where was the data collected, where was it stored etc.


Ask a lot of questions about data before you start using the data. And interpretability. Interpretability is another aspect of data. For example, you know somebody is talking about age. Age is one of the variables and some data, the age is 2000. Do you believe that somebody's age can be 2000? So this is common sense but it is not interpretable.

There is some issue. You immediately make out that there is some problem here. We saw that database will enforce certain rules, when you actually input say a time stamp etc but that actually takes care of this problem. So many of this data quality concern can be addressed through tools which are available for so-called anomaly detection. Anomaly detection.

Anomalies in the data can be detected automatically today through tools. There are tools available but what goes into these tools is you know, concerns and questions like this what we are dealing with. Before I move to the next slide, let me also talk about another trivial but potential problem when we look at data. This is a question I ask students. If somebody writes, how much money is this? Of course this is in Euros.

You can easily interpret the Euro sign. How much money is this? So as Indians typically we will respond that this is 1 Euro, but ask a German how much money is this? And they will immediately make out, this is 1000 Euros. This is 1000 Euros. The dot has a different meaning as a decimal separator in different countries, in different regions and

one must be aware of it. The comma separator is used in many countries like India, Britain, Americas and so on, but comma separator is not used in certain European countries. It is a dot that is used instead of comma and if you are not aware of it, again your interpretation of results will be misleading.



INTRODUCTION TO STATISTICAL LEARNING AND DATA PRE-
Data problems in the real world

- ▶ **Missing data**
- ▶ **Noise**
- ▶ **Inconsistency**
 - ▶ Units of measurement
- ▶ **SOLUTION: Discrepancy detection, Data pre-processing**

So data, data, data. Data is the basis. Let us move on to the next lesson. So we discuss data problems in the real world and I try to summarize that in the previous slide and I try to highlight the problem as a big problem. Data quality is very important in analytics. In class exercises it may not be important because you are only learning how to use algorithms but in analytics, you are applying algorithms for real life problems.

Missing data. There are techniques for addressing missing data. One is missing data substitution depending on the extent to which data is missing. So you can substitute with the average of the neighbors or average of the row etc. There are different techniques that are recommended in textbooks on data preparation for data mining.

And noise, you can use filters. You can use different techniques to address outliers. You must have learned those methods in data analysis, basic data analysis which is a prerequisite for this course and we have talked about inconsistency and other issues already. So let me move on. So data pre-processing is a very important step for data mining and analytics. And this becomes critically important because often times the analyst does not have any control in the production of the data.



INTRODUCTION TO STATISTICAL LEARNING AND DATA PRE-PROCESSING

Data pre-processing

- ▶ **Data cleaning**
 - ▶ Missing values → ignore the tuple, manual replacement, replacement following a method
 - ▶ Noisy data → smoothing techniques
- ▶ **Data integration**
 - ▶ Database normalization
- ▶ **Data transformation**
 - ▶ Scale, normalization
- ▶ **Data Reduction**
 - ▶ Aggregation

Data production is done somewhere. You are actually accessing secondary data, that is data which is already captured and stored, based on certain design of the database and data warehouse or some other source and you are given, granted this data. You did not design an instrument to collect this data, as you do in survey or experiments where the instrument is designed by the researcher and therefore you have complete control on the range of data, on the scale of data and you know what the different variables together constitute or construct etc. There you have a lot more control on the quality of data whereas here you do not have, because there is large volumes of data collected by someone stored somewhere, you need to actually work on that to make it useful. And therefore you apply techniques like this.

You know most of this we have already discussed. There is also data transformation that is required often. We are going to discuss that in another slide.