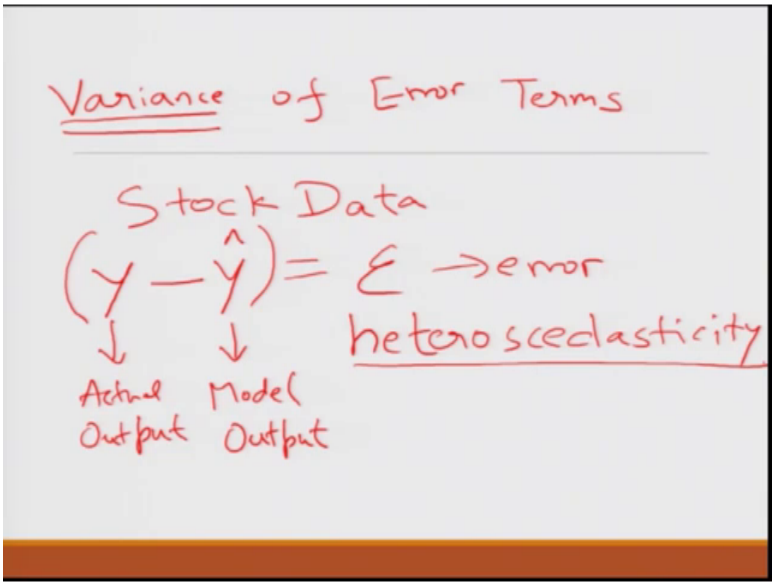
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics
Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 25
Machine Learning – (Part-2)

I welcome you all to another session in this Practitioners Course on Analytics Descriptive, Prescriptive, Predictive Analytics. In the last session, we were talking about machine learning the intuition behind it. The trips, the traps, the tricks, the intuitive understanding of machine learning, if you recall, we talked about two important traps; one was when? What if the relationship is non-linear? In case you are doing multiple regression, which is one of the most common forms of machine learning methods.

One of the most common forms of analytics methods or if you talk about the tool kit, it is one of the most used, tool in the tool kit. Second, we talked about correlation of error terms. Let us now, look at some more traps in regression.

(Refer Slide Time: 01:13)



Variance of Error Terms

Stock Data

$(y - \hat{y}) = \varepsilon \rightarrow$ error

heteroscedasticity

Actual Output

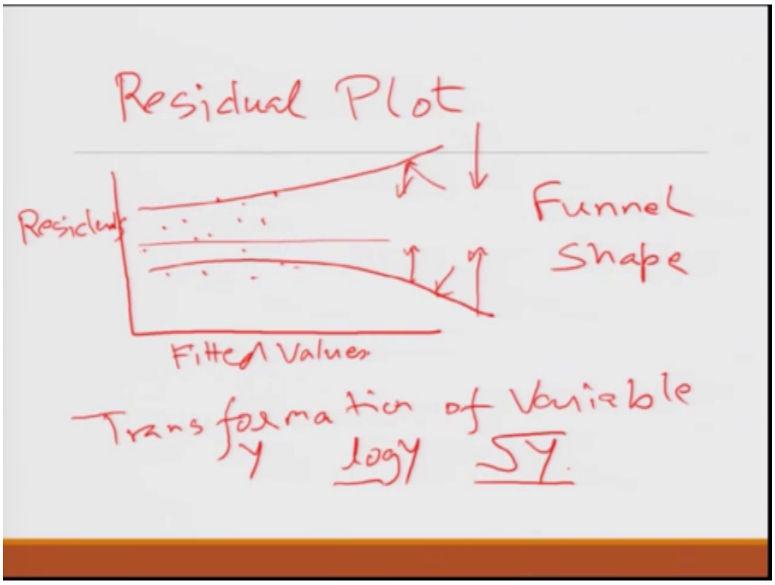
Model Output

So, another very important trap or one of the issues, which is often ignored, but is very important, is about variance of error terms in regression. Now, this becomes important, this is very-very common in for example, stock data, to, to talk in terms of more Layman terms. What this variance means, you know what is variance right. So, variance is basically a measure of the, the error, the, the measure of deviation, the square of the standard deviation is what is variance. Now, often what happens is that, the error terms.

So, for example, if Yb were output variable and Y caps is normally, this is your model output and this be actual output. So, you do a subtraction of these two terms, for all the data points, on which you want to run your model. Now, this is the error. Now, typically this error should remain constant, but what often happens is that, the error is actually not constant. So, this is something called, it is very, it is a tongue twister. This is called hetero sce dasticity.

So, heteroscedasticity means that your error terms, show a variance, the error terms are not correlation. Ideally, when you do a regression the assumption is that all your error terms, all your errors are same, there is no variation across error. So, the variation is almost minimal, but in case you have variance in the error terms, this is again an issue and one of the ways you look at, it is through something called residual plot.

(Refer Slide Time: 04:02)



So, what you do is that, if you recall in the previous slide, we had this errors. So, you, you plot these errors. So, these are called residuals, which is the difference between what you wanted and what you expected and these be for example, the values that you fitted. So, what you see is that, though the mean is there, if you, the mean may be following the straight line, but if you say look at the outer quartile range, if you look at the, the extremes. Say for example, the upper quartile and the lower quartile, if you see some kind of a funnel shape.

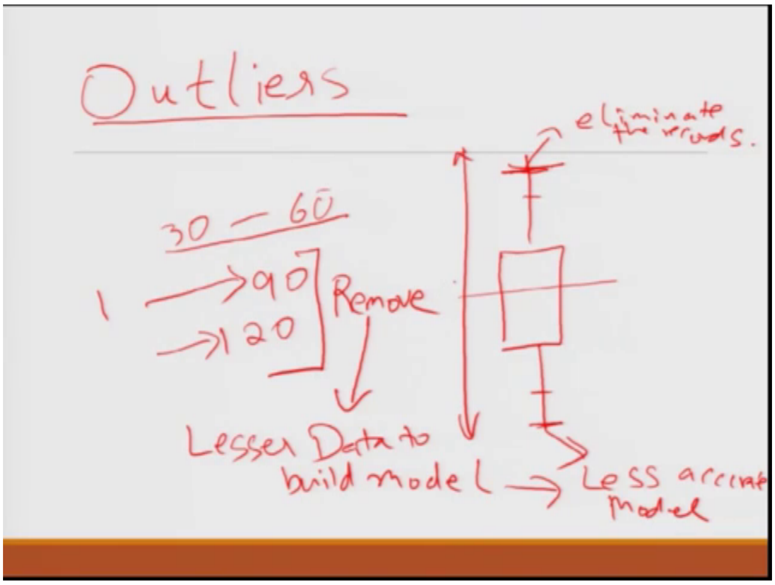
Now, this funnel shape basically shows that, in reality what is happening is that your errors are increasing. So, over a period of time the previous errors; So, if often if you analyze, you will find that the current error is the function of the previous error in a way. So, that is why the errors they, they expand a lot and this happens a lot in a stock market, the commodities market, because they are, they are sentiment driven.

And when, when things are sentiment driven, your previous sentiments actually impact your current sentiments and that is why sometimes the market crash or sometimes there is, what they call, runs in the market. The beer runs, the bull runs, that is primarily in a way. Due to this variance in error terms, this hetero scedasticity. Now, how do you deal with this? They are various ways. So, depending on the kind of plot, you see one of the simple ways to deal with. It is you do, do a transformation of a variable. This is one of the simplest way.

So, for example, earlier you were trying to plot Y now this showed a lot of hetero scedasticity. Now, instead of Y you plot log Y or depending on how this shape looks like you can go by something like root Y. So, the whole idea is these extreme points where the deviation is highest you tend to reduce them. So, you try to bring more and more reduction as the values inflate.

So, this is one of the common ways another huge consonant many a times even these transformations do not work and that is where you have to look beyond regression based model the other common issue very popular issue is that of outliers.

(Refer Slide Time: 07:15)



Now, outliers are simply values, which are not in the normal range of values that you would expect. So, for example, you are doing an analysis of, some kind of a score and the scores typically vary between say 30 to 60, but then suddenly you find the score of 90 or you find the score of 120. So, one of the ways is if you recall in your data analysis, when we do our box plots you, you tend to depending upon the amount of conservativeness. You want to have in this range; you try to eliminate the values, which are beyond the upper quartile range or the inner quartile range.

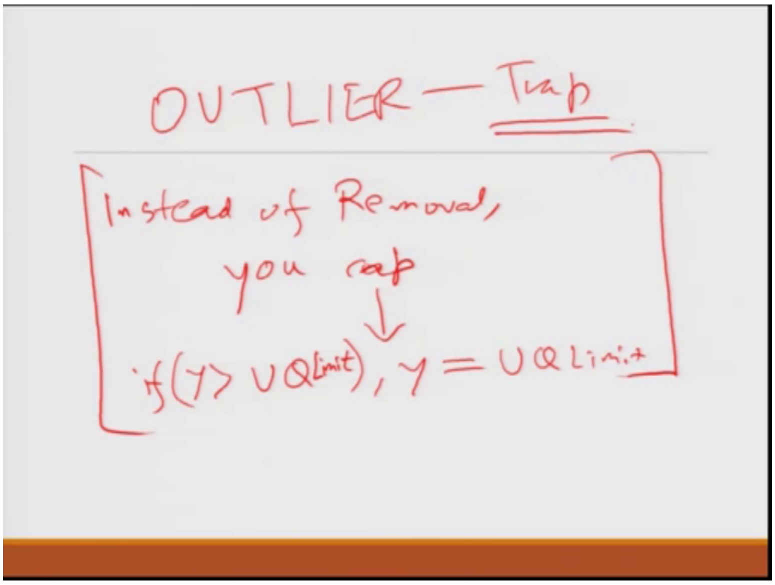
So, depending on different levels of conservativeness, you want, you simply one of the ways is to eliminate the records. So, for example, if you have some records where the scores are really. So, most of the scores are between 30 to 60, but you have one record, which is having a 90 value. Another record, which is having a 120 value; so, one of the simple ways is to simply remove.

But this often is not recommended, there are concerns with this and one of the major concerns is that, when you remove this means, you have lesser data to build model and this means less accurate model. So, in a way removal is a compromise you do. So, so think of this, what was the impact of an outlier, outlier scudder model, presence of outlier midair model less accurate; so, but when you remove the data point, because the number of data point reduce the model is again less accurate.

So, the overall benefit of removing an outlier, does not tend to be very high and often many a times, the regressions are done on data sets, which are not very huge and the data sets are not very huge. In that case every data points becomes precious.

So, removal is not often, though it may sound the easiest, but it is not best, of the methods. There is a whole domain in analytics, the whole field of study, which primarily focuses on outlier analysis. How to deal with outliers? One of the simple ways is instead of doing a removal is instead of removal you cap.

(Refer Slide Time: 10:21)



OUTLIER — Trap

Instead of Removal,
you cap
↓
if(Y> UQ(limit)), Y = UQ Limit

Cap means; you basically say that anything which is above my upper quartile range, if Y be greater than then, Y is equal to upper quartile limit.

So, if Y is greater than this limit you cap, it to upper quartile limit. So, in this way you do not lose the data; however, this, this will work only if you believe that it is a genuine outlier, many a times the outlier is, because you got. You got bad data, it was a printing mistake, it was a data that came from wrong measurement.

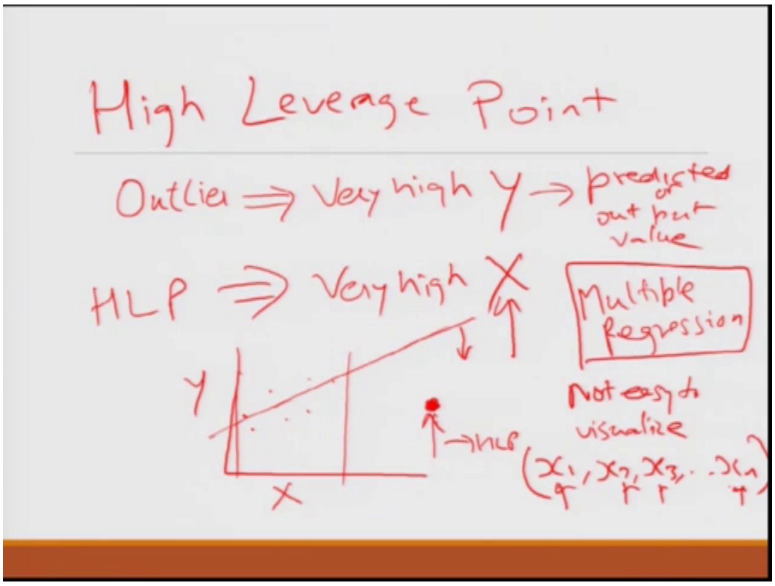
In those cases perhaps, elimination is the best. So, you have to look in to the data, but outlier often and specially, when you are dealing with data in the domain of, say survey analysis, a lot of this business related the marketing related data. Where some of the data comes from, not from only automated sources, there are a lot of sources of outliers. The errors that creep in, measurement errors type o errors.

For example, a very classic case is and this often happens in banks is that, when you apply for a loan, you have to fill a handled form and then there is a fellow, who then transcribes that, whatever you write in your form, with pen and paper, in to machine readable format. So, the transcription at times brings errors. So, those kind of errors you may well better eliminate it.

There are other ways of dealing with it. For example, you try to estimate, what could be a reasonable, way, reasonable guess and one of the ways to do this guess is, for example, you build a model on a data without the outlier, then try to estimate, what the parameters look like create a similarity between. The data point, which had an outlier versus, versus what the, the rest of the model predicts and then try to fit in somewhere, there are lot of ways itself.

It is a, complete field in itself, but the point is this OUTLIER is a trap; you need to take care of even before you build the model.

(Refer Slide Time: 13:21)



High Leverage Point

Outlier ⇒ Very high Y → predicted or output value

HLP ⇒ Very high X

Multiple Regression

Not easy to visualize

$(x_1, x_2, x_3, \ldots x_n)$

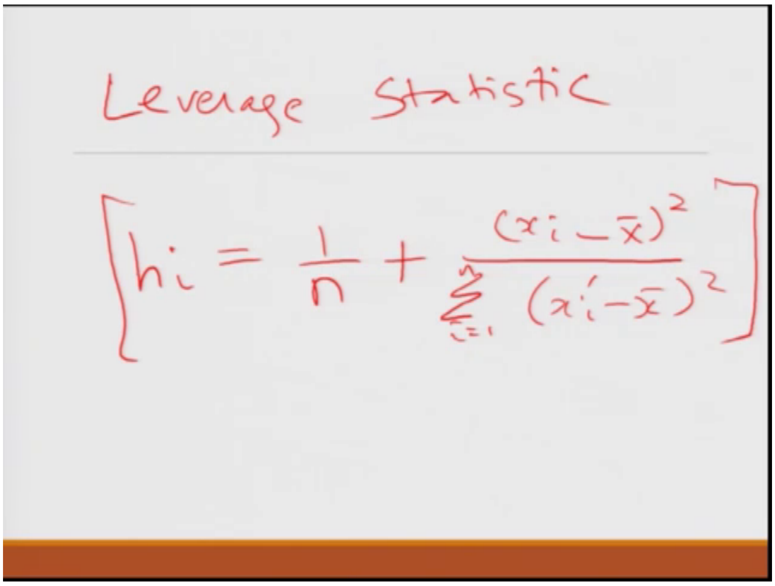
Now, very closely related to outlier is another trap, another concern, which is high leverage point; so, outlier very high Y, the predicted or output value. Now, high is something with very high X. So, in simple case of say a bivariate regression, this means X is a value. You know, for example, if you are plotting your Y versus X and this be the typical range and now, you have one value, which is sitting somewhere here.

So, rest of say for example, this was the regression line, but this value sitting out here. So, most of the X values are within this range and this is, this is kind of an outlier. The outlier is typically on the Y, but this is outlier on X which is called a high leverage point., there is a maths, mathematical reason, why it is called a leverage point, but you can intuitively understand that, just because of this particular high leverage point, this has the, impact. This can potentially skew this regression line.

So, these values again just like, outliers are. Traps high leverage points becomes trap, it becomes more of a trap, when you are dealing with, multiple regression, because in multiple regression, because there are multiple dimensions involved, you may not be able to visualize these high leverage points, not easy to visualize reason, being that the value may be within range for each of the different size.

So, multiple regression means, but instead of 1 x you have now, x 1, x 2, x 3, x n; Now, for each of these x 1, x 2, x 3, xn. The value is within range, but if you look at all of them combined, it is somewhere out of sync. Now, this is one of the reasons, why it is very difficult to identify high leverage points and yet they are the one that skew your regressions big time. They calculate something called a leverage statistic.

(Refer Slide Time: 16:39)



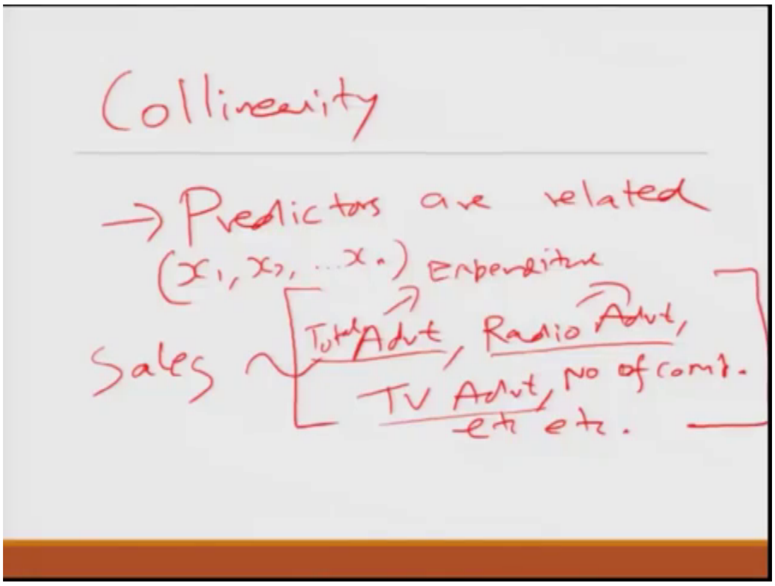
Leverage Statistic

$$\left[h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_{i'} - \bar{x})^2} \right]$$

I will, I will give the formula, but I will not get in to details of it, because it becomes, too maths, it is, may be those of you who are interested, they can get into the maths of it, though many softwares will simply calculate it for you the whole point.

However, is to keep in mind that you need to take care of, high leverage points. You need to eliminate them and you need to before, you eliminate them, you have to smartly track them.

(Refer Slide Time: 17:39)



Collinearity

→ Predictors are related

$(x_1, x_2, \ldots x_n)$ Expenditure

Sales ~ [Total Advt, Radio Advt, TV Advt, No of comp. etc etc.]

One final pit fall that we face is collinearity. Now, collinearity means your, predictors. By predictors I mean, your x 1, x 2, if these be the predictors, they are related. So, it means for example, you are trying to build a regression model on, let us for example, take the case of sales. Now, you do a regression of sales for multiple variables. So, from your different sources of data you have 100s of variables.

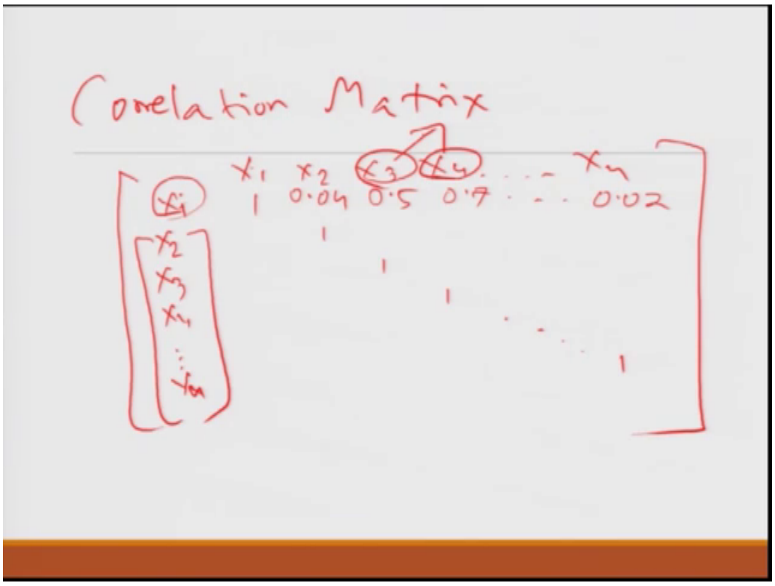
So, one of the variable may be advertisement and you also have a variable, which is radio advertisement, TV advertisement. So, this may be the amount of, expenditure made in a year or the amount of, people involved or whatever. So, most cases, expenditure will be, the most popular expenditure will be the most common proxy used for each of these. So, you have radio advertisement, you have TV advertisement, you have advertisement and then it may be lot of other things. For example, number of competitors etcetera, etcetera.

Now, if you notice these, this may be, let me be more express it. So, this is the total advertisement, there is radio advertisement, this TV advertisement. There may be other forms of advertisement there. Now, you may build a model and each of these variables can come out to be significant; however, if you look total advertisement is in a way correlated to your TV advertisement and your radio advertisement, which also means that the TV advertisement and radio advertisement can also be correlated, given that, that the company has limited budgets for advertisement.

So, in a sense, there can be multiple variables, which are all correlated or in the sense try to predict the same thing in a different way. So, these are, this is the tricky situation and yeah one of the simplest way to deal this is, if you find many of the these variables are correlated, first of all the issue is how do you find? Which of the variables are correlated?

In this case yes, from intuitive business stand point, you could, make, analysis and you can figure out that ok, these variables look seem to correlated, what if you have huge number of variables slightly more technical variables or maybe you are doing an analysis, where you do not understand the business very properly.

(Refer Slide Time: 20:54)

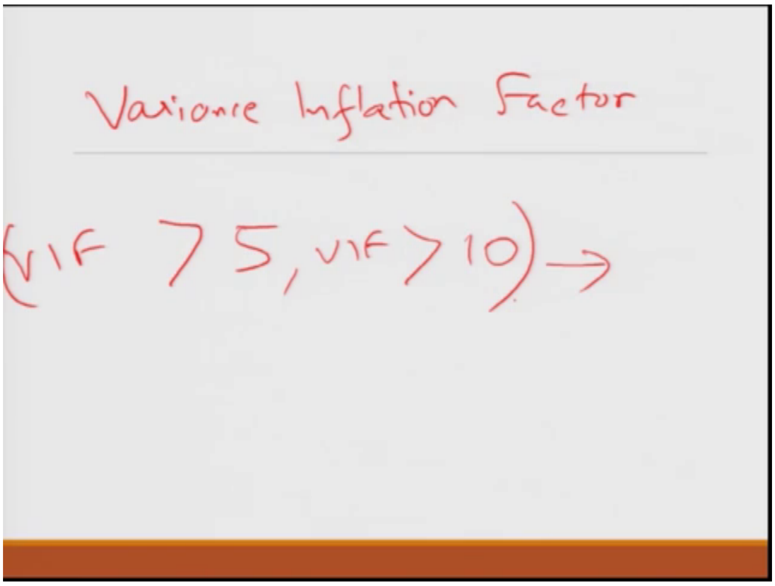


So, what do you do. So, one of the methods is correlation matrix. So, in correlation matrix, what you do is all the variables x 1, x 2, x 3, x 4 up to xn, xn. So, you find out, what is the correlation between these. So, this may 0.04, this may be 0.5, this may be 0.7, 0.002. So, you create this matrices, this matrix and then each of those variables, where you find high correlation.

For example, this seems to be high correlation, this also seems to be reasonably high correlation, and then you try to eliminate these variables. You try to you decide whether, what to use x 1 or x 3, just one of them, not both x 1 and x 3 or you use x 1 and x or x 4, but not both of them. So, in a typical case what to do, is to build up model with each of them separately and then you try to see that the model works or not; however, let me tell you, because why the variation picture of the maths of, multiple regression, you may find out that if you eliminate one of this variables. So, sometimes this variables also gets eliminated and what is left is that some of the variables become important.

So, just removing one collinear variable can also remove the variable, which it was correlated and then the rest of the variables become important. So, this is why you have to take different variables at a time and then do this same analysis. So, one of the ways is just choose one of this variables, the other approach to deal with this is, something called variance inflation factor.

(Refer Slide Time: 22:59)

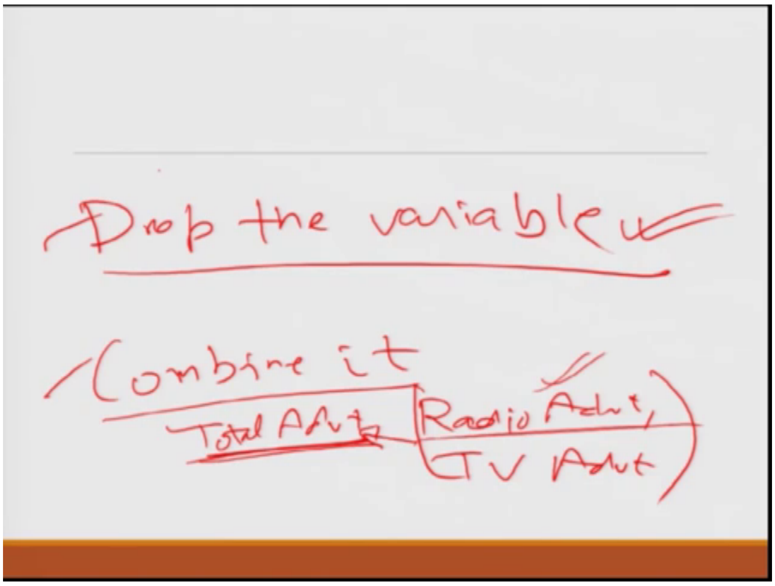


Variance Inflation Factor

(VIF > 5, VIF > 10) →

I will not get in to the maths of it, but if it is a, it is a metric to determine what is the importance of each variable, in itself, ignoring rest of the variables, which are there in the model; So, typically if your variance inflation matrix is greater than 5 then it is a concern for some people, they get liberal and they go up to 10. So, if your VIF is greater than 5 or VIF is greater than 10 depending upon the kind of problem that is a variable of concern and then you try to deal with it, you at least realize. So, this is the first step that you realize that yes, there is collinearity there. Now, how do you deal with it? One simple method, easiest and quickest is, drop the variable.

(Refer Slide Time: 24:05)

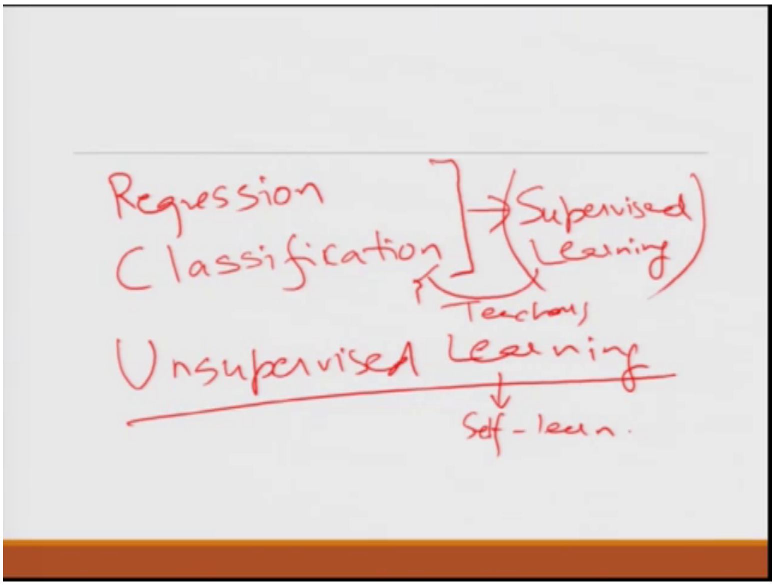


If you ask me, this will work in most of the cases, because you will find that if you remove these variables, because in any case, the two variables together work, because they were correlated, there was collinearity. They are together, were also not able to predict too much of the output. So, removing them does not impact too much of, your prediction.

So, one simple ways to simply drop it, the other way is to combine it; So, if we recall in our case, we had this, total advertisement, we had radio advertisement, we had TV advertisement. So, if you combine these together, these are primarily represented by total advertisement. So, you can choose to go with this one and remove this or you can do a total advertisement minus TV advertisement, which may give radio advertisement assuming that, it may be the case that radio advertisement itself is, responsible, for lot of sales to get.

So, in that case, this may be another variable. So, whether you combine it or you drop it, these are the two of the most common methods, which are used ok. So, we talked about regression, we talked about, the concept of, machine learning. We, we talked about regression being one of the, most important or one of the, may be the, the first areas of problem, first domains, they are machine learning started being used apart from regression, there are; so, we talked about regression. There are two more areas, where machine learning is used.

(Refer Slide Time: 26:05)



Regression

Classification

Supervised Learning

Teachers

Unsupervised Learning

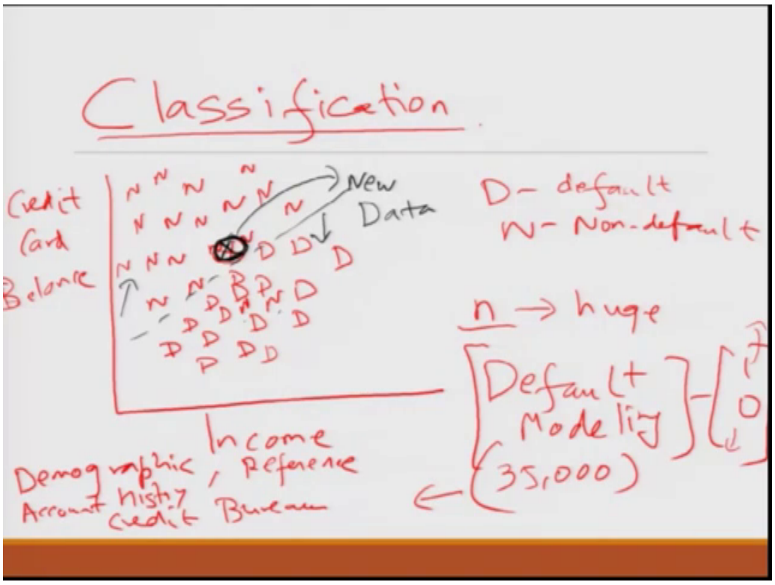
Self-learn.

So, this regression and then there is something called classification. So, regression and classification, they are typically called supervised learning, they are called supervised learning, because, when you build the models, you, you have a data set, where you know the output, because you know the output in a way, you are supervising for each data set. You are saying ok.

This was my desired output, this was my, models output and let me try to tweak the model. So, that might desired output, is as close as to my actual output. So, this is in a way trying to supervise it. It is like, teaching a child, how to speak, and whenever. So, he says A B C D and if he makes a mistake, you supervise it, say no, this is how you pronounce it. This is the, how you pronounce me.

So, this is the most common form or intuitive form of learning that, we do, this is supervised learning and then there is another class of model, which are called unsupervised learning. Now, this is again, supervised learning is what the way we learn from our teachers. This is the way we self learn just by observing, looking into patterns, trying to make sense of the patterns.

(Refer Slide Time: 27:55)



So, let us look at the next set of methods in machine learning. These are classification methods, we will talk off about these in much more detail, but, let us have an intuitive grasp of what it is. So, for example, I plot income versus credit card balance and this is a very common problem in the financial sector, and let me mark by these Ds, the plots of different people would be defaulted and be the people who did not default.

So, let this be a simple plot, am just plotting. So, this, this D means default N means non default. So, this credit card companies are often worried about, which of the fellows will default and not pay monthly due, because they are profitability operates. You know very-very, dangerous zone. I would say between you know trying to, for most of the companies trying to incentivize the customer to delay the payments, because when you delay the payments, you charge a interest on that and that is the source of company, but if somebody is delaying too much, that may be, because it is not even unable to pay.

So, it is a very-very dangerous zone in which they play this game of maximizing the money and this, this is one of the favorite plots that they always try to make and try to make sense out of it that, who are the people, who for different parameters and different, variety of parameters, who would default and not default,

So, I have just taken two parameters out there. In reality there may be 100s of parameters. So, one of my objectives may be to be able to correctly. So, for example, I have one customer, who is somewhere say, who comes out say here, for example,. So, let me just make it with the different color. So, this is a new case, which came, you do not know this is a new data. So, it is may be a new customer, about whom you do not know whether you will default or not default you want to now, decide or make an estimate of whether they will default or non default.

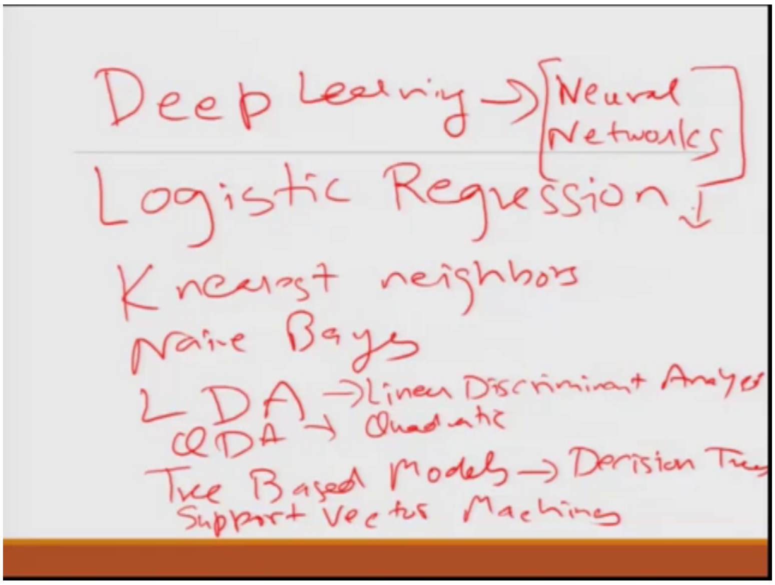
So, how do you go about it. So, intuitively, if you look at it, well it makes sense that roughly, it looks as if you know, this can be clearly partitioned here and people who are here are mostly defaulters people, who are here are non defaulters, but then there are some exceptions out there. And then of course, you know there will be always the errors in the model that, you do not want to be 100 percent correct that. So, in this case, it is simple, what if it was much more jumbled. So, these kind of problems are called classification problems.

Now, depending upon whether you can easily separate the boundary or not. So, in this case, because I had to plot, I used only two variables. In reality, there may be large number of variables, which are. So, there the, the number of variables can be huge, if you talk of, a typical bank, which is doing this kind of, for default modeling. So, this is often called default modeling, which is just trying to classify people in to a 1 0. Bracket 1 means non defaulters, 0 means defaulters or vice versa, depending upon the way you build the model.

So, the number of parameters; for example, for some of the banks are even 35000 parameters. So, these may be parameters, like demographic, then account history, then credit bureau, the reference, the reference with whom he was made a customer, the history of those people. So, and variety for example, even if you look at account history that may be the minimum balance he keeps, the average balance he keeps, the number of year he has kept the number of cheques that bounced the number of times. He has met his commitment, the number of delayed payments.

So, there can be huge number of variables, it can run in to 1000s. Now, depending upon the n that you have depending upon, how cleanly separable data is depending upon lot of things. There are different types of methods, which are used for classification in the banking sector.

(Refer Slide Time: 33:58)



Deep Learning → [Neural Networks]

Logistic Regression ↓

K nearest neighbors

Naive Bays

LDA → Linear Discriminant Analys

QDA → Quadratic

Tree Based Models → Decision Tre

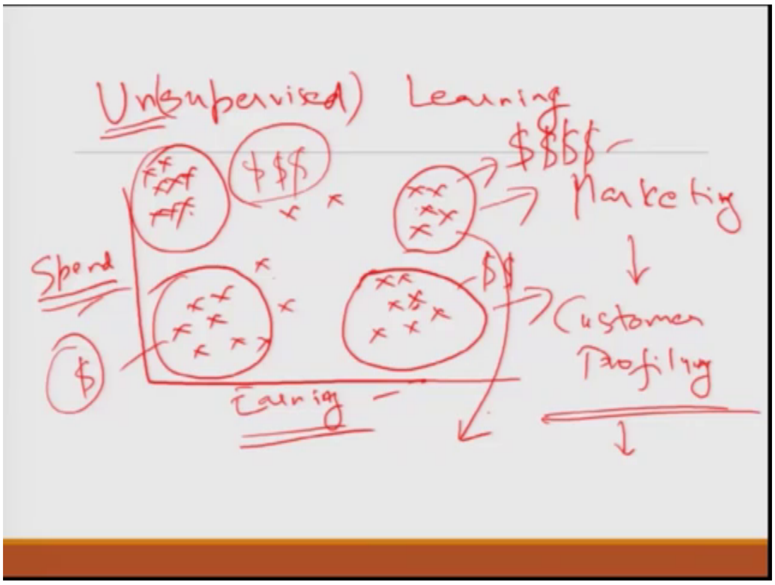
Support Vector Machines

One of popular ones is logistic regression and, we will talk about in details, then there is K nearest neighbors, there is Naïve Bays, which is an, extended form of a more generalized LDA model, which are linear discriminant analysis.

Then you have QDS, which are basically instead of linear use of quadratic, then you have tree based models decision trees then, you have support vector machines and of course, these days very popular, deep learning or nothing, but neural networks.

So, this neural networks is the generic set of tools, not only used for classification, but all forms of, analysis even for regression and even for unsupervised learning. They are used, but, their use in, classification has seen a lot of uptick trends, in recent times.

(Refer Slide Time: 35:57)



So, this is the classification based models and then we have unsupervised learning. Now, unsupervised learning means see. So, for the models that we eh, looked at you know, there was always a target variable. There was always something good or bad weather, somebody is defaulting or not. So, whether his height is high or low or whether his height is 6 feet or 7 feet whether, the sales will be x or x equal to 1 million or 1.5 million. So, these are all what we talked about supervised learning. Now, we had unsupervised, it means there is no, you are not looking at, your, there is no target, there is no goal, but you still want to understand the data.

So, let me give a very intuitive example and often people fail to understand. What unsupervised learning it, but if you look in life, I mean fact unsupervised learning is the most common way that you learn, you take lessons of learning life and you, you you try to create patterns out of what you see and observe nobody tells you, but that is how, we humans are the entire animal can kingdom is designed to act like it. It is all unsupervised learning, the supervised learning is may be a very small part of, what we do, think of, very simple example. Suppose, I have two parameters for example, let me talk of spend versus earning. So, you look at data of, people.

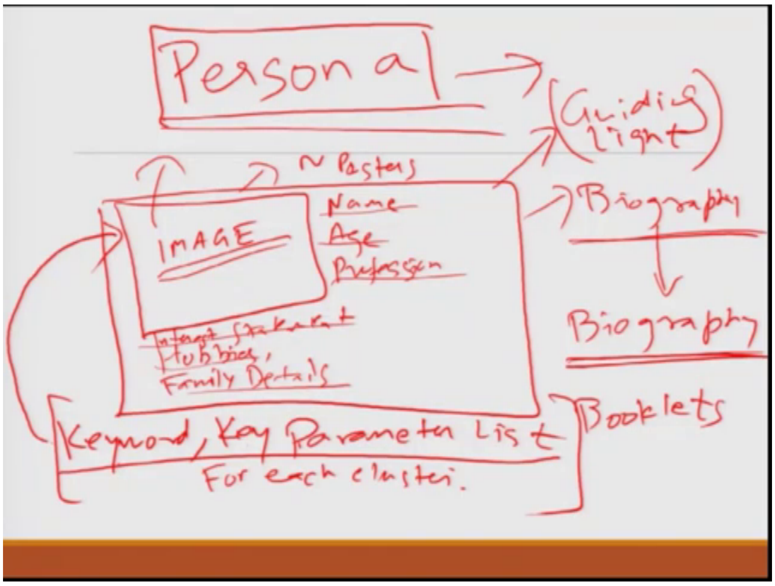
So, just two variable that can be multiple variables, the people, the amount of money that they are spending, every month versus the amount that they are and suppose, you get. Suppose, this is the kind of data that you look; There are people, you know could do, earn a lot, but spend this. There are people who earn little, but also spend little, there are people who earn very little, but then this spend a lot. We not many people is this and there are people, who earn a lot spend a lot. So, just by looking at this data intuitively you know, just on a two dimensional graph we, you can sense out that ok.

This seems to be like one category of people. This seems to be another category of people probably; these are another category of people. So, these are other category of people and then these are may be people in the middle. They are kind of the average, not getting into an recent specific category or we may want to extent this and include this here and then try to make the, this a category at least 4 of the categories are very obvious.

Now, this is the way, we look at data and try to understand, this is what we call for looking at, forest among trees. So, just looking at these patterns and trying to find out what the data looks like breaking, bringing out meaning out of it, this is what is unsupervised learning. This is very-very popular, in marketing.

So, in marketing, what they do is, when they have to do customer profiling and this is very-very powerful tool used there. So, for example, if you go in to; so many of these, marketing forms, many of you, this brand based products you know the, the marketing headquarter of this people, you will find a lot of, holdings of.

(Refer Slide Time: 40:30)

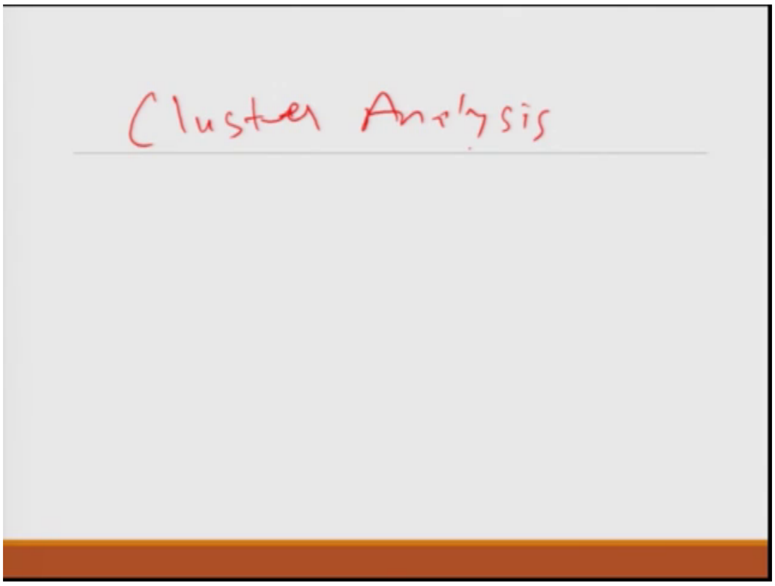


But the posters of, people you know. For example, let me just may be, for some fun sake and some more insides into the, how machine learning etcetera is used in practiced (Refer Time: 40:39) Let me just draw poster. You, you might have seen these posters somewhere here. So, here typically, there is an image, it is the name of the fellow, there is age gender is; obviously, from the image, then there are this profession and there is some interest statement. There is for example; often police office hobbies, family details.

So, in a way, kind of an entire biography of a person; so, who is, what is he like and there can be multiples. So, this is one poster and there, the for same person, there can be N posters and many companies, not only posters are actually have so called biography of a person. So, if you have actually those, booklets, which detail about what kind of book series? What is the kind of food that he loves to eat? What does he do when he is free? What kind of colleagues he have and all sort of details. Now, many a times, if you visit, many shops and many service centers etcetera, you will find something similar out there, you know at least some of the posters out there.

So, the whole idea the reason they make him, this is not something yeah, it may there may be companies, who are doing it based in their gut feeling, but actually, the way they have, they came out of, this was in many cases, this through what is called cluster analysis.

(Refer Slide Time: 43:01)



Cluster Analysis

So, going back to the this slide. So, this is nothing, but a cluster analysis. They looked into multiple parameters, we looked into, just spend and earning they looked into much more parameters and they found out what these clusters look like. Now, depending upon the, from a marketing stand point, depending upon the money dollar, dollar, dollar, dollar. This may be dollar, dollar ok. This is may be this.

This is again dollar, dollar, dollar. So, depending upon the money they give, they find out this cluster. They get into a deep dive, they try to get more information about it as many parameters that supports. So, in this case, we talked about something about collinearity. In this case, let us look at the same thing in a different way, in this case, you do not want to eliminate variables, which are similar or predict the same, whoever not predicting anything you want, all those parameters which are similar.

So, you may, you may do a unsupervised learning with a small set of parameters and then you try to add upon parameters, which are strongly correlated and hence, you create a kind of I would say a keyword and key parameter list for each cluster, try to find out if there is some similarity, in these. So, depending upon these keyword, these key parameters list, for each cluster identify, which are the customer segments. They need to target and to do that this is where, then their marketing team sits, based upon this, creates this image of the fellow they look at.

So, for example, once you have these keywords, key parameters, this cluster, they get to different sources and then try to find out, which are the actual people, who closely meet these kind of parameters. So, they create what is called in marketing lingo persona. So, this is, this is not an actual per person, you get these key parameters list etcetera, then based upon this. For example, you go on line, go into various sources of data and try to find out actual people, who look, behave, feel like the and then based upon all that collective information. You fill in those details.

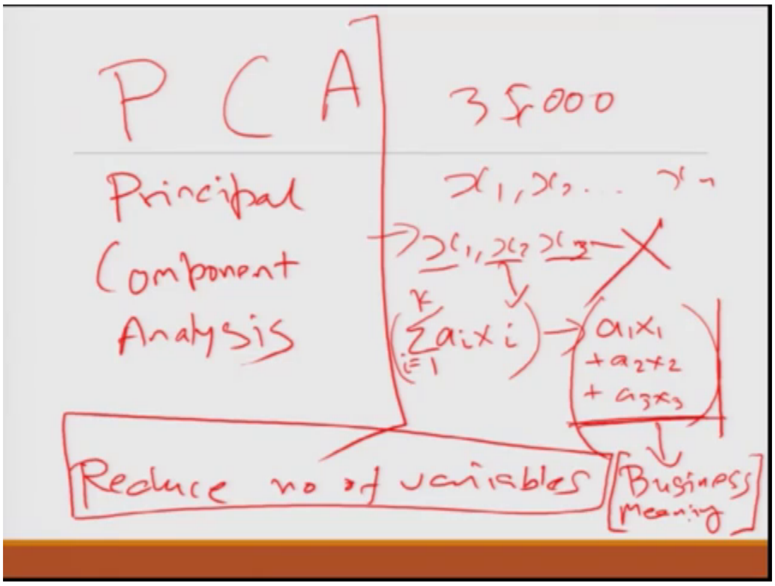
Now, the reason, they do this is that now, this becomes kind of their guiding light, for all their marketing strategies. So, all their marketing strategies now, are based upon, these things. So, often they would divide the team into separate groups, each of them is responsible for one particular persona.

So, often you in many of these very highly brands, sensitive products the, the marketing managers they try to live with it, they try to kind of feel this, persona apart from biography and bio data and etcetera. God knows they may also create, models and they may create 3D images and, holograms and stuff like that they can do, all kind of crazy stuff and like if you, if you recall that, fellow, who, who played that role of joker in that movie, dark night batman movie.

So, he said to have lived as joker for 6 months, closed in one room, just to live like joker, do only that stuff and that, that joker inside him and that is why he is performance is rated to be among the most extraordinary performances in film history. So, it is not fortunate that, that, that, that took him time and it took a great tall on his, life and he died prematurely, but the point am telling is that you know they, they try to live this and the, the, the source of this thing is what is called cluster analysis.

And strong cluster analysis is the entire marketing strategy that flows right from the way. The teams are organized to the way they target, their customers. The kind of, product propositions they give, remind of marketing of the kind of sales and incentives they give, that is often in many state of the art marketing of the organizations is driven through this, persona based marketing analysis. This persona based customer analysis, which is nothing, but clustering or unsupervised learning.

(Refer Slide Time: 48:19)



PCA

Principal Component Analysis

35000

$x_1, x_2 \ldots x_n$

x_1, x_2, x_3

$\left(\sum_{i=1}^{k} a_i x_i\right) \rightarrow a_1x_1 + a_2x_2 + a_3x_3$

Reduce no of variables

Business Meaning

Apart from clustering another unsupervised learning, which you may have done, if you have done course, on marketing research is PCA or nothing, but principal component analysis. Now, principal component analysis again, for those of you who are, comfortable with, matrix algebra. Who have done course on linear algebra, they can get into the maths part of it. Let me just give you a very high level intuition about, what it is, what we are trying to do.

So, what we are trying to do basically, is that, from a practitioners perspective is that we have for example, as I said in many cases 35000 variables, now these. So, many high number of variables are again a night way to deal with it, how do deal with. So, many variables and there, is all sort of collinearities there and more variable, does not mean always more information, often more, often there not it means more noise. So, principal component analysis is try to see, when you can summarize some of these variables and then have one variable represent these.

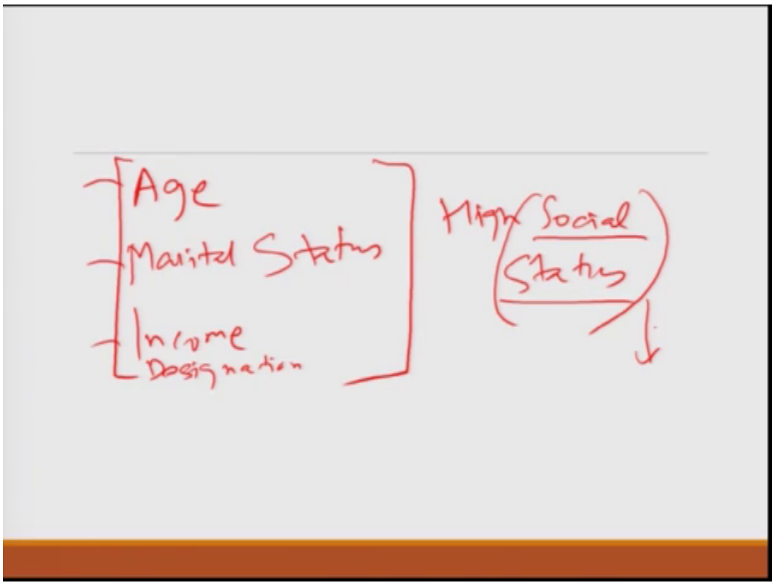
So, in a way what we are trying to do with principal component analysis is that say x 1, x 2, xn are your variables. So, can there be say for example, I choose x 3, 3 of the variables. Is there a way I can make, 1 to k can I make a linear combination of these variables, that linear combination which is. Let me try to; so, which is thing, but a 1 x 1 plus a 2 x 2 plus a 3 x 3. For example, if it is, there are 3 variables. So, can I, may be write this expression and then remove these 3 variables. So, in a way, what am trying to do is to reduce my number of variables. Now, I will lose some accuracy in that.

So, what principal component analysis does is that it gives you the linear combination of variable along with a metric of how important is this. What kind, what is the kind of loss of information that you have, if you combine this variables. So, instead of having x 1, x 2, x 3, if you have all these three combined; obviously.

So, this linear, because life is not linear; so, there is some loss of information that have come. So, we will talk about this, when we discuss it in slight detail, but the concept is that, can this linear combination replace this x 1, x 2, x 3 and if it can replace then let me try to reduce my number of variables. So, the whole idea is to reduce number of variables.

So, I am not doing any regression, etcetera all I am trying to do is, trying to find out, which of the variables can be combined and in a way it is again looking at patterns. So, for example, and, and often by the way this principal component analysis is again; we, we talked about this cluster analysis. We, we talked about this persona principal component analysis again used in persona. So, for example, in case, the three variables that you found useful were like age, then marital status and income.

(Refer Slide Time: 52:15)



So, well you can say that, if somebody is, is a matured age and is married and has high income, he is probably somebody mature or high social status. So, you may that social this, this three variables, kind of represent social status not very accurately, may be if income apart from income, you also have, variable called designation. You find he has a very high designation.

So, you may say that well these 3 variables. I combine them together and I create a new variable called social status. So, this is again very intuitively, easy to understand and feel. So, from a marketing perspective, from decision making perspective, it gives a way to get hold of, let a business sense, let a intuitive sense of, what these things are talking. So, often what they do is, from an operation stand point, they try to create this principal component and based on this principal components, they try to see if I can give some kind of business meaning to this.

So, if you can give a business meaning to it from, practitioner stand point from management stand point, that means, a lot of my marketing, lot of my strategy, lot of my financing, lot of other things I do. In the organization to increase my shareholder value, that can directly flow from there. So, primarily principal component analysis and cluster analysis, they form part of what is called, unsupervised learning. So, we looked into, classification cluster analysis. We have already looked into regression based analysis, which covers primarily, the most of the way machine learning is being, used in industry, practitioners to use it.

In the next class, we will talk about another, emerging field. It is not emerging, but it is becoming very popular, even in a business domain, which is the reinforcement, learning based machine learning tools. We will look about that and then we will also look about some more aspects of doing good machine learning based analysis.

Thank you very much.