

Management Information System
Prof. Saini Das
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Module - 02
Foundations of Business Analytics
Lecture – 08
Introduction to Data Mining

Welcome back! So, in the previous lecture, we had spoken about data warehouses, and how they are used to derive business intelligence in an organization. Today's lecture will focus more on 'data mining' and its various applications.

(Refer Slide Time: 00:30).



Why data mining? Because data mining we had discussed is one of the important tools that is used to derive business intelligence from data that is stored in the data warehouse.

So, this speaks all about you know why data is required in organizations, in god we trust all others must bring data was mentioned by W. Edwards Deming and who is he? He is a great engineer statistician professor and a management guru. So without data nobody will trust you. So, the world is all about data and data is supposed to be the gold today.

(Refer Slide Time: 01:13)

"There is a striking correlation between an organization's analytics sophistication and its competitive performance."

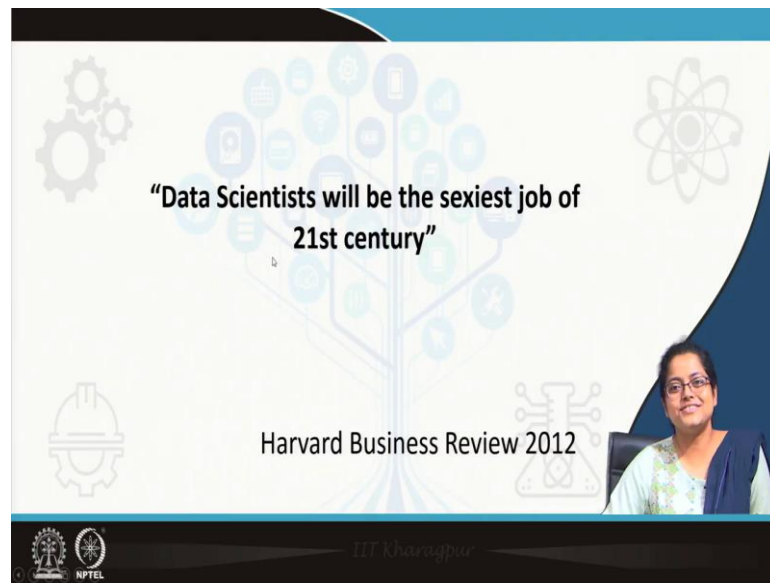
10 Insights: A first look at the new intelligent enterprise survey on winning with data, MIT Sloan Management Review, Vol 52, No 1, 2010

Dr. Kharagpur

So, moving ahead, this is an insight from MIT Sloan Management Review which came in 2010, 10 years back, which says that there is a striking correlation between an organization's analytic sophistication and its competitive performance which is very-very obvious in today's date.

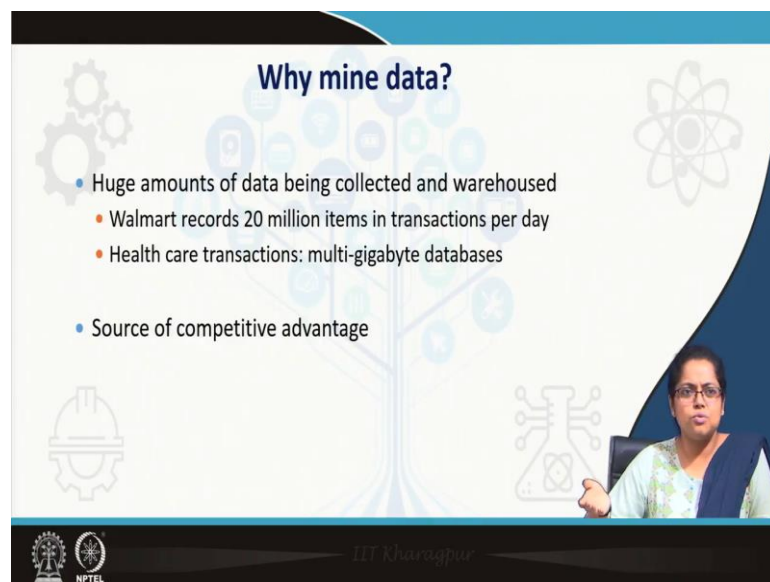
Because you know every organization is producing huge chunks of data. So, if you do not have the sophistication or the capability to analyze that data and use it for deriving your strategies or making your business decisions, you will lag behind your competitors. So, if you have better sophistication ah in analyzing your organizational data that will obviously need to better competitive performance of your organization. So, this is very true in today's date.

(Refer Slide Time: 02:14)



Harvard Business Review 2012, so this is very important for all the aspiring managers who would want to be data scientists in future ah. So, 21st century is going to be all about data science.

(Refer Slide Time: 02:33)



So, all of these you know make us think why is there so much of hype about data mining? Why do people mine data today why did not they then there was nobody spoke so much about data mining say 15 years back. So the reason lies here, a huge amounts of data are being collected and warehoused everywhere around us.

Earlier nobody thought about you know collecting so much of data, we did not have the technical sophistication the storage space the processing power to handle so much of data, but today we do have. So for example, Walmart collects and you know records 20 million items in transactions per day 20 million Walmart alone, health care transactions produce multi gigabytes of data.

So and of course, you would realize that in the; you know in the times of covid-19, analysing healthcare analytics has become so very crucial, so very important. Because without healthcare analytics we would not be able to predict trends, we would not be able to find clusters of you know you know red zones which are highly susceptible to the corona virus.

So, healthcare analytics is not only used in covid-19 scenario which is a pandemic, it was also very popular you know it was used a lot for multiple diseases. Because huge amounts of data are being you know collected on the fly healthcare data being collected on the fly and being analyzed on the fly. Today all of us wear smart watches. So, smart watches collect data on human being about you know your health related data continuously and stores it somewhere right, stores it and you can analyze that data.

So, that is why since huge amounts of data is being collected and warehoused, therefore why not mind that data because data is gold right. So as gold is you know there is a gold mining similarly data mining would actually help you in you know in coming up with insights which are not at all less precious than gold today right, they are not less precious than gold.

Moreover mining of data has become very important, because data is a source of competitive advantage for every firm; firms have a lot of data. So, if you analyze your data better and if you helped it you know make data will help you make better decisions data will help you come up with better strategies and all of that would give you a source of competitive advantage over your competitors.

(Refer Slide Time: 05:19)

The slide features a light blue background with faint icons of gears, a network, and a person. The title 'What is data mining and KDD?' is at the top. Two bullet points are listed, with the words 'valid', 'potentially useful', 'understandable', and 'actionable patterns' circled in red. The presenter, a woman with glasses, is visible in the bottom right corner. The NPTEL logo is in the bottom left, and the name 'Dr. Khuram' is at the bottom center.

What is data mining and KDD?

- Knowledge discovery in databases (KDD) is the non-trivial process of identifying **valid**, potentially **useful**, **understandable** and ultimately **actionable patterns** in data.
- Data mining is a **step** in the KDD process of applying data analytics and discovery algorithms

Dr. Khuram

So, moving ahead we have often heard about the term data mining, but are we aware of its actual exact meaning. Along with that today we will also discuss another very important term called KDD Knowledge Discovery in Databases, we will also try to understand what the difference between the two terms is.

So, what is the difference between data mining and KDD? Knowledge discovery in databases is the non trivial process of identifying valid here; I would be highlighting a few points valid potentially useful understandable and ultimately actionable patterns in data.

So, KDD is the non trivial process which means it is a very significant process of identifying valid. So, why did we highlight valid, because there could be a lot of you know patterns that you could get from data. Because as we mentioned the data is getting collected on the fly huge amounts of data.

So, there could be a lot of patterns which are irrelevant to you they are not valid they are not relevant. For example, you know if I say that there is a relationship between the number of trees in my neighbourhood and the number of pens I possess.

So, you know my limited understanding tells me that there is it could not be a relevant pattern, I cannot derive much insight or value out of this pattern. So, patterns that are derived should be valid or relevant potentially useful which is which is why we actually

go ahead and mine data. So, they should be potentially useful understandable why, because there could be a lot of patterns which are not easy to understand and even a more important not actionable right.

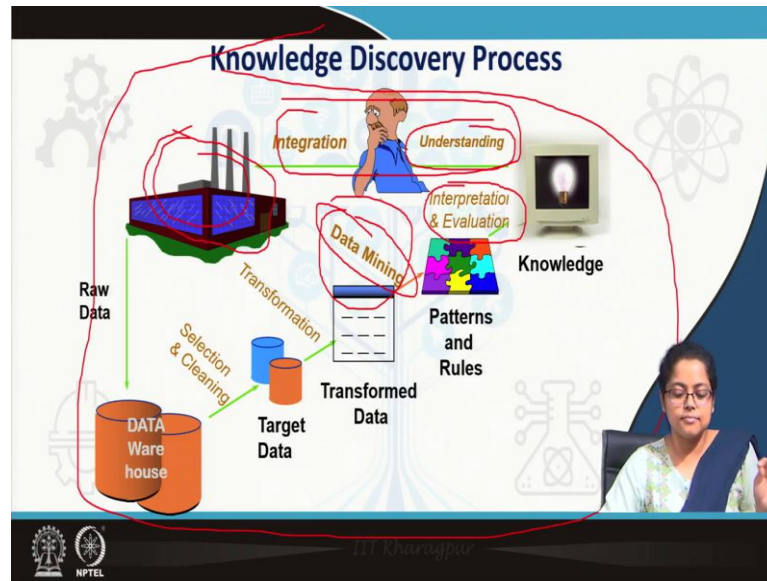
So, if you come up with an insight or if you come up with the pattern based on which you cannot take any action, it is just a pattern but you know it is its it is a pattern out of which it an organization cannot take any action; then that is not knowledge discovery in database. So, that does not amount to it.

So, what should happen is ideally, the pattern that we find out should be valid or relevant of course useful understandable. So, that you know the data analysts or the business persons are able to understand the pattern and they are able to take some action based on the pattern. That is you either take some business strategy out of the pattern or you make a decision out of that pattern.

For example, if you get a pattern that you know there are there are a group of users for a telecom provider who are at a high risk of churning or who are at a high risk of switching to another provider. You can take an action based on that, you can either give him some incentives or you can try to you know give him some lucrative offers to retain him or if you think the customer is of extremely low value you can let him go.

So, here we have a pattern or we have a finding or an insight based on which we can take some action. Therefore, it is very important to be able to take some action based on the pattern that we derive. Now data mining is just a step it is very important why I highlight, this it is just a step in the KDD process of applying data analytics and discovery algorithms. So KDD is a huge process, out of that data mining is only a step of finding some patterns. So, what are the other steps let us have a look.

(Refer Slide Time: 09:16)



So, here this schematic diagram will explain everything about data knowledge discovery in databases and where data mining stands. So, here we see a small factory it is it could be a factory or it could be a departmental store, it could be any departmental store that we see around us. Wherein there are huge amounts of data being produced every day transactional data storage data warehousing data distribution data a lot of data being stored every day.

Now, that raw data from the from the store departmental store gets transfer extracted transferred and loaded from the multiple databases of the departmental store as we had discussed in the previous lecture, all of them now get loaded into the data warehouse.

So, from the data warehouse the next step is selection and cleaning of data. Selection because there could be a lot of data that you collect from the departmental store, but not all of the data would be relevant to you; earlier also we had said that the data should be relevant.

So, you select only the relevant data that is suited to you and you then clean the data. So, what is data cleaning? Data cleaning also amounts to you know it is a part of pre processing of data, wherein you take certain actions such as you know you may want to remove some outliers, if your data has a lot of outliers or you may you know if your outliers is that the presence of outliers could be skewing the data. So, you would want to eliminate the outliers or you could be having a lot of missing data incomplete data. So,

you may want to take care of that or tackle that problem by either dropping some of the records that has missing data or taking an average of some of the records of the records ah, you know which the attribute which is having missing data and then filling the missing cells.

Or you; there are; there are multiple ways in which missing data is handled that itself is a; you know could take 3-4 lectures but we are not getting into that. So, since our post selection of data is cleaned as we have discussed and the selected cleaned data is called the target data.

So, the target data is the data on which the transformation operation is then performed. So, what is this operation called transformation, it is another very important operation in which you know data is transformed from one format to another.

There are a lot of you know and let me give you some examples. So, for example, you could have data in Fahrenheit. So, temperature data in Fahrenheit, but prior to analysis you would want to change it to Celsius, so that is one way of transformation. Similarly, say you have two attributes, but you do not want to use those two attributes for your analysis, rather you would want to use a third attribute which is derived out of these two attributes. So, the third attribute is called a derived attribute and the derived attribute or the transformed new attribute is going to be used in your analysis.

For example, I may have, you know GDP data of a kind of multiple countries and there the population, but I would want to use GDP per capita; as my; you know a variable. So, what I would do is input variable. So, what I would do is I would you know use these two attributes GDP and the population and find a new attribute called GDP per capita which is the derived attribute that I will use in my further analysis.

So, there are a lot of other examples such as logarithmic transformation which is used for data sets, maybe you know which are which are of distributions which are highly skewed to reduce the skewness prior to data mining.

So these are certain examples of transformation, there are a lot of other examples of transformation. But they are very context specific and I would not want to discuss all of them here because they are beyond the scope. Now, post transformation the transformed data is then mined. So, here we see that out of this entire process that we see here data

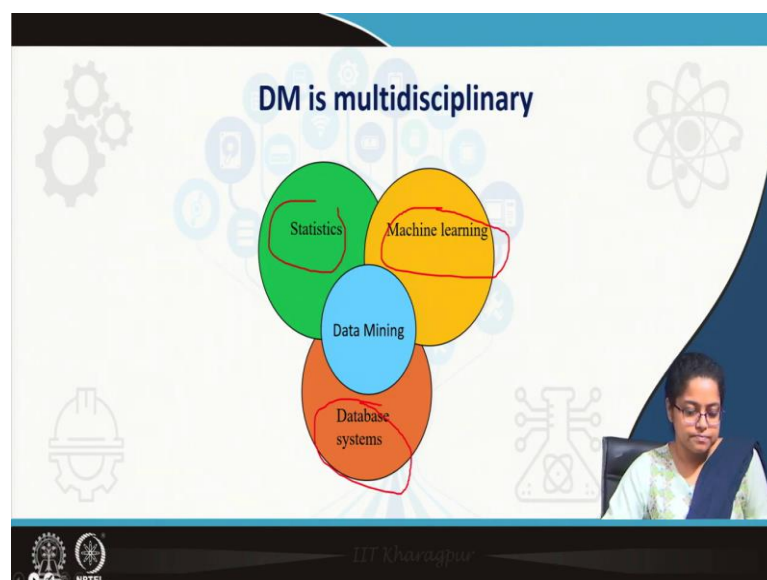
mining is just one step, it is just one step. It is one step it is the particular step of finding patterns and rules in the data. So now the transformed data is fed to some algorithms are run on the transformed data and you do identify some patterns and rules out of the data.

Post which once you find out the patterns or rules or the outputs of the algorithm, the output is interpreted and evaluated to derive knowledge. So, finding the output is not enough the data analyst has the rule of interpreting and evaluating the output to derive knowledge out of it.

Post the knowledge it is very important to understand how this knowledge can be useful for the organization. So this is the step where finally understanding and integration happens or the knowledge that you derive out of the entire out of the process is now transformed to something that is actionable.

And that actionable output or knowledge is now sent to back to your departmental store, so that it can add some value to the departmental store. This entire process is called knowledge discovery in databases and data mining once again I would like to reiterate is only a step. I hope the process of KDD and the role that data mining has in the KDD process is now clear all right.

(Refer Slide Time: 15:28)



So, let us move ahead and here this particular slide talks about the fact that data mining is multidisciplinary. So, data mining draws from different disciplines as we see of course

database systems play a very important role, we saw the entire data is stored in databases and then sent to you know data warehouses; from where we are actually using the data for data mining. Statistics and machine learning both play a very important role in analyzing the data to come up with outputs. So, of course, data mining is multidisciplinary and draws from different disciplines moving ahead.

(Refer Slide Time: 16:12)

Data Mining Applications

Typical Applications

- Customer Segmentation**
- Propensity to Buy
- Profitability Modeling & Profiling
- Customer Attrition
- Channel Optimization
- Fraud Detection

What are my market segments and who are my customers by segment?

Personalize customer relationships.
Higher satisfaction = Higher retention

The slide features a background with gear and atom icons, a woman at a computer, and the NPTEL logo at the bottom left.

Let us see some very interesting applications of data mining in the business context. Customer segmentation, in marketing very important role played by data mining. So, what are my market segments and who are my customers by segment? So, you find out your market segments and then you find your customers who reside in each of those segments, post which what you do is personalize customer relationships.

(Refer Slide Time: 16:54)

Data Mining Applications

Typical Applications

- Customer Segmentation
- Propensity to Buy**
- Profitability Modeling & Profiling
- Customer Attrition
- Channel Optimization
- Fraud Detection

Which customers are good candidates for our new long distance calling plans ?

Targeting customers based on their needs.
More product sales = Greater loyalty

The slide features a background with gear and atom icons. A woman is visible in the bottom right corner, and a man is shown at a computer workstation. The NPTEL logo is in the bottom left.

So, if you know who are your customers by segments you could give them appropriate strategies. Propensity to buy is another area where data mining has a very important role, which customers are good candidates for our new long distance calling plants very important for telecom a telecom sector. So, targeting customers based on their needs, that is what you need to do.

(Refer Slide Time: 17:14)

Data Mining Applications

Typical Applications

- Customer Segmentation
- Propensity to Buy
- Profitability Modeling & Profiling**
- Customer Attrition
- Channel Optimization
- Fraud Detection

What is the life time profitability of my customers ?

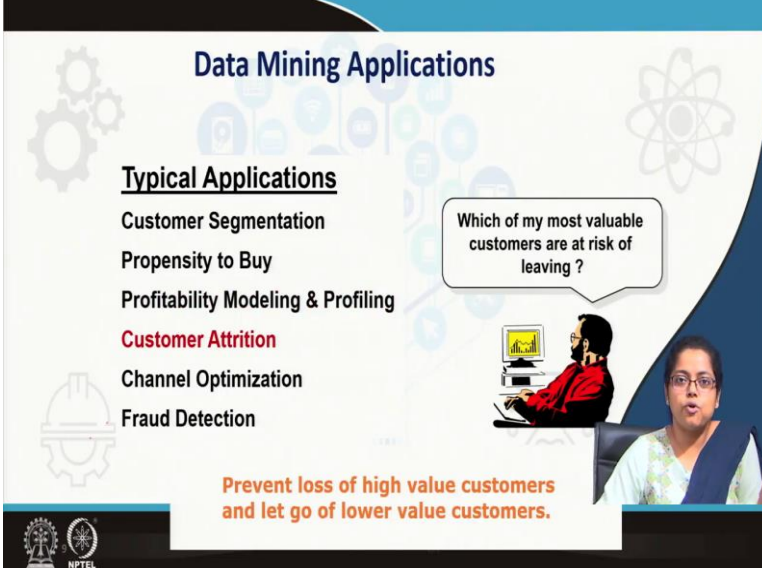
Increase high value customers based on current & future profitability.

The slide features a background with gear and atom icons. A woman is visible in the bottom right corner, and a man is shown at a computer workstation. The NPTEL logo is in the bottom left.

Profitability Modelling and Profiling, so what is the lifetime profitability of your customers of my customer. So, once I understand the lifetime profitability of my

customers, I would be able to increase high value customers based on current and future profitability. And if I see there are some low value customers I may either try to convert them to high value customers or if I you know if I see that you know there is no value I may actually let them go.

(Refer Slide Time: 17:46)



Data Mining Applications

Typical Applications

- Customer Segmentation
- Propensity to Buy
- Profitability Modeling & Profiling
- Customer Attrition**
- Channel Optimization
- Fraud Detection

Which of my most valuable customers are at risk of leaving ?

Prevent loss of high value customers and let go of lower value customers.

NPTEL

Another very important application is in predicting customer attrition or churn. So, who are my most valuable customers; who are at a risk of leaving? If I can understand this I can try to prevent the loss of high care value customers and let go of lower value customers.

(Refer Slide Time: 18:06)

Data Mining Applications

Typical Applications

- Customer Segmentation
- Propensity to Buy
- Profitability Modeling & Profiling
- Customer Attrition
- Channel Optimization**
- Fraud Detection

What is the best channel to reach my customers in each market segment?

Interact w/customers based on their preference.

The slide features a list of typical data mining applications. 'Channel Optimization' is highlighted in red. A speech bubble asks, 'What is the best channel to reach my customers in each market segment?'. Below the list, there is an illustration of a person at a computer and a woman speaking. The NPTEL logo is in the bottom left corner.

Channel optimization; what is the best channel to reach my customers in each market segment? So, this is again very important. Finally fraud detection, so data mining there is a huge section of data mining that is dedicated to fraud detection.

(Refer Slide Time: 18:17)

Data Mining Applications

Typical Applications

- Customer Segmentation
- Propensity to Buy
- Profitability Modeling & Profiling
- Customer Attrition
- Channel Optimization
- Fraud Detection**

How can I tell which transactions are likely to be fraudulent?

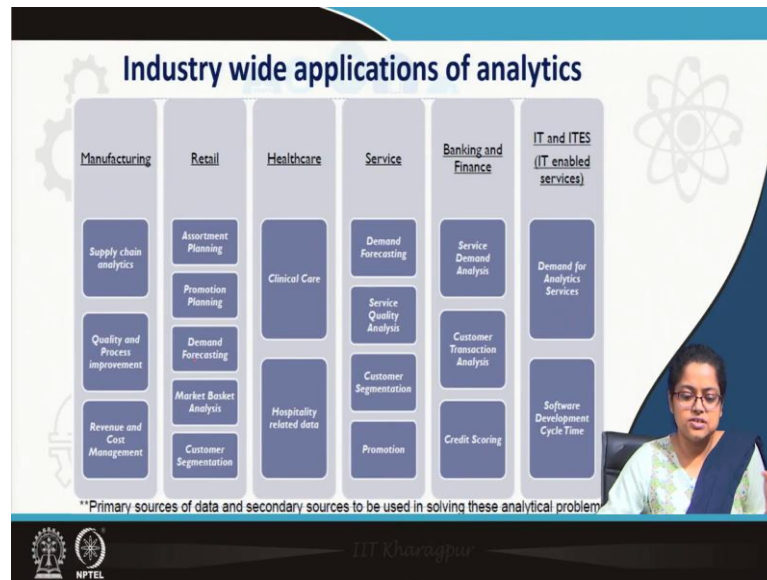
Detect and prevent fraud to minimize loss.

The slide features a list of typical data mining applications. 'Fraud Detection' is highlighted in red. A speech bubble asks, 'How can I tell which transactions are likely to be fraudulent?'. Below the list, there is an illustration of a person at a computer and a woman speaking. The NPTEL logo is in the bottom left corner.

So, we see fraud an all-around right in marketing; there is fraud in finance; there is fraud. So, you know fraud detection and data mining go hand in hand. So, how can I tell which transactions are likely to be fraudulent. So, if I can predict transactions that are likely to be fraudulent in future, I could take a lot of measures to detect and prevent the fraud in

order to minimize loss. So, if there is fraud of course, there is a lot of loss to the organization. So, it is very important to try to detect the fraud a priori, so that measures can be taken to prevent the fraud.

(Refer Slide Time: 19:02)



So these are certain applications, now here industry wide applications of analytics. Here we see that in every domain manufacturing retail healthcare service banking and finance, IT ITES, everywhere analytics and data mining play a very important role.

(Refer Slide Time: 19:25)



Moving ahead here there are some other applications which I would really like to focus on in marketing. Which customers are likely to respond to a particular campaign? So, if you know that you can so in the world of digital marketing this is very important today, because every organization is into digital marketing.

So, if you would be able to predict the customers that are likely to respond to a campaign, not only in digital marketing in the world of physical marketing also you can target the campaigns accordingly.

Which customers are likely to be profitable of course very important to know? Who might want to buy a particular product? So, if you want to cross sell or sell or up sell. So, either you want to sell some complementary products to a customer who has bought a to a certain product or you would want to sell some higher value product. If you can predict you can target accordingly, so that helps in you know wasting a lot of calls or you know a lot of money on targeting a customers who might not be interested at all right.

Similarly in you know there are a lot of other applications of data mining and analysis analytics in marketing. In telecom of course, we had already discussed which customers are vulnerable to attrition or at a risk of churning. So, this would help us in identifying them and in order to devise strategies to prevent them from churning, based on these symptoms where our problems located in the network.

So, if they are there is a section of you know of zone a particular zone or a particular region, where customers are churning more we would be able to identify that you know the look into deeper into that particular region and check out the network problems that could lead to the churn ah.

In finance and insurance again analytics has a very important role. So, which customers are credit risks or insurance risks? So, you may then decide whether you would want to you know let us the customer take a policy or take let the customer take a credit based on the risks which claims or credit transactions are fraudulent very important right. Similarly there are a lot of other examples in finance and insurance.

(Refer Slide Time: 21:47)

Some other applications (Contd..)

- **Healthcare**
 - Which patients may take longer to recover ?
 - What is the likely cause of this illness ?
 - Which patients are at risk of disease (and might benefit from medication)? Pfizer pharmaceuticals used data mining to construct a predictive model that was then embedded in their online cholesterol health risk assessment, which tells patients their cholesterol risk score. High risk patients can consult their doctors and request Lipitor, Pfizer's cholesterol medication.
- **Retail**
 - Which products do customers buy together (or in sequence) & which do they not buy together ? ('Category management'.)
 - What characterizes customers at various stores ?
 - What items are bought for cash, on credit, or by check ?
 - What type of customer buys this item, or this product type ?

NPTEL
Dr. Khuram

Health care we have discussed that you know in the world of covid 19 healthcare analytics plays a very important role. Otherwise also in the world of healthcare analytics has a major role to play. So, which patients may take longer to recover; what is the likely cause of a particular illness; which patients are at risk of disease?

So for example, Pfizer pharmaceuticals use data mining to construct a predictive model that was then embedded in their online cholesterol health assessment risk assessment, which tells patient patients their cholesterol risk score. So, once you know your risk score you can take measures as to not eat that egg that you would like to eat every day ok.

So, jokes apart coming back in retail ah. So, ah analytics help or the mind data mining would help you find out which products do customers buy together or which products they do not. So, this would again help you come up with strategies. So, I will discuss this in detail in my in one of my subsequent lectures. So, what can you do if you find out that certain products are purchased together and certain products are not?

Now what characterises customers at various stores ah. You know what type of customers buy certain items or product type, which items are generally bought with cash which are bought on credit cards on credit or which are bought by cheque. So, if you know such if you have such insights from data you can take measures a priori which would help you minimize a losses later.

(Refer Slide Time: 23:29)

The slide is titled "Data Mining Applications (Contd..)" and features a background with various icons including gears, a tree, a hard hat, and a circuit board. A woman is visible in the bottom right corner of the slide. The content is as follows:

- **Quality Control**
 - Which shipments are high-risk and need to be inspected ?
- **Customer Support**
 - Which tasks schedule (ordering) is optimal (or good enough) ?
 - Which customer service representative should I assign to a task ?
 - What documents or people are likely to be helpful to the customer in solving their problem ?

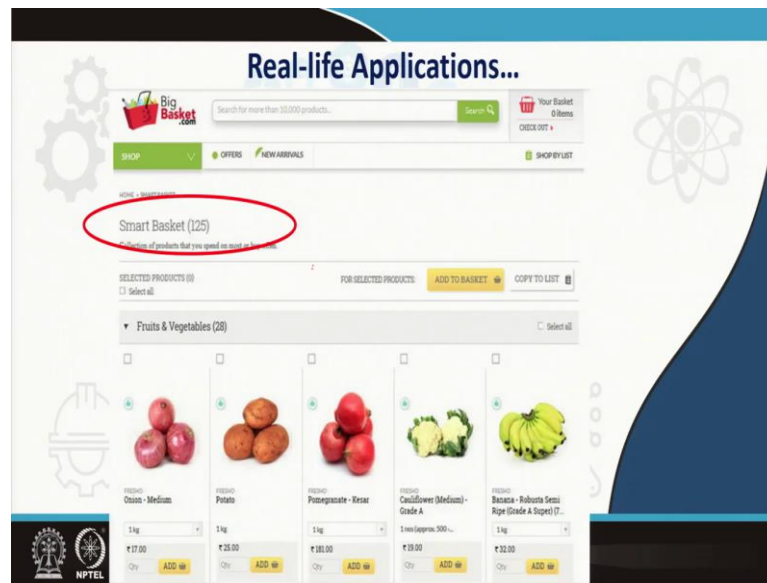
At the bottom left, there are logos for IIT Kharagpur and NPTEL. The name "IIT Kharagpur" is written in the center at the bottom.

Quality control very important, because all organizations today take measures to control quality of their products or shipments very true for E commerce companies today. So, if you know that you know especially in the marketplace model, where many a times you know the marketplace provider has to keep a tab on the quality of the product that is shipped by the supplier. So, for the marketplace provider it is a lot of you know consumption of resources in terms of manpower time etcetera, to inspect each particular shipment for quality.

Therefore if analytics could predict in some way shipments which are at high risk, then companies could actually you know try to check only those high risk and medium risk shipments and maybe you know not check the low risk shipments. So, this would save a lot of resources in terms of time, money, manpower, etc. And again in customer control customer support everywhere there are a lot of examples here I think you can go through them.

So, here also data mining and data analytics play a very important role. So, here we have extensively discussed a lot of areas in business where data mining and analytics play a very important role.

(Refer Slide Time: 25:02)



Moving ahead some real life applications, the earlier ones were also real life, but these are you know this is taken from a particular website. So, here you know this is a website called big basket which is a very popular online grocery delivery you know platform in Indian context. So, they have this feature called a smart basket, which is a feature it is a collection of products that you spend on most or you know so what you know buy most.

What happens is when you are trying to shop grocery from this particular forum or platform and you would want to maybe you know you have shopped for you are a regular here. So, one particular day you miss out some products that you regularly shop in general.

So, this will give you this will prompt you and this will this will use analytics to say that here to find out that you know this particular product is something that you shop regularly, but today you have missed it out. So it will give you a prompt, so that if you have actually forgotten you may quickly want to include that in your basket.

(Refer Slide Time: 26:17)



(Refer Slide Time: 26:19)

At Flipkart...

- Forecast demand for each SKU.
- Predict customer cancellations and returns.
- Predict what a customer is likely to purchase in the future?
- How to optimize the delivery system?

The slide features a background with a network diagram of nodes and lines, and decorative icons of gears, a hard hat, and a circuit board. The Flipkart logo is also visible in the background.

Similarly, in other; Flipkart is another very popular E-commerce company in India, home-grown Indian E-commerce company. So, in Flipkart, analytics again plays a very important role. So, you can forecast demand for each SKU in advance, so that would help you warehouse and manage inventory better. You can predict customer cancellations and returns this is very important because, E commerce companies in general lose a lot on product cancellations and returns.

So, research suggests that 30 percent of products ordered or purchased are returned to E-commerce vendors. So, if you can predict customer cancellations and returns you may take measures to minimize or to reduce those cancellations and returns and that would help you prevent losses.

Also it would you can try and you know at Flipkart analytics is used to predict what a customer is likely to purchase in the future. So if you know that you can store accordingly, you can recommend those products to the customer. So, that there is a higher probability that the customer actually purchases.

Again you know at Flipkart that the entire delivery system is optimized using analytics. So, that helps in saving time that helps in better scheduling and that helps keep your delivery persons happy, because delivery personnel's have a very role to very important role to play for any commerce vendor. So, if they are happy your customers would be happy and your business would be happy at the end of the day all right.

(Refer Slide Time: 27:58)

The slide features a white background with a blue header and footer. The title 'Diaper-beer syndrome' is in a large, black, serif font. Below the title, there are three paragraphs of text in a smaller, black, sans-serif font. The first paragraph reads: 'IT'S PART OF the folklore of data processing. A retail chain put all its checkout-counter data into a giant digital warehouse and set the disk drives spinning.' The second paragraph reads: 'Out popped a most unexpected correlation: sales of diapers and beer.' The third paragraph reads: 'Evidently, young fathers would make a late-night run to the store to pick up Pampers and get some Bud Light while they were there.' In the center of the slide, there is a photograph of a Pampers baby wipe pack and a bottle of Kingfisher beer. To the right of the text, there is a small inset video of a woman with glasses and a blue sari, who appears to be the presenter. The slide is decorated with faint icons: a gear, an atom, a toilet, and a circuit board. In the bottom left corner, there are logos for IIT Bombay and NPTEL.

So, finally, this is a; this is called the Diaper-beer syndrome. So, this is an area where analytics or data mining predicted something that was out of the blue that nobody had thought of before and this is an area this is a classical case of the rule of data mining in business.

So, this explained you know that you know when you have a lot of data you can come up with patterns which are unthinkable and but actually happens and that can help you in deriving huge benefits. So, it is part of folklore of data processing retail chain put all its checkout counter data into a giant digital warehouse and set the disk drive spinning.

Out popped the most unexpected correlations it was it was not at all expected was unthinkable, but it popped out. Sales of diaper and beer what turned out is that your young fathers would make a late night run to the departmental store to pick up pampers and get some bud light while they were there.

So this let me explain this a little more. So, evidently when data from departmental stores in the in a particular region were was analyzed, it was found that especially or specifically on Friday evenings late night when young fathers would visit departmental stores, they would buy diapers and beer together.

So, this is an; this is; this was an; this was a pattern which was unexpected absolutely; unexpected some way you know, you may even think; it as a spurious correlation. But what turns out is there is there is a reason behind this or it is part of a folklore of course but it is very important to know the reason behind this.

So, it was made you know people take a lot of guesses, but in my opinion the most possible reason could be that a young mothers wanted to party on Friday evenings leaving the young fathers to babysit. And while babysitting they would definitely have to purchase diapers and along with it to entertain themselves they would purchase some beer. So, this was a incredible insight or unexpected insight unexpected correlation that happened from and that came from data analytics.

And based on this, it was found that these two products – beer and diapers, were purchased together and this drove a lot of marketers to device a lot of strategies for store layout in certain departmental stores; we will discuss those strategies in the subsequent lectures; all right; thank you!

So, in this session, we have discussed knowledge discovery in databases, and what data mining is; the definition of both the terms, the difference between the two, and we have also discussed a lot of applications of data mining in the context of business. In the next

lecture, we will talk more about the various categories of data analytics and the various techniques of data analytics with their appropriate applications; ok.

Thank you! Till then, see you around!