Predictive Analytics - Regression and Classification

Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 01
Introduction

Welcome to the Predictive Analytics – Regression and Classification course. I am Sourish. I am an Associate Professor of Chennai Mathematical Institute.

(Refer Slide Time: 00:25)

Reference

Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer
NPTEL
cmi

In this course, I am teaching in this course in CMI for about 8 years now and most of my materials are based on mainly two books by James, Witten, Hastie and Tibshirani's Introduction to Statistical Learning and Hastie, Tibshirani, Friedman's Elements of Statistical Learning.

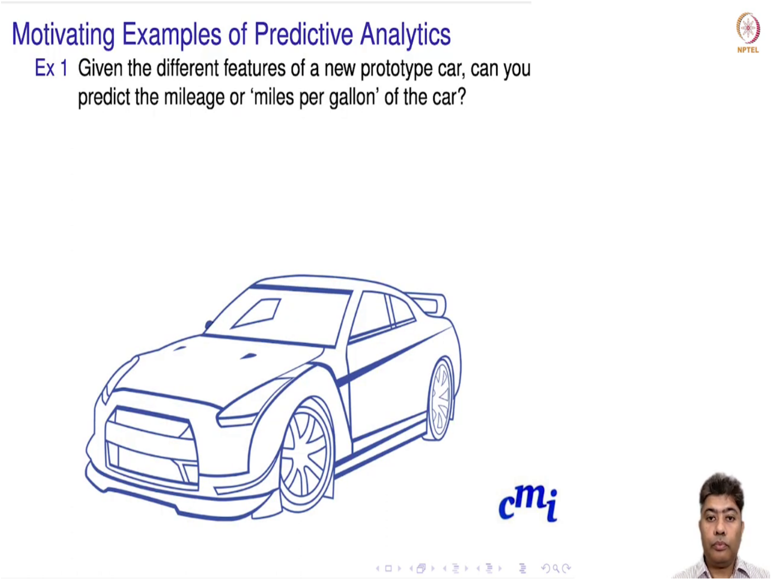
(Refer Slide Time: 00:51)

Reference

Analytical Methods for Social Research
Regression and Other Stories
Andrew Gelman, Jennifer Hill and Aki Vehtari

Now, recently Andrew Gelman and his colleagues Jennifer Hill and Aki Vehtari has written a new book called Regression and Other Stories. I like this new book as well. So, most of my material in this course are going to come from these three books.

(Refer Slide Time: 01:12)



Motivating Examples of Predictive Analytics

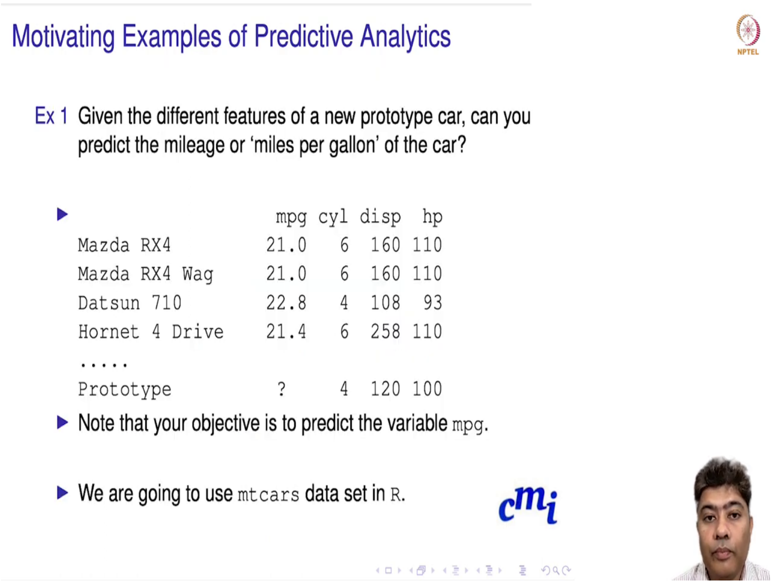
Ex 1 Given the different features of a new prototype car, can you predict the mileage or 'miles per gallon' of the car?

I will motivate you for a predictive analytics problem with this example. Imagine yourself you are working in a automotive car manufacturing company and you are supposed to design a new car. Now, you have due diligently designed this new car and now in this car you have specify every bit of things in this new prototype car, but you do not know what would be the miles per gallon of the car because maybe horsepower or displacement.

And you know other engineering feature you can define using your engineering concepts per miles per gallon is not really a engineering concept in that sense is just a measure of fuel efficiency. So, effectively you want to measure the fuel efficiency, all you have to do you have to really make a car and then you have to run the car for a while and then you can make a measure of fuel efficiency for that car.

But, that is very costly expensive exercise and many car manufacturing companies might not try to do that. So, what you can do instead, we can develop a mathematical predictive model which will tell you based on all the engineering features and based on historical data that what would be the miles per gallon for the prototype car.

(Refer Slide Time: 03:04)



Motivating Examples of Predictive Analytics

Ex 1 Given the different features of a new prototype car, can you predict the mileage or 'miles per gallon' of the car?

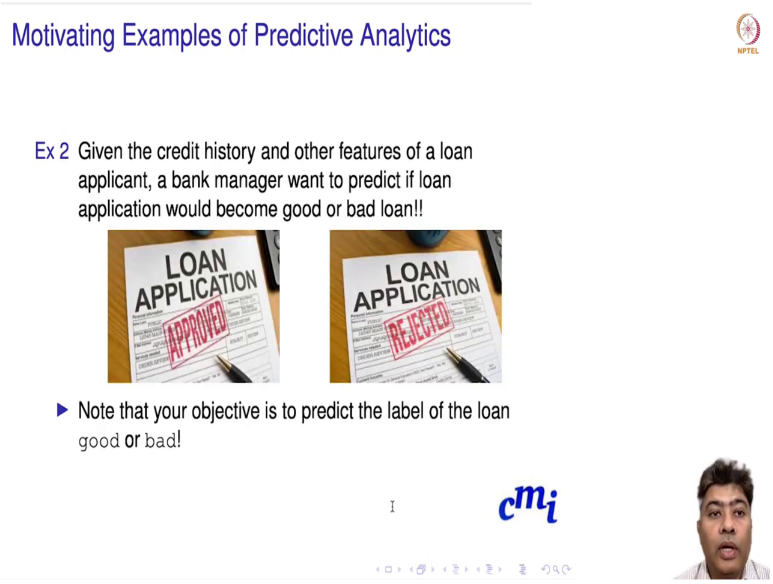
- ▶
```
                  mpg cyl disp  hp
Mazda RX4        21.0   6  160 110
Mazda RX4 Wag    21.0   6  160 110
Datsun 710       22.8   4  108  93
Hornet 4 Drive   21.4   6  258 110
.....
Prototype          ?    4  120 100
```
- ▶ Note that your objective is to predict the variable `mpg`.
- ▶ We are going to use `mtcars` data set in `R`.

Here is an example data for a new prototype car like for this is a one model particular model Mazda RX4 which is which is a 6 cylinder car, displacement is 160, horsepower is 110. There is another car maybe Datsun 710 which is a 4 cylinder car with displacement of at 108 and horsepower is 93 and the miles per gallon is 22. Similarly, Hornet 4 drive is another new car which is a miles per gallon 21.4, 6 cylinder car, displacement of 258 and horsepower of 110.

Now, you have a prototype car new prototype car which you have thought of ok let me design this car with 4 cylinder car with a displacement of 120 and horsepower of 100. Now, what would be the miles per gallon this of this new prototype car? Can you predict this value? Can you predict the miles per gallon? This is a precisely the kind of problem we will be talking about in this entire course.

This particular data set is typically known as mtcars data set it is available in R. Also I will make this available along the course, so that you can do your exercise in Python.

(Refer Slide Time: 04:44)



Motivating Examples of Predictive Analytics

Ex 2 Given the credit history and other features of a loan applicant, a bank manager want to predict if loan application would become good or bad loan!!

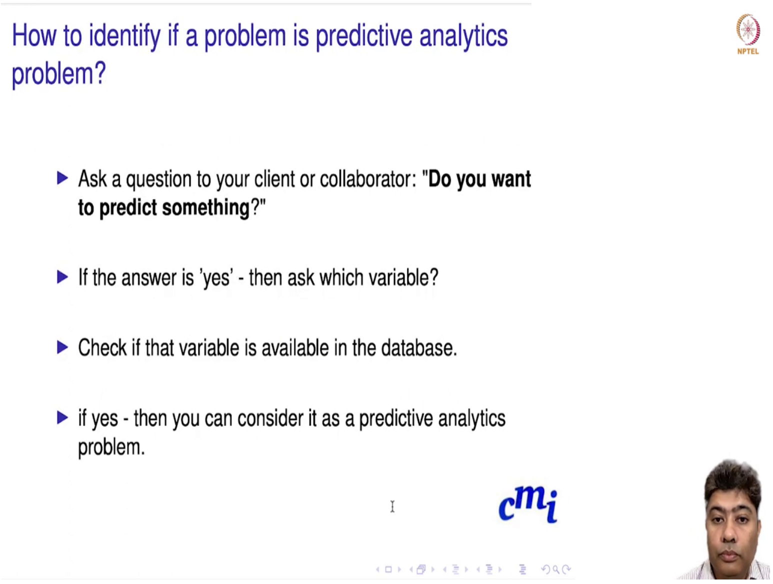
- Note that your objective is to predict the label of the loan good or bad!

Here is another second example on predictive analytics. Imagine yourself a bank manager of a bank now somebody applies for a bank loan. Now, given the credit history or a housing loan or a car loan now given the credit history and other features like you know maybe the educational qualification, whether the person works or not etcetera etcetera given the other these kind of features of the loan applicant, bank manager wants to predict if the loan application would be would become good or a bad loan.

So, our objective is to predict the label of the loan either the potential customer is going to be a good customer or bad customer. If it is a good customer, they will take the loan they will pay the interest and or EMI absolutely on time and they will return the entire capital on time. But, if it is not good customer they may default on their loan and that in that case bank may have to incur a big loss.

So, this is a very important problem for bank manager. So, he she or he wants to predict whether a loan applicant is a good loan applicant or not.

(Refer Slide Time: 06:15)

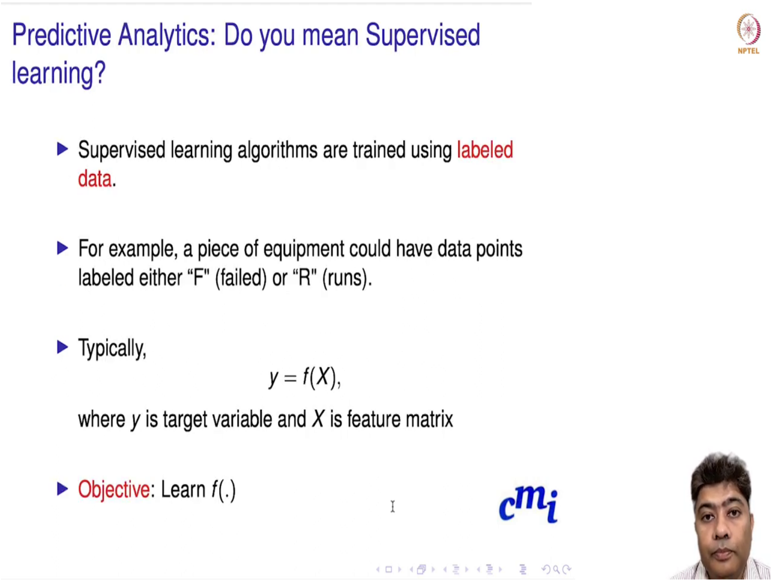


How to identify if a problem is predictive analytics problem?

- Ask a question to your client or collaborator: "**Do you want to predict something**?"
- If the answer is 'yes' - then ask which variable?
- Check if that variable is available in the database.
- if yes - then you can consider it as a predictive analytics problem.

Now, how to identify if a problem a business problem statistical scientific problem, any problem is a predictive analytics problem? Ask question to your client, your collaborator do you want to predict something. If the answer is yes – then ask which variable you want to predict. Check if that variable is available in the database. If yes – then you can consider it as a predictive analytics problem.

(Refer Slide Time: 06:55)



Predictive Analytics: Do you mean Supervised learning?

- Supervised learning algorithms are trained using labeled data.
- For example, a piece of equipment could have data points labeled either “F" (failed) or “R" (runs).
- Typically,

$$y = f(X),$$

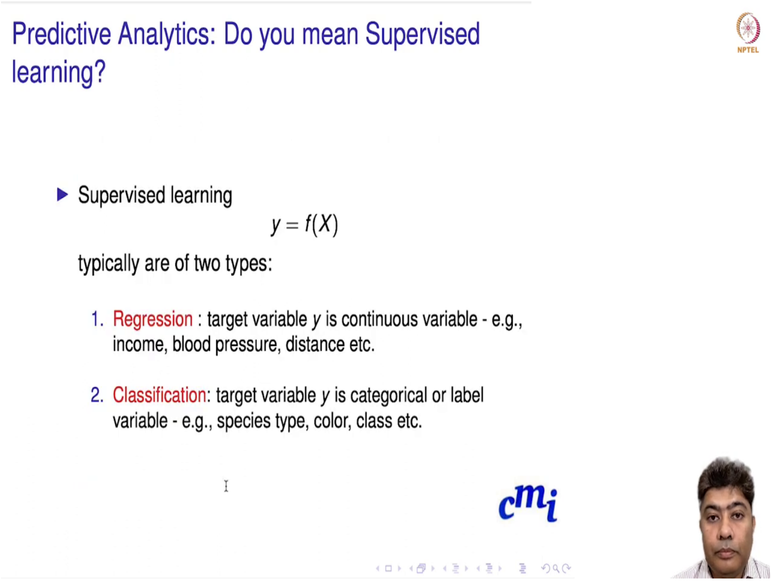
where y is target variable and X is feature matrix

- Objective: Learn $f(.)$

So, when we say predictive analytics you can ask me a question do you mean supervised learning in the machine learning literature. Supervised learning algorithms are trained as using labelled data. So, for example, a piece of equipment could have a data points labelled as failed or it runs. And, based on the features of the equipment you want to predict whether the machine is going to fail or going to run for next one month.

So, here objective is to learn the F that is our objective. So, yes predictive analytics is exactly the supervised learning in typically statistics and in business we call it predictive analytics and in machine learning literature we call it supervised learning.

(Refer Slide Time: 07:55)



Predictive Analytics: Do you mean Supervised learning?

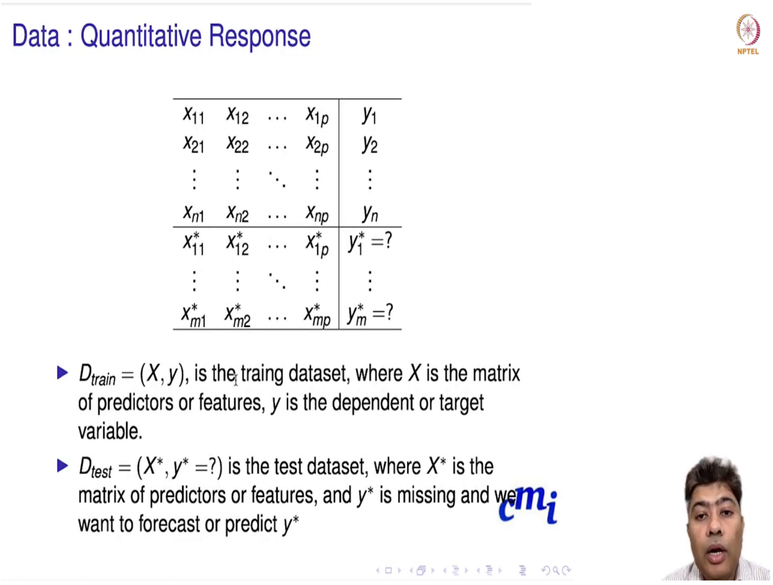
- Supervised learning
$$y = f(X)$$
typically are of two types:
  1. Regression : target variable y is continuous variable - e.g., income, blood pressure, distance etc.
  2. Classification: target variable y is categorical or label variable - e.g., species type, color, class etc.

In supervised learning, typically there are two types of problems: the first problem is called regression and the second problem is called classification. In the regression target variable y is continuous variable. For example, you want to predict somebody's income or somebody's blood pressure or you just want to predict the distance from one place to another place.

Classification on the other hand is target variable when it is categorical or a label variable. Suppose based on the feature of a animal you want to predict the species type or the subspecies type of the animal or the plant, based on the you want to predict the colour of the car that will be most by most that will be most bought in the market or you want to predict class.

So, this is whenever it is a like categorical variable, this is a classification problem. Your target variable is categorical variable it is a classification problem, but if your target variable is a continuous variable, then it is a regression problem.

(Refer Slide Time: 09:27)



Data : Quantitative Response

x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n
x^*_{11}	x^*_{12}	$\cdots$	x^*_{1p}	$y^*_1 = ?$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x^*_{m1}	x^*_{m2}	$\cdots$	x^*_{mp}	$y^*_m = ?$

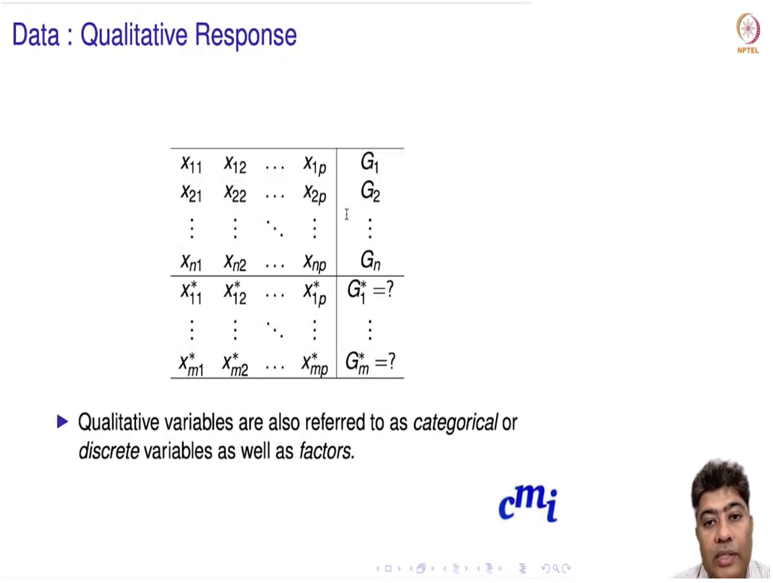
- $D_{train} = (X, y)$, is the traing dataset, where X is the matrix of predictors or features, y is the dependent or target variable.
- $D_{test} = (X^*, y^* = ?)$ is the test dataset, where X^* is the matrix of predictors or features, and y^* is missing and we want to forecast or predict y^*

Typically, what we do? Our data set will have bunch of features like here in this case I am assuming we have p features each column represent a particular feature and each row particular a sample represents a particular sample or instances. So, in machine learning literature it call each row will be called instances and typically in statistics literature it will be called samples.

Now, in a typical predictive analytics or the supervised learning setup, each features these are called typically X will be the features and y will be your response or target variable and we are here considering n many samples or n many instances. And, this is this part of the data typically called the train data, and then there will be a test data where you may have m many samples or instances and where you do not know the value of the responses.

So, based on the train data we will fit the model, we will train the model and then based on the test data we will try to predict what would be the value of y for these new test data point or test points. So, this case in this case what we found, it is the typically the here it is a quantitative response. So, it is a it will be a regression kind of problem.

(Refer Slide Time: 11:14)



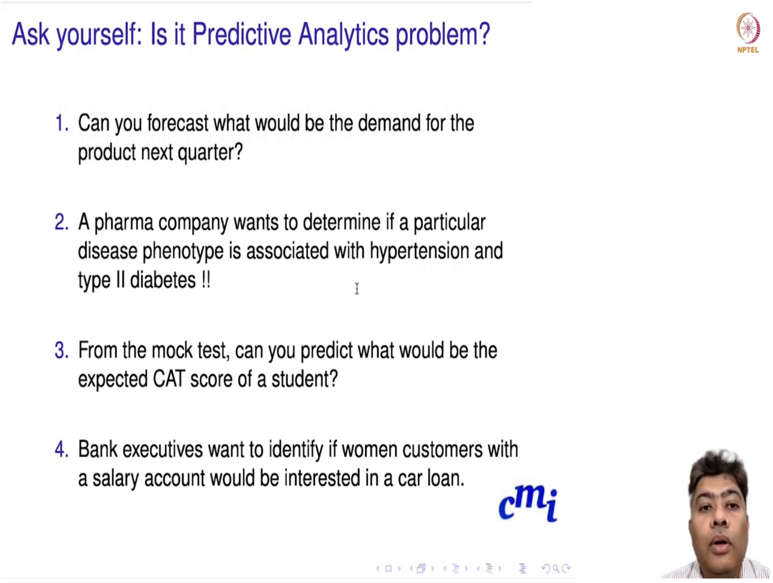
Data : Qualitative Response

$$\begin{array}{cccc|c}
\hline
x_{11} & x_{12} & \cdots & x_{1p} & G_1 \\
x_{21} & x_{22} & \cdots & x_{2p} & G_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np} & G_n \\
\hline
x^*_{11} & x^*_{12} & \cdots & x^*_{1p} & G^*_1 = ? \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
x^*_{m1} & x^*_{m2} & \cdots & x^*_{mp} & G^*_m = ? \\
\hline
\end{array}$$

- Qualitative variables are also referred to as *categorical* or *discrete* variables as well as *factors*.

And, in this case when it will be qualitative response it will be classification problem, but we will solve it exactly in the same way, we will have bunch of features and many samples and we each sample will be have some categorical response. And, we will build a model, based on the model for the response for the test cases we will predict what would be the categorical response.

(Refer Slide Time: 11:44)



Ask yourself: Is it Predictive Analytics problem?

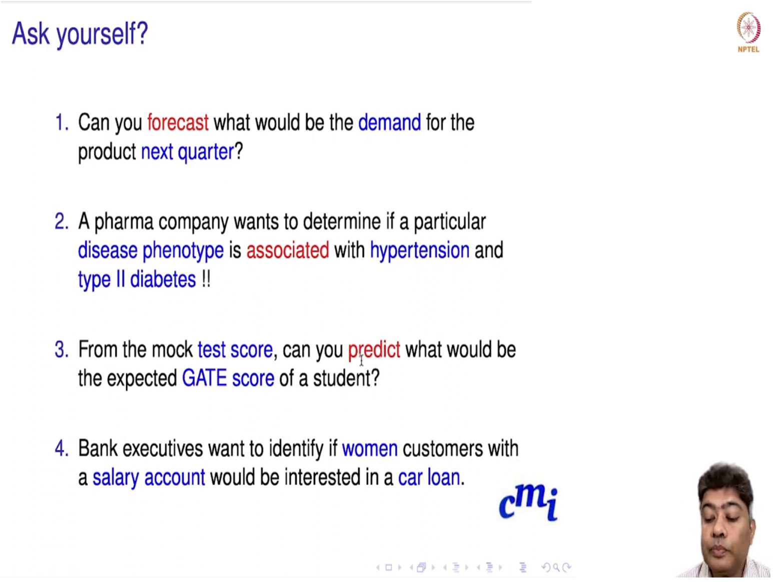
1. Can you forecast what would be the demand for the product next quarter?
2. A pharma company wants to determine if a particular disease phenotype is associated with hypertension and type II diabetes !!
3. From the mock test, can you predict what would be the expected CAT score of a student?
4. Bank executives want to identify if women customers with a salary account would be interested in a car loan.

Now, I am going to set up ask yourself question session. Pause the video, read these questions carefully and think for a while and come up with the answer if it is a predictive analytics problem or not. So, the first question is can you forecast what would be the demand for the product next quarter?

Second question, a pharma company wants to determine if particular disease phenotype is associated with hypertension and type II diabetes. Third question, from mock test can you predict what would be expected CAT or CAT score for a student? And, the fourth question bank executives want to identify if a woman customer with salary account would be interested in car loan.

So, pause your video for a while, think about it and write down your answer and then start the video. I am going to give you the answer in the next slide.

(Refer Slide Time: 13:09)



Ask yourself?

1. Can you forecast what would be the demand for the product next quarter?
2. A pharma company wants to determine if a particular disease phenotype is associated with hypertension and type II diabetes !!
3. From the mock test score, can you predict what would be the expected GATE score of a student?
4. Bank executives want to identify if women customers with a salary account would be interested in a car loan.

Can you forecast what would be the demand for product next quarter? Now, you see the forecast I marked it as a red colour. So, it is going to this is a keyword that will give you the that is giving you the answer that yes, it is going to be a predictive analytics problem because we want to predict, what we want to predict? Demand in the next quarter.

Now, the second question is a little bit tricky. A pharma company wants to determine if a particular disease phenotype is associated with hypertension and type II diabetes. So, it is not predictive analytics problem; however, maybe we will be interested using regression type model or a classification type model to figure out if indeed hypertension and type II diabetes is associated with disease phenotype.

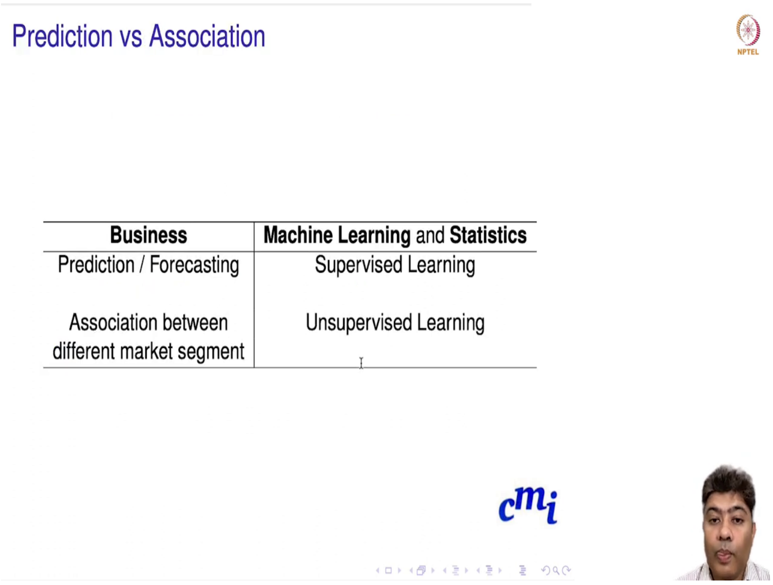
Now, this particular problem can be approached by two methods either you just see what are the association and correlation between these variables disease phenotype hypertension and type II diabetes or you can set up a regression model also and from the regression model try to figure out if there is in any association between these variables.

So, we are not interested in this problem, we are not interested in predicting anything, but still the same predictive model we will see future that same predictive models can be used to understanding association between the variables. From the mock third question from the mock test score can you predict would be what would be the expected GATE score of a student?

Now, here clearly you want to predict something. So, it is indeed predictive analytics problem. And, the final fourth and final question bank executives want to identify if woman customer with bank salary account would be interested in car loan. Here also you will see that association is something that is what we are. Interested in similarly in banking and you know incorporate it will be considered as a market segmentation problem.

You can use simple market segmentation technique or the data based segmentation technique to identify if they are interested or not; you can use regression or classification type model also update to use this. So, directly no it is not predictive analytics kind of problems, but what you were interested in little bit more on the association finding the association between the woman customer if they have a salary account and would be interested in the car loan or not.

(Refer Slide Time: 16:39)

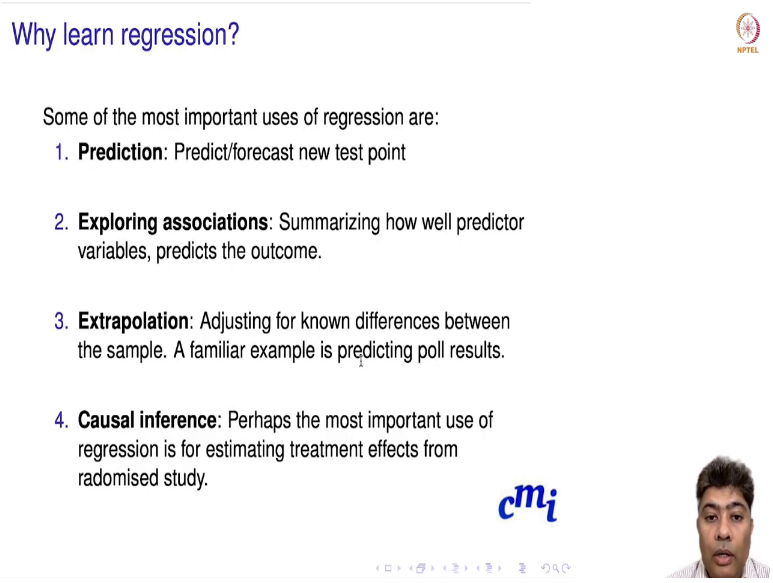


Prediction vs Association

Business	Machine Learning and Statistics
Prediction / Forecasting	Supervised Learning
Association between different market segment	Unsupervised Learning

Now, natural question that comes to our mind now, so, what is the difference between prediction and association. So, in business what we call prediction and forecasting or business in statistics in machine learning and statistics we call it supervised learning. And, in business typically often time we are interested in association between different market segmentation. This comes under what is what we call unsupervised learning or machine learning or in machine learning and statistics jargon.

(Refer Slide Time: 17:17)



Why learn regression?

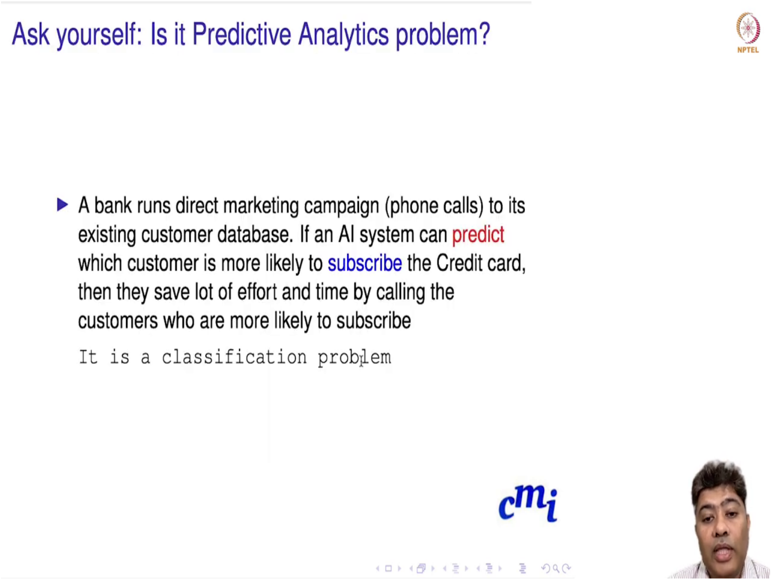
Some of the most important uses of regression are:

1. **Prediction**: Predict/forecast new test point
2. **Exploring associations**: Summarizing how well predictor variables, predicts the outcome.
3. **Extrapolation**: Adjusting for known differences between the sample. A familiar example is predicting poll results.
4. **Causal inference**: Perhaps the most important use of regression is for estimating treatment effects from radomised study.

Now, you must ask a question to yourself why we should learn regression. Some of the most important use of regressions are I would say the most four important uses of regressions – prediction or forecast a new test point; second is exploring the association or typically call we call it inference statistical inference in statistics summarizing how well a predictor variable predicts the outcome; exploration adjust for known differences between the samples a familiar example is predicting poll results.

And, the finally, the causal inference perhaps the most important use of regression is for estimating treatment effects from randomized study. If you have a randomized clinical control study from the randomized clinical control study, if you can make causal inference and in that case a regression classification model play a huge role.

(Refer Slide Time: 18:36)



Ask yourself: Is it Predictive Analytics problem?

- A bank runs direct marketing campaign (phone calls) to its existing customer database. If an AI system can predict which customer is more likely to subscribe the Credit card, then they save lot of effort and time by calling the customers who are more likely to subscribe

It is a classification problem

Now, one more ask yourself moment. So, is it predictive analytics problem? A bank runs direct marketing campaign phone call to its existing customer database. If an AI system can predict which customer is more likely to subscribe the Credit card, then they save lot of effort and time by calling the customer who are more likely to subscribe.

Now, please pause your video and then think about it for a couple of minutes and identify if it is indeed a predictive analytics problem. Yes, indeed it is a predictive analytics problem. It is actually classification problem and with that we are going to stop this video and we will move on to the part B of this lecture.

So, thank you and please carry on to the part B of the lecture. You may take a small break here and come back and let us continue on part B.