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:06
Lecture:22

INTRODUCTION TO CLASSIFICATION | Business Intelligence & Analytics

Hello and welcome back. Today we are going to start a discussion on one of the dominant and very important types of problems and in that way, types of techniques that are available to address that problem that is known as classification. Classification as an area of application of business intelligence and analytics techniques. So in real life in different domains, there is always a need to classify, to group objects and when we classify or when we hear these terms grouping, classification, categorization they all sound very similar. Yes they are very similar but there are subtle differences among these concepts. We will be intently focusing on the concept of classification in this session and we will learn certain specific techniques for classification subsequently.

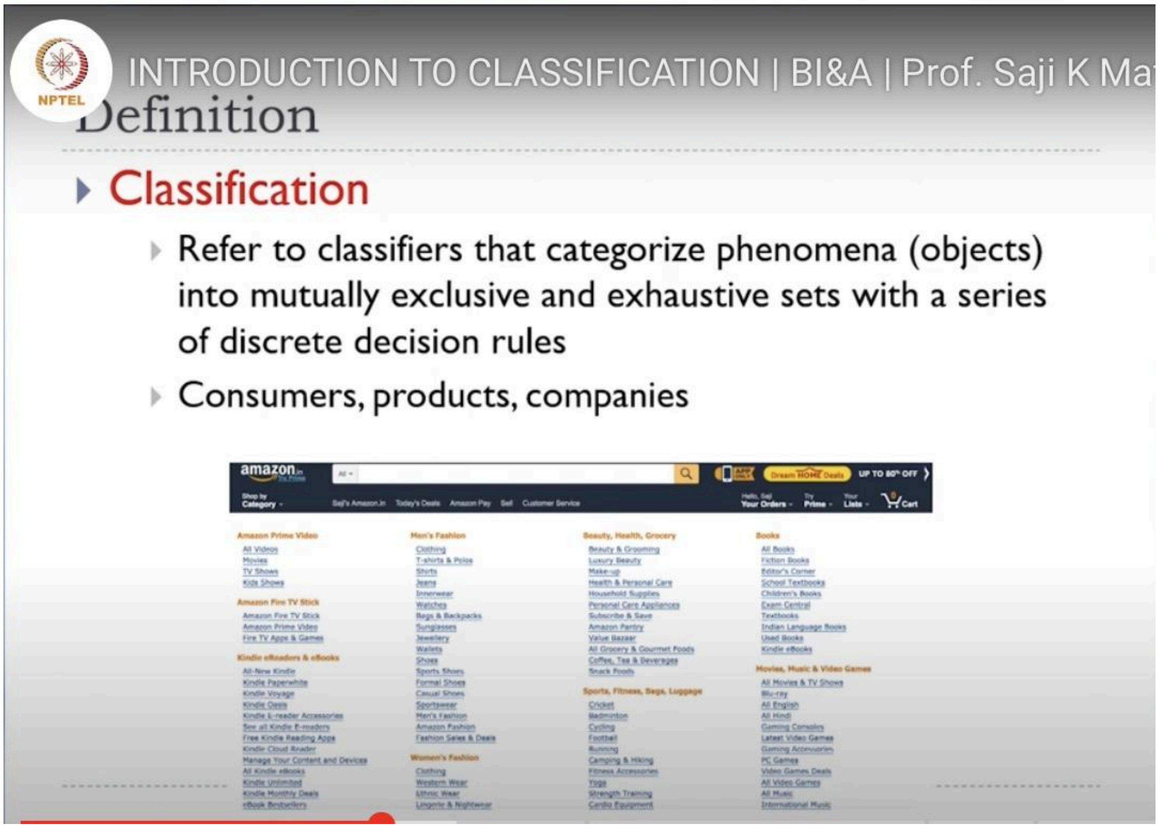
A class, when students come to a room and attend to lectures and complete a course, we call each session as a class. It was a class because they are all entering a classroom where there is a class. So they belong to a class and what is that class? They are studying one common subject. So that subject has certain common attributes or certain common characteristics.

They belong to a class. So how do these people or objects get into the class or get classified, is the objective of our enquiry as far as this particular topic is concerned. So classification, if I give a very simple definition of classification, if you refer to textbooks you will find that classification is that task or that activity which assigns objects to certain predefined classes, certain predefined classes. So therefore in classification the understanding is that there are defined class labels or class categories or whatever you want to call. For example if I make simplest possible, you know, classes in a classroom I would ask students the male students and the female students.

Of course those are not the only two class labels possible but for simplicity, only for the sake of simplicity let us assume that there are two types of students in the class. So based on something known as gender. So the variable there or the basis there or the base variable there is gender and you also notice that gender has two values. It is a categorical variable. Of course it is a special case.

In my case I am treating it as a binary variable. So it has two class labels, a variable with two values or two labels, which is predefined. So I am asking the class or instead of asking, I am sort of grouping each object into one or the other and that is known as classification. Assigning objects to predefined classes is classification. Assigning objects to predefined.

So therefore please notice that the class labels are predefined or the labels, it is a, in you know in data mining this kind of problems are called or this kind of data are called labeled data sets. They have a predefined label available. And therefore you know, you can also call it a sort of supervised learning technique where there is some target variable which has class labels predefined and that is the objective. You can group every object into those predefined classes or put them give them some class label, where do you belong to. That is what essentially classification is.



INTRODUCTION TO CLASSIFICATION | BI&A | Prof. Saji K Ma

Definition

- **Classification**
  - Refer to classifiers that categorize phenomena (objects) into mutually exclusive and exhaustive sets with a series of discrete decision rules
  - Consumers, products, companies

I should say that the foundations of classification at a philosophical level is in the concept of essentialism which was proposed by Greek philosophers like Plato. Implying that for an object to be an object or for an object to be defined as an object, it should have certain essential characteristics. It is those characteristics that makes that object an object. So objects are defined by characteristics. So it is again in classification, you will see that the basis for classification is the characteristics of objects or the attributes of objects.

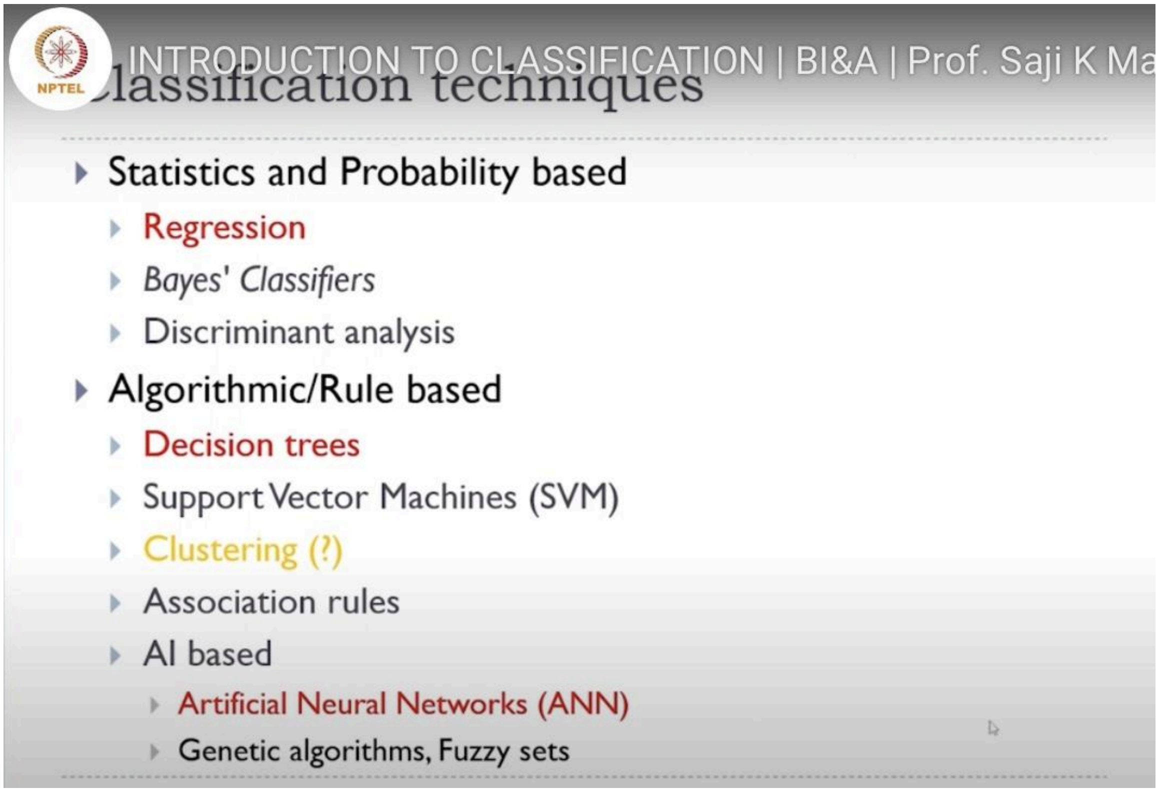
So we draw upon some of those foundations and that is how this concept is going to be presented in the session. Now classification has several applications. You often times in social life, there is a need to classify, you know as I said you knowy based on gender or based on region, based on economic status or educational status, you always you know think of an HR firm recruiting people, you know they may classify people based on their expertise or their educational qualification or their experience etc. So therefore, so and look at you know problems where risk profiling of people have to be done. For example, in credit business, the credit worthiness of an individual is based on certain profile of the individual or the risk, if you associate that word risk.

Low risk versus medium risk versus high risk. Well, risk is the target variable and the target variable in my definition has three values or three levels low, medium, high. So every prospect or every customer falls into one of the three classes based on certain characteristics of the object. So that is a practical problem and there are you know, you can also imagine in biology, you know if the species are classified based on their characteristics and you know that you know, you also call it typology. So, practical applications of classification do exist in different domains- business and non-business.

In this class of course, we will be focusing on business problems where classification is very important. Look at product classification or customer classification, classification of you know, firms within an industry etc. etc. etc. All these are problems where the issue of classification becomes important.

And if I have to classify a small group into two classes or three classes you know, you do not need a session like this because that can be manually done. If the features are visually identifiable with certain accuracy then you yourself, your mind is the classifier. You can actually put objects into different buckets or group people into different classes. That is done by visual observation of the characteristics of the objects and then the mind actually does the assignment. So you are the classifier in that case.

But when you are working with data sets which are, which runs to millions in terms of size then manual classification is virtually impossible and therefore you need algorithms or you need a more systematic approach. So that is what we are going to see. So look at product classification in e-commerce. So products are classified into different classes and how is that done. So it is done using algorithms and there are different types of algorithms for this purpose.



INTRODUCTION TO CLASSIFICATION | BI&A | Prof. Saji K Ma

lassification techniques

- Statistics and Probability based
  - Regression
  - *Bayes' Classifiers*
  - Discriminant analysis
- Algorithmic/Rule based
  - Decision trees
  - Support Vector Machines (SVM)
  - Clustering (?)
  - Association rules
  - AI based
    - Artificial Neural Networks (ANN)
    - Genetic algorithms, Fuzzy sets

I will give you a general overview of different types of classification algorithms and in the subsequent sessions, we will focus on one of them in particular. So let me give an overview of classification techniques that are available in literature and also available in practice, in terms of algorithms available in the form of software or in the form of functions in libraries in the programming context. So classification algorithms can be classified into two. One is the statistics and probability based algorithms. Regression is an example, logit particularly is very useful technique for classification.

Why? We have seen this already in one of the early sessions where we discussed bizocity score, scoring every object. Every object gets a probability value between 0 and 1. And hence once every object has a score, then using that score as the basis you can build classes. Suppose you want to classify your objects into every decile. So you can define 10 deciles.

So 0 to 0.1, 0.1 to 0.2, 0.2 to 0.3 and so on till it is 0.9 to 1. So you define deciles and each object falls into one of those deciles.

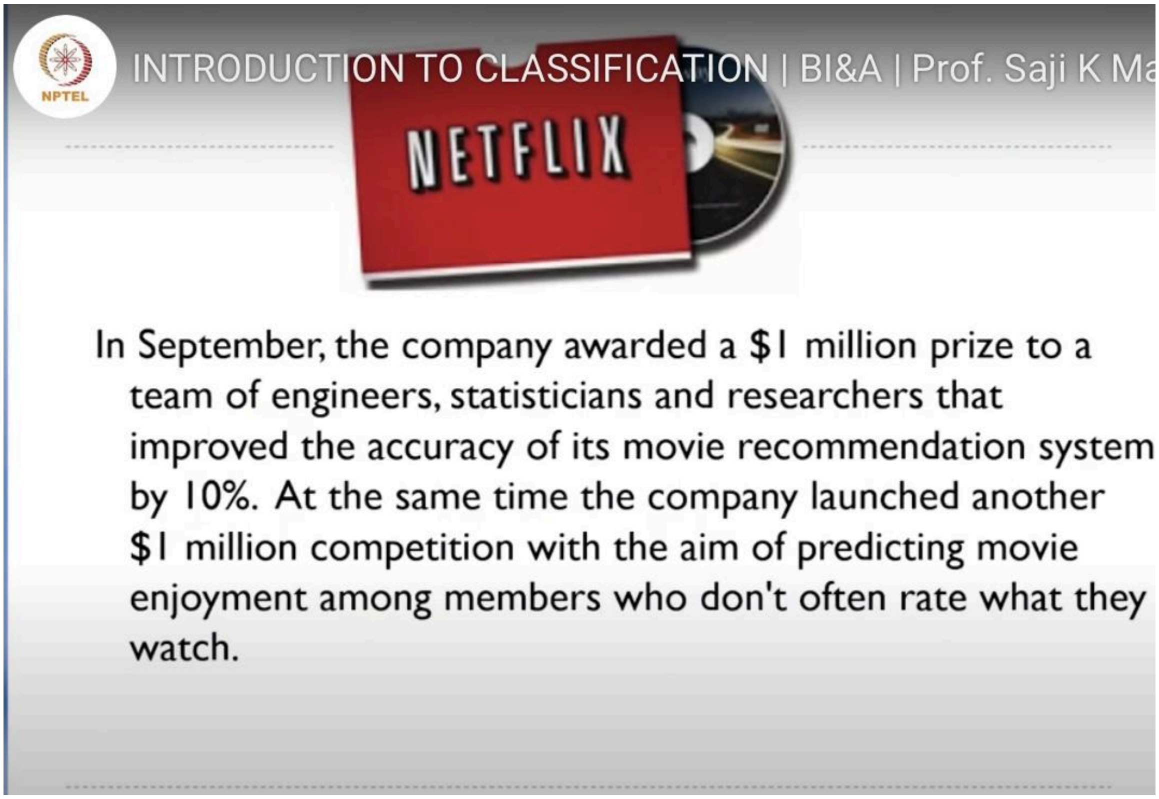
So classification typically is mutually exclusive and collectively exhaustive. You have to classify all objects or all objects have a class and all objects have a singular class, one class. They do not belong to multiple classes. Although in some techniques like Fuzzy classification, the membership of an object is by degree. So that is a different method but we are discussing the statistical techniques and also algorithmic and rule based techniques.

So regression is one method. Then based on Bayes theorem, there are Bayesian classifiers, Naive Bayes classifier is a good example which is very simple to understand but widely used. Discriminant analysis is also used and it is a statistical technique for classification. And I have put regression in red because this is something that we have discussed in some detail and we will also be discussing a problem based on regression, although not specifically for classification but we have already seen the principles of using regression, logistic regression as a classification technique in connection with the case of bizocity scoring.

So that is one thing. The other class of classification techniques is the algorithmic or rule based techniques. Decision trees is a good example and we will be discussing decision trees in a subsequent class and then, there are other techniques like support vector machines, clustering. I have put a question mark there, I will tell you why, association rules, then there are artificial intelligence based techniques like the artificial neural networks, genetic algorithms, fuzzy sets etc. So you can see a range of techniques that are available, developed by the CS community, computer science community which are algorithmic in nature and as far as this course goes we will be discussing three of them from the algorithmic category, decision trees, clustering and neural networks and, but the disclaimers are neural networks we are not discussing, we are discussing but we are not discussing for classification but you can extend the application of neural networks to classification as well, based on the lecture. And clustering is it a classification technique? Strictly speaking no, but it is a bit grey area. For example, in clustering, the clusters are formed based on the proximity of the objects with respect to attributes.

It is not based on any predefined class labels. There is no predefined classes in the case of clustering and therefore it does not follow, well I pick an object and based on its characteristics, I assign it to one class or the other class. No, that is not the way clustering works but in clustering algorithms will group every object based on their attributes, not based on a target variable and its class labels. So that way clustering is different but once you build clusters from a data set based on the attributes of the objects and say cluster A, cluster B, cluster C. Now each cluster has certain characteristics.

Each cluster is characterized or profiled based on the attribute values. Certain attributes are having high value in certain cluster and they have medium or low value in another cluster. So clusters are discriminated. Now having built clusters, it is possible to assign new objects to those clusters. Then it becomes like a classification problem.



INTRODUCTION TO CLASSIFICATION | BI&A | Prof. Saji K Ma

In September, the company awarded a $1 million prize to a team of engineers, statisticians and researchers that improved the accuracy of its movie recommendation system by 10%. At the same time the company launched another $1 million competition with the aim of predicting movie enjoyment among members who don't often rate what they watch.

So therefore, purely clustering as such is not classification but once labels are identified using clustering technique, you can use it for classification. Let us come back and go more specifically into classification. So in order to motivate classification as a business problem, I have here given the example of the announcement of a company Netflix. I am sure many of you are familiar with Netflix and Netflix oftentimes announces problems, particularly algorithmic problems as a sort of competition.

So look at this case. The company awarded 1 million dollars as price to a team of engineers, statisticians and researchers that improved the accuracy of its movie recommender system by 10 percent. At the same time, the company launched another 1 million competition and 1 million dollar competition with the aim of predicting movie enjoyment among members who do not often rate what they watch. What is to be noticed here is that there is a huge reward, of the order of million dollar. For what? For improving the performance of an algorithm by certain percent. And of course, you have to keep in mind that this is business.

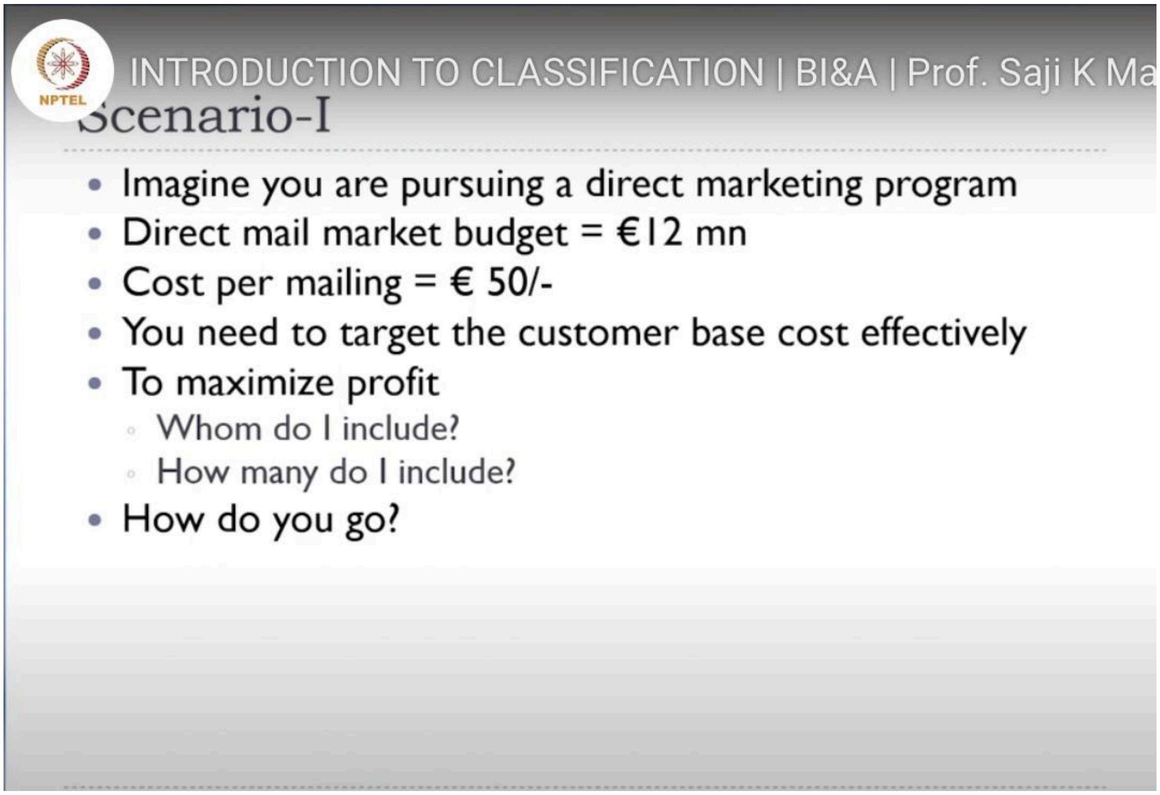
In a business context when you spent money, you expect returns from the money. There should be return on every penny that is invested. So we can reasonably imagine that the company is expecting much more returns than what they invest, from that 10 percent improvement in the performance of an algorithm. It gives us certain indication. And let us see in the subsequent discussion, how this could be true.

So I am presenting a couple of scenarios to you, taken from textbook on marketing analytics. So imagine you are pursuing a direct marketing program. What do you mean by direct marketing program? In direct marketing, you have a product or a service. And you reach out to the end customer or the end user directly as the producer of the service or the product. You sell the product directly to the customers.

There is no distributor, there is no retailer. It is a direct connect with the customer and that is known as direct marketing. For example, well some of you are very creative and you start a lifestyle magazine. And you want to reach out to potential customers. And what will you do? You need to get a list of potential customers or prospects.

And then you need to identify whom you should send your promotional materials to. And of course, you need to get a build a customer base. So, here is a small problem that is given to you. Number one, the direct mail market budget is euros 12 million.

I have converted the currency to euros here. So 12 million is the marketing budget for direct mail marketing. And that is fixed. For everything is having a cost and a budget in business. You understand that.



INTRODUCTION TO CLASSIFICATION | BI&A | Prof. Saji K Ma

Scenario-I

- Imagine you are pursuing a direct marketing program
- Direct mail market budget = €12 mn
- Cost per mailing = € 50/-
- You need to target the customer base cost effectively
- To maximize profit
  - Whom do I include?
  - How many do I include?
- How do you go?

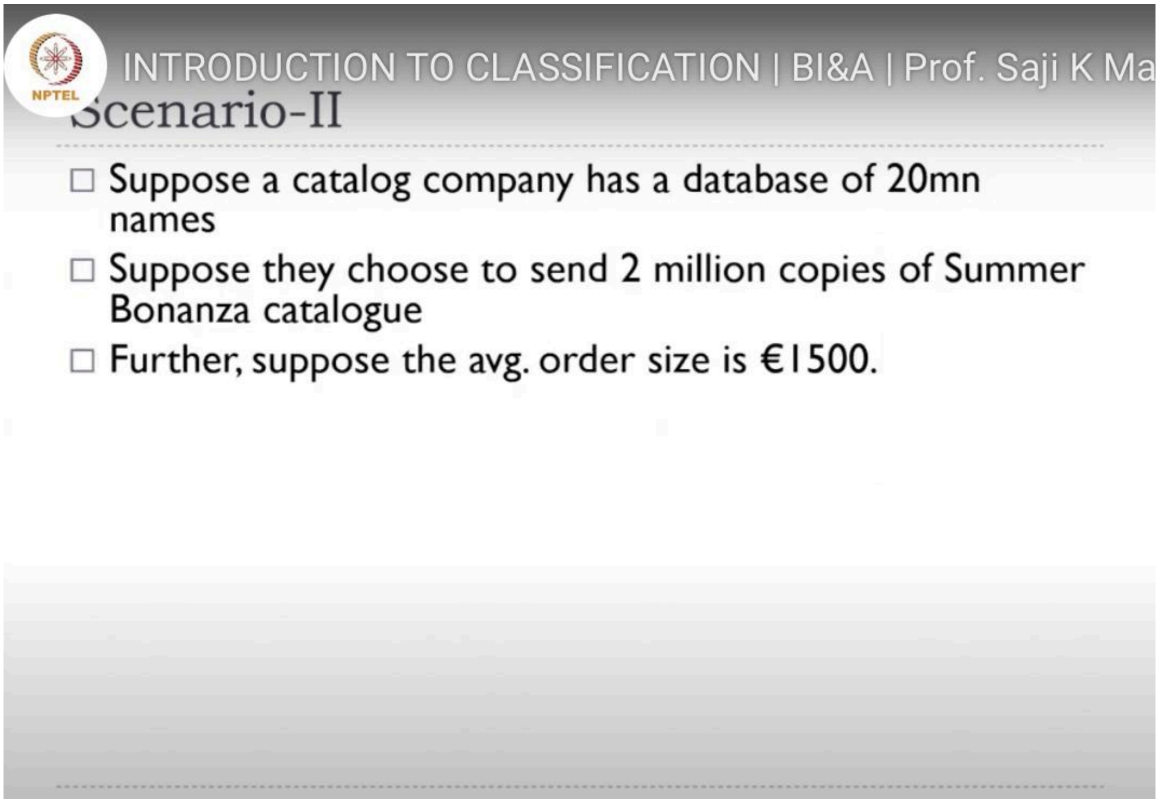
So 12 million is maximum that one could spend. And the cost per mailing is 50 euros. What is cost per mailing? Suppose in the example I provided, say lifestyle magazine. So if you mail it to one prospect, the cost of that mail packet reaching the customer, reaching the prospect is 50 euros, which include all the costs involved, including the design costs, the development costs, the administration cost, the packaging costs, the mailing costs, there are a lot of activity based costs. Let me stress that again. This kind of analytics requires activity based costing, or you should know the cost of activities, only then you can actually apply this kind of algorithms.

So cost per mailing is calculated as 50 euros. And now here your objective is to target the customer base cost effectively. That is one and cost effectiveness apart, the objective of the marketing manager is to maximize profits. If the manager spends 12 million, that 12 million should actually return and the profit should be as high as possible or one tries to maximize profits from the investment. So the question here is, whom do I include? The business wants to maximize the profits.

So suppose you have a list of prospects, which is a large number, which you can buy from outside as a database. So whom do you include and how many do you include? And how do you go from here? You have a budget, you have a cost per mailing. Now look at the second question. How many do I include, is a simple question.

You know the budget, you know, cost per mailing. So it is 12 million divided by 50. It is solved by a simple division. This problem is solved. But the more interesting and difficult problem is whom do I include? For example, you buy 10 million customer contact details, but you cannot reach out to 10 million.

You do not have the budget. Out of this, you have to choose a subset, which is this number. And what is that subset that you should choose, so that you can maximize profit, is the problem here. The problem here in other terms, in business terms is a targeting problem. Whom do you target? How do you identify the target group from a large database? That is a question in front of us.



Scenario-II

- □ Suppose a catalog company has a database of 20mn names
- □ Suppose they choose to send 2 million copies of Summer Bonanza catalogue
- □ Further, suppose the avg. order size is €1500.

Let us go. Okay, let me give another scenario. This is complimentary. Suppose a catalog company has a database of 20 million names. And so that is the general database that they have purchased. And suppose they choose to send 2 million copies of summer bonanza catalog. And this is of course, you can understand that this is based on their budget.

They have calculated only 2 million, is what they can afford to reach out. And further, suppose the average order size is euros 1500. What does that mean? If you reach out to 2 million prospects, and some of them get converted. So based on the historical data, the average order size of those who are converted is 1500 euros. That is the so called expected value of order size, in other words.

And given the scenario, let us ask a question. Suppose somehow, that somehow is in italics, that somehow is where the science comes in or the analytics comes in. Suppose somehow, you could increase the response rate from 5% to 6%. The question is, what is the impact on orders, orders or what is the impact on revenues? You say the average order size, order is, this is not order size, order value you can rather put it. The revenue is 1500. It is not the volume, but it is given in euros and therefore, obviously, it is the revenue, not the profit.

Let us take it that way. So 5% to 6% increase in response rate. What does that 5% mean? If you reach out or if you send your packages to 100 people, earlier 5 people used to respond. But you make an effort to increase the response rate somehow, say using analytics techniques, then it becomes 6%. Because you chose the target group more scientifically. So there is a 1% increase, 1% increase in the response rate.

What will be the impact on revenues? That is a question. And that is not difficult to calculate, instead of calculating that. So you know that it is 1% increase. So 1% of 2 million. So that is 20,000 more orders.

And 20,000 into average order size is 1500 or 1500. And that is equal to 30 million euros or the revenues increase by 30 million. We do not know what is the profit, but is a small increase of 1% in response rate, leads to a great change in the revenues. This is known as lift. This change or this increase is known as lift. Lift is the improvement in business performance caused by certain intervention based on analytics or data mining.

Certain intervention based on analytics or data mining improves the business performance and lift is a measure to calculate that. So now you can imagine, why a company like Netflix promises huge rewards like 1 million for a 10% increase in the performance of algorithms. And you can also imagine when you know this is all platform based business. So it is volume business. So when the numbers go up, the sort of impact on revenues would be very high.

So that explains the importance of performance of algorithms and the role of algorithms in certain businesses and particularly the platform based businesses and direct marketing businesses. So all these platform based businesses you know are direct marketing. They reach directly to the customers. Of course, they may be multi sided, but there is a direct contact with different markets or different categories of customers. And hence the importance of algorithms in this whole business.