Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:05
Lecture:21

ANALYTICS PROCESS CASE | Business Intelligence & Analytics

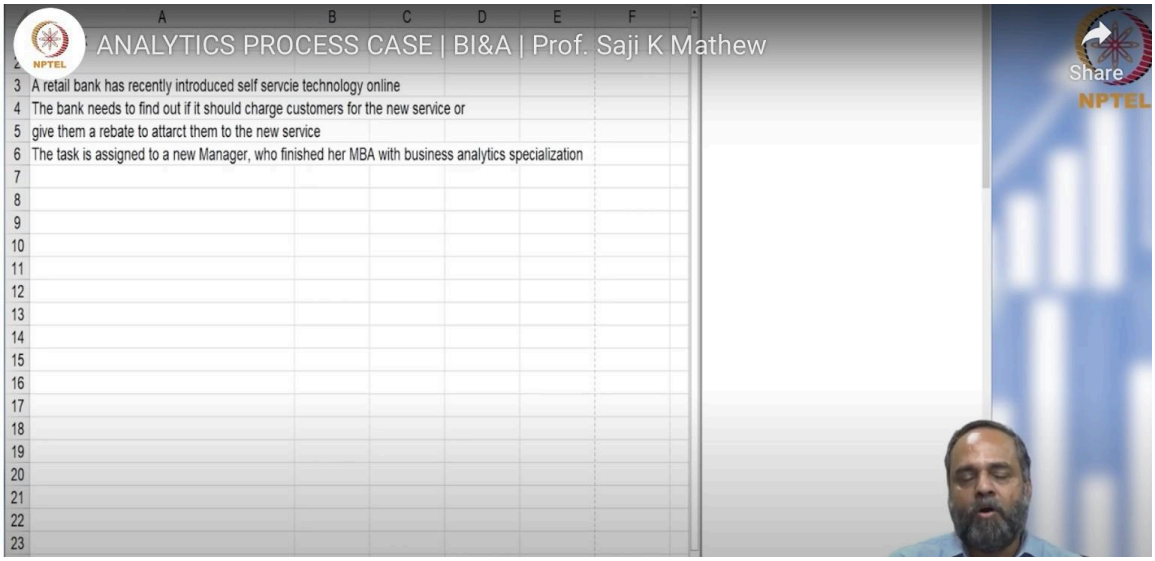
Hello and welcome to this session, which is an extension of our discussion. Related to analytics process or data mining process. We learned certain fundamental lessons related to the third process, identifying variables, selection of variables, and subsequently collecting data, building a model, testing a model, if it is a prediction model, and then deploying the model, taking feedback and so on. So that is the process. So in order to understand the nuances involved in the process of analytics, right from problem to solution, I am today going to discuss a case. This is a case which I have sort of customized for my audience, so that you understand the problem broadly, but by looking at the data, which I am going to demonstrate to you, you will be able to appreciate the problem and the nuances or the details of the problem.

And then see what solution path one could follow. And then when you follow that solution path, where do you actually sort of stop and then realize if the path that you follow is right or wrong, etc. So it is basically to illustrate analytics process, you know, and the importance of following a sound process, particularly the thought process, that is the purpose of this particular case. So the case in point is that of a bank, a retail bank, I have titled the bank as XYZ bank.

And obviously, it is a fictitious name, I have camouflaged the original title. So just assume that there exists a retail bank. You know, in retail banking, there are customers who deposit money or bank takes deposit or bank loans. And they have actually interest income from these transactions that the bank conducts on its customers. And the context is that the retail bank, which is a growing bank, wanted to keep itself updated with current technology.

And you know, banking technology has been advancing over the past several decades, right from the use of ERP systems for automation, business process automation, to use of BI analytics, etc. alongside. And then with internet, you know, there came online banking, online banking actually added convenience to customers, meaning a customer instead of going to a branch could do banking sitting at home or wherever one is, that is

the next step, of course, but one could actually access one's bank account remotely through internet and that is known as online banking or internet banking. And then anywhere banking, of course, with smartphones and you know, with advancements in devices, you can you know that you can also do banking through your phone. So those are, these are generally known as self-service technologies, self-service technologies.



What does that mean? That means that you do not need the service of a banking staff. You do not need a bank staff to service yourself on banking. You can service, you can be the service provider for your own banking requirements. You log in, you check out, check the account balance, you do not have to go to the branch, you deposit money or you withdraw money. And all that actually, withdraw means you transfer money and so on.

So I was talking about self-service technologies for banking. And what that means is the customer actually substitutes the staff or essentially you can imagine the bank can cut down on human resources or head counts can be reduced and obviously self-service actually reduces cost for the bank and that is why they invest in self-service technologies. And as far as the customer is concerned, it is you know convenience, it is customer's convenience. So you cannot go for it because it adds convenience to the customer. So both way there is a win-win situation and therefore you know these technologies have been increasingly adopted by banks.

And you can see that XYZ bank has invested in a particular self-service technology. And let me not be very particular about what is that technology, but for simplicity let us take it as online banking. Online banking is nothing new, but let us take a typical case

which is about providing online banking as a new feature or a new service channel for customers. So this is something the bank is introducing. Now when the bank has invested in this technology and the technology is rolled out, the bank management has a question.

Well, it is a new investment, it is a new technology and the bank has incurred certain costs for in this new technology. So therefore, in addition it is also going to create value for the customer. What is the customer's value? Customer's value is in convenience. In literature this is known as transaction utility. The transaction itself generates certain utility in terms of convenience and comfort for the customer.

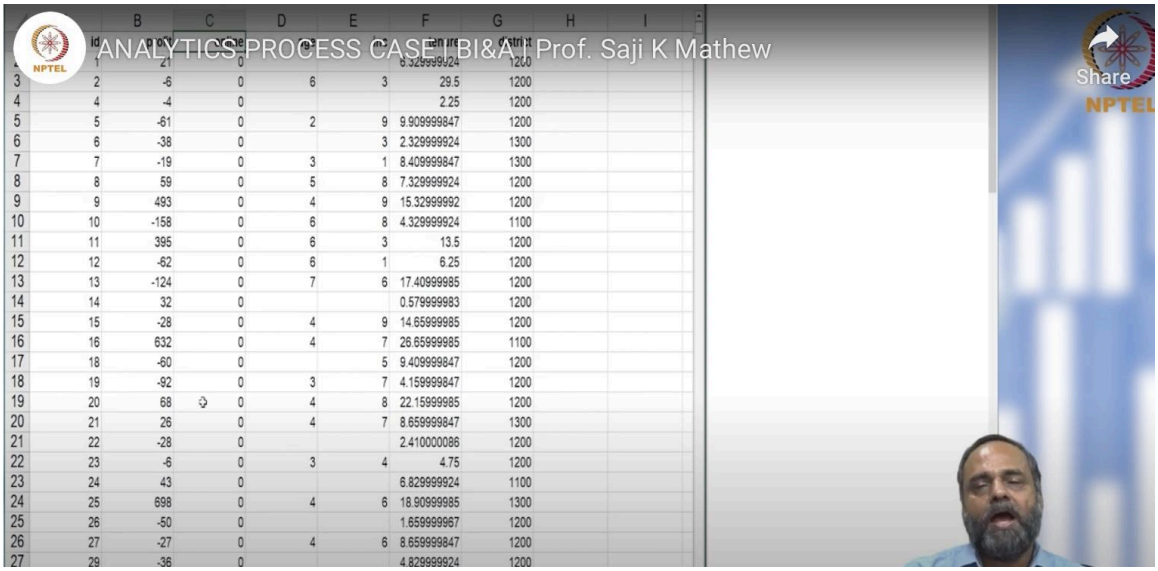
So there is a certain utility in the transaction itself, which is a value created for the customer. And therefore, why not charge the customers for it? Because there is a new value that is created. And of course, there is a cost that the company has incurred, etc. So that is one argument that the bank has. The other argument is, well, will people adopt the new technology? Like in the case of ATMs, you know, if you if you actually look at the history of ATMs, when ATMs were first deployed, users refuse to accept it because they are used to going to the branch and talking to the staff and you know, doing their banking transactions.

And they therefore, they were not comfortable using the ATMs. You know, maybe I am talking about the 60s or 70s during that time, but later on, you know, users adopted it. So there is a chance that many customers will not adopt online banking, they may actually still go to the branches, in which case the bank's original purpose of reducing headcounts and reducing costs actually does not materialize. And therefore, the bank says, why not actually give it free and not only free, give them a discount or give them some benefits, if they use online banking as compared to branch banking. So these are two arguments that is going on in the bank.

And the decision makers need to take a decision, give a rebate or charge. That is the question. And it is at that point in time that a new manager passed out of a business school of a top business school with a specialization in business analytics, joined the bank. And since the new manager is really good in working with data and solving problems, the bank decides or the middle level management of the bank decides at this point, why not employ the new managers and ask him to do some sort of use some sort of analytical approach to this problem and see which decision is more, you know, which decision is better, which leads to better outcomes. So what is the decision that the company should take? Give a rebate or charge.

So that is the problem. The business problem is for the new self-service technology, should the bank charge the customers or should the bank actually encourage by giving a rebate to adopt the new service. And now the new manager thinks about the problem. We have seen there is a thought process in analytics. This problem is understood, but how to solve this problem using analytics? That is actually bothering the manager.

And she goes back home and thinks about the problem. And of course, this is a new job and one has to prove oneself. And this is a big opportunity to apply one's analytical skills. And hence, the so she frames something in the mind and goes back to the organization the next day and ask for some data. So there is some idea that the new manager has figured.



And that is to look at the customers or look at a subset of customers or a sample of customer base and look at how customers are performing and particularly using some measure of performance. So what is done here is that the manager goes to the IT department and ask for some sample data. And of course, the IT department will ask what is the sort of data that you want and then she said maybe share with me some 30-35,000 records of customers and the customers should be both online and offline. Of course, the online is rolled out and some customers have already adopted it and many have not adopted it yet. But basically the idea of the manager is to see what does the data speak about the profitability of online versus offline customers, which category is more profitable? Well, if any of us are given this problem, we will all think in similar ways, which customers online customers are more profitable or offline customers are more profitable.

Look at the thought process. If online customers are more profitable, it makes sense for the bank to make more customers online, attract them because the profitability of the bank will go up, that is the assumption. On the other hand, if there is no difference in profitability, then probably the bank can charge online service because it is sort of cost to the company or it is an opportunity to generate another source of revenue. So this is the thinking, this is the thought process of the manager and let me act as the bank manager, the young bank manager who thinks in this line and then collect data and try to do some analysis. And let us see where we will be heading to.

So she asked for data and she got it. And along with that data, she also said, you know, if you can share some demographic data of each customer in terms of age, income, tenure and which district or district code of the customer, then that will be also useful, maybe demographic variables also has influence on profitability. And therefore, share all that data that you can share with me, give me a pilot, let me do a pilot study, share with me a sample data set. And that is the data set that is actually depicted here, you know. So you can see that in this spreadsheet, you see the ID of each customer that goes from 1, starts from 1, profit.

And obviously, in the first look itself, we can see some customers make profit and some customers make loss because there are negative values there. So there are both profitable and loss making customers in the customer base. And some customers are online, some customers are offline. 0 means not online or offline. There are customers who are online who will have the code 1.

And let us go down and see how many records are there. A lot of customers are there towards the end, who are actually online customers, it is 1. And we see that there are 31,633 records or 31,633 customers data were shared, and some of them are online, some of them are offline, their ages and as soon as you look at the data visually, you can see that not everyone has disclosed their age. So there is missing data. That is one issue that you see here as well as and even for income, there is missing data.

This is about income brackets. So one may be the lowest bracket, one is the lowest bracket, and then the bracket goes up till name as we visually observe the data. And of course, how long the customers have been with the bank in years that is also given and the district code of the customers. This is the data that the young manager received from the IT department. So you know, one is actually going to undertake a small pilot analytics project to generate some insight so that that insight can be used for decision making. So let us see as an analyst, what is some of the descriptive analysis one could do.

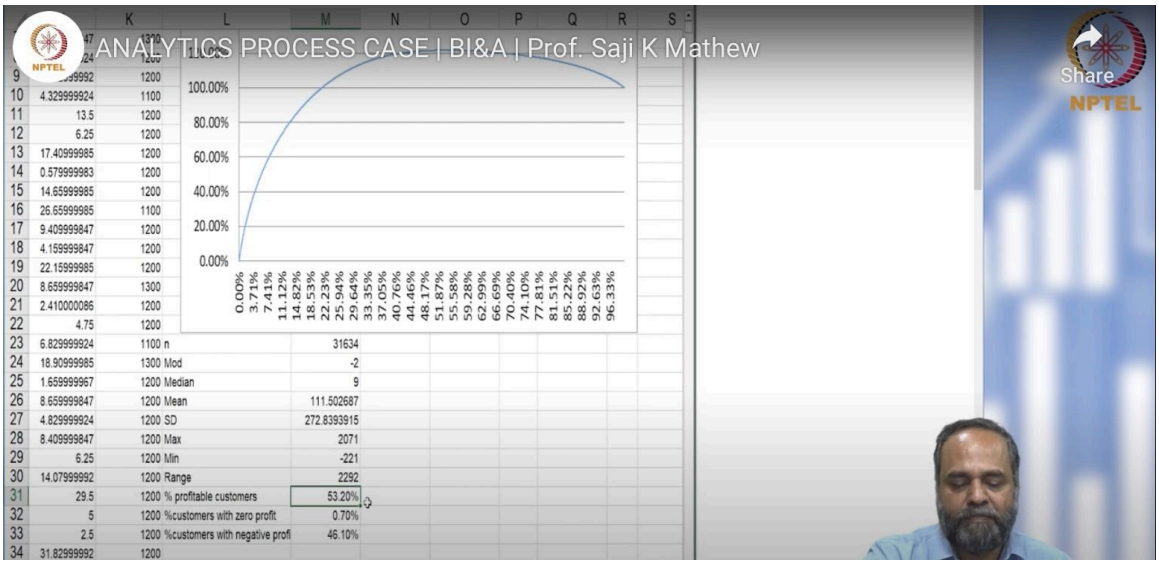
So in the next stage, what I am trying to depict here is something known as a profitability skew. A profitability skew, you can see the x axis and the y axis. X axis is about the percent of customers and y axis is about the percentage of profitability. And you when you, you know, when you move from left to right in the x axis, you can see that the cumulative, the cumulative profitability is going up and up. And when it reaches almost 30%, 40% here, you know, you can 44% you can see, you can see here it become, the graph flattens and then it starts falling.

What does that show? And you also see that, you know, the overall profitability, cumulative profitability goes beyond 100% up to 120% and then it starts falling. When you of course include all the customers, the profitability is 100. But in terms of the distribution of the probability, you can see that when you move from about 50% and beyond that the graph starts falling, which only explains that the customers, almost 50% of the customers constitute the whole profitability, the whole profitability of the bank. And then there are a lot of loss making customers, a lot of loss making customers. And that is why, you know, that explains why it is going beyond 120 and so on, because of the loss or the negative values in the profits.

So a lot of customers are actually generating loss for the bank. And of course, whether those customers should be fired, whether those customers should be retained, etcetera is another discussion. It is another problem, but we are not going there. You can see that in terms of exact values, when I did a descriptive analysis of the data, the sample size is 31,634, the mode of the profitability is minus 2, median is 9, mean is 111.50, and standard deviation is 272.83. And you see here that the standard deviation is higher than the mean and maximum value and minimum value, 2071 is the maximum profit, minus 221 is the minimum profit. And you can see the range here and percentage of profitable customers just 53.2, just 53.2 percent and 46.1 percent of the customers are, or 46 plus percentage of customers are not profitable.

Large customer base is not profitable in the data set. And there may be customers who do not do, those who do not make any loss or any profit or zero. So that might be explaining the difference between 53.2 and 46.1, it is not adding up to 100 percent. But in any case, this is a bank with large number, almost half the customers are lossmaking. Do you think that that is a small insight? Well, the new manager actually have brought a very interesting insight. It is just a descriptive graph, but it actually can serve as an eye opener for the banking management, bank management, because the bank is running with a lot of lossmaking customers. How to convert them into profit making customers? That is one question. And how to retain the really profit making customers? For example, you can see that if you look at the 80 percent value, you can see that about 12

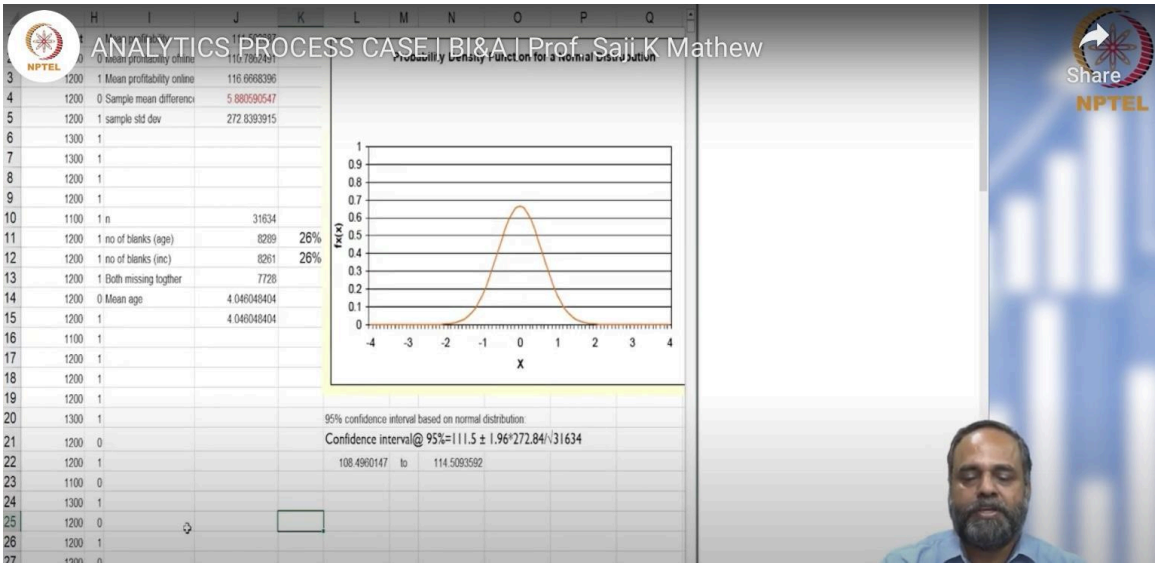
percent of the customers, about 12 percent of the customers contribute to 80 percent of the profit of the bank.



Is that a small insight? That is a big insight as far as this pilot data set is concerned. So it is critically important for the bank to retain those 12 or 15 percent of the customers because they are the major profit making customers. This is not related to the problem that we are discussing, but descriptive analysis of data provides or generates very useful insights for further action or further investigation as we discussed in one of the early sessions. Now our effort here is to understand according to the assumption that the manager has made, that is, look at the profitability of online customers and offline customers and understand which category is more profitable. And before doing that, I decided to create a confidence interval for the profitability.

What would be the sort of profitability for the population or for the entire customer base? If this 111.5 is the mean value for the sample, and that is done using a confidence interval estimate which is based on the t value at 95 percent. So I am adding a range, a positive range, a positive side and a negative side to that 111.5 to arrive at a range of values that would between which possibly if I take large number of samples from this population as the central limit theorem states, then the profitability, the average profitability will lie between 108.49 to 114.50. And we notice that it does not pass through 0, so well, it gives us some interesting information about the range of the profitability for the population. And we move on from there. And our aim ultimately is to understand whether online customers are more profitable or offline customers are

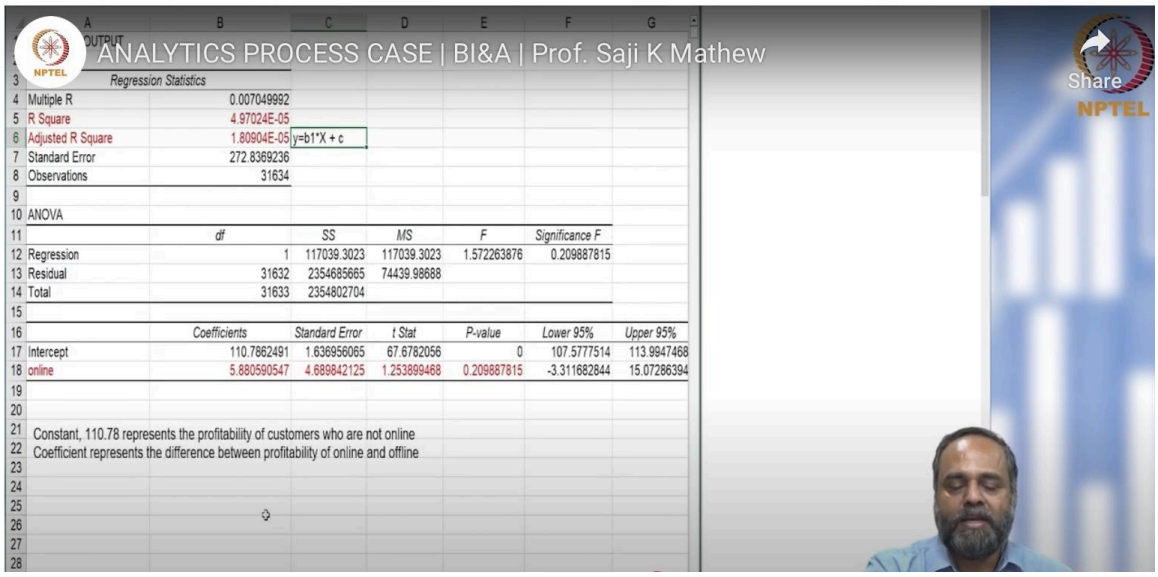
more profitable. Then you will also look at the category wise profitability. I did not show you some other data or some other results in the previous worksheet.



Let me do that before I move to the next worksheet. That is, I looked at the sample, I calculated the mean profitability of the entire customer base which is 111.50. Of course, I set up the confidence interval also. And then I filtered for offline customers alone. And the offline customers average profitability is 110.78. That is offline, those who have not adopted self-service banking yet. And mean profitability of the online customers who have already adopted online banking is 116.6 which shows for the given sample, the online customers are more profitable than the offline customers by a value of 5.88. 5.88 is the difference between the mean profitability of online and offline with a positive value meaning online is more profitable than offline. What does that mean? That only means that for this given sample, the online customers are more profitable than offline customers. But is that enough? Well, that is a starting point. But what worries us as analysts at this point is would this be the same for the population or would this be the same for the entire customer base. In analytics, we need to ensure that the results are generalizable.

The results are generalizable in the sense we look at, we need to draw inference about the population from the sample. So, we need to get into the inferential statistics. When we set up the confidence interval itself, it was a sort of inference of the population mean from the sample mean. Now instead of stopping there, we need to see if this difference will be significant. Will the positive difference between online and offline be significant for the entire population? Is it a significant result or did we obtain it by chance or we

obtained it by some reason? That is what we are trying to test in a hypothesis testing like this.



So, in order to test the hypothesis whether online profitability is greater than offline, you can do it in different ways. One is to do a t-test, other is to do a dummy regression. I am doing a dummy regression where I am doing a dummy regression because I can start with one variable which is, you know, two variables, the dependent variable will be profitability and my independent variable or the explanatory variable would be the online status, online or offline. In dummy, when you prepare a dummy variable for regression, you need to have only one variable if it is a binary valued variable.

You need to add only one variable which is online. Online equals 1 means it is an online customer, online equals 0 means it is not an online customer, it is an offline customer. So, one variable can actually stand, take two values and therefore one variable is enough to be defined in the model. And hence, that is what I am doing. So, I am actually using Excel to run my regression and I have online as the independent variable or in a model like this what is described here. I am actually suggesting my profitability is equal to b1 which is my coefficient into online.

Online is my variable which is a binary variable or a dummy variable here which takes either 0 or 1 plus an intercept which is c. And you can see the values for b1 and c which I obtained here in the regression model.

My intercept is 110.78 and my coefficient b1 is 5.88. You can see that we already saw what is 5.88 which is the difference of means between the two groups. And that is how the coefficient is estimated when you have a dummy variable, when you have a single

dummy variable.

And then look at the remaining results. The standard error is 4.68, t ratio is 1.25 as soon as you see this value, you would doubt whether the result will be significant and the p value you can see that it is 0.2 meaning it falls within the rejection region of a hypothesis. And with you know with alpha as 0.05, we need to reject this null hypothesis because p value is much greater than the alpha which is 0.05. Here it is 0.2. Sorry, I made a mistake. You need to, you cannot reject the null hypothesis which only shows that the difference between the two categories online and offline is not significant. p value 0.2 only means that the difference is not significant. You cannot accept the hypothesis that there is a significant difference because you, your null hypothesis is that there is no significant difference. Your hypothesis is that you have a, you know, let me write the hypothesis, your h1 there is a significant or let me write it as profitability, profitability of online customers is greater than profitability of offline customers on, on versus off.

My null hypothesis is there is, there is no, there is no difference between them or, or you are nullifying that there is a greater difference or there is actually a greater value for online customers. In this case, there is, you know, it is equivalent to saying there is no significant difference between them. They are the same. Or even if you got a difference in your sample, since p value is 0.2, we are saying that you just got it by chance. It is just a noise. You cannot actually generalize that. So these are the implications of a hypothesis test. You cannot establish p online is greater than p offline based on this hypothesis test. So you have to sort of assume or you have to actually accept the fact that there is no significant difference between the two. Now, I have put it in this way, but I am essentially saying no significant difference.

All right. So, what do I do? I am the manager who has to actually give certain results to the decision making body of the bank. And here I am finding that an online does not lead to more profit. And therefore, is there a point in adding more customers to online? Well, not in terms of profit, but if more customers are added to online, it reduces cost. It reduces headcount and that way there is a rational, but not for profitability.

There is no added advantage. And since there is no added advantage, should the bank actually give rebate? Doubtful. So, the manager, as a manager, you are sort of, you know, you are not able to give any convincing results or you are not able to draw any clear insight. You expected that there is a significant difference, but you did not get it. So, what I am trying to do again now here is, well, instead of stopping my analysis here, let me go one more step.

Let me add one more variable to the model. Here what I am trying to do is that in my regression model, instead of ending with one variable x1, which is online, I am adding an x2. My x2 here is, let me put it correctly.

There is a + c here and that is my model. My x1 is online. My x2 is age. I am adding age. So instead of ending here, I am actually profitability is equal to b2 × age + c. That is what I am saying. That is my new model.

So, I added that variable and re-estimated the regression coefficient. And then you can see that my, first of all, my first observation is that all my values, all my model parameters changed now. The coefficient in the previous model when I used only online, the b1 value was 5.88. Now my b1 is 27.18. So one lesson that we learn here is when we add more variables to a model, the model parameters of the previous variables, pre-existing variables also change. The estimation actually is done together for all the coefficients and it affects. There is an interaction effect here.

And you can see that age has a coefficient of 25.85 and online has a coefficient of 27.18 and intercept values 17.08. Now the interesting point to notice here is when I added one more variable which is age, the t statistics has changed and so are the p values. What do you see here? The online variable which was not significant previously has now become significant.

The p value is 8.58×10^{-7}. And the p value for age is 1.08×10^{-115}, meaning the p values are much lower than 0.05, the alpha. What does that mean? You have to reject, you reject this hypothesis. You reject this hypothesis, but you accept this hypothesis. When you add one more variable or essentially meaning that alongside age, there is a significant difference between profitability of online customers and offline customers.

How do you interpret this model? Let me actually help you interpret this model. Suppose there are two customers, two online customers, two online customers. One customer belongs to the age bracket 2 and second customer belong to the age bracket 2. One customer age bracket 1, another customer age bracket 2. Both are online customers. Who will be more profitable? The customer with age 2 will be more profitable because there is 25.85 × 2 comes there. So as age grows, the profitability in online mode increases. So if customers are online, for higher age groups, they are more profitable. That is the sense that you get from this model. Let us also give the alternate interpretation. Suppose there are two customers with the same age bracket, say 4.

Two customers belong to the same age bracket, but one is online, the other is offline. Which customer would be more profitable? The online customer will be more profitable because it is, there is a multiple of 27.18. For age remaining the same, an online customer will be more profitable than an offline customer by 27.18 times. That is the interpretation of this model. There are many other aspects in interpreting this model, but for our specific application, we are doing a basic level analysis. We have not even looked at the assumptions of regression here. This is a large data set, but relatively large data set, not a very small sample. So we just analyze this data and with the assumptions of regression, we interpret the results.

We find that age has an influence on profitability along with online status. So one insight that you get here is, well, since more aged customers who use online will be more profitable, you know, the manager might suggest, well, why do not you give rebate to customers who are senior citizens or in the higher age group? Well, that is something that comes to your mind. So you get something more. When you work with this model, you can add more demographic variables and see categories where profitability is higher for online and offline or for online.

Not offline, for online. So that is a further exercise, but I am not going to do it now. But reflecting on the problem, if we define the problem as the task of finding out whether online channel should be incentivized through rebate or it should be charged. As a manager, the third process was if online customers are more profitable than offline customers, well, give rebate and attract them. If not, do not do or do something else. Now, with these results, we have seen with age, well, the online profitability increases.

So you know, along with age, go for some sort of more specific action ideas. But am I right? As I said, analytics follows a thought process. So far, we are pretty good in doing the data analysis. We just, you know, do some early sort of, we are trying to do an early work.

It is not final. This is not what you are going to present because you have more variables. You also have to look at missing values because age has missing values. So, you know, this is an early attempt to see if, you know, this kind of approach will work or not. But then as you look at the data, you start worrying, you start thinking, well, a bank has to take decision based on my results. Am I heading in the right direction? Is my thought process right? And then a thought strikes me. Even if online customers are more profitable than offline customers in the database, can I make a recommendation that give rebate to the offline customers so that they become online so that profitability would increase? Let me say that again, can I make a recommendation to the bank that because online customers are more profitable than offline customers, give them rebate so that

more offline customers become online and hence becomes more profitable? Well, what is that idea? Can it be true that just by converting offline customers to online customers, they will simply become more profitable? Or what is the underlying assumption here? So, there is an issue here that is issue of causality.

I am assuming that the online customers are more profitable because they are online. How can I say that? This is only a correlation. Correlation is not causation. Correlation is not causation. Two things occur together does not mean that one is causing the other. There could be a third cause or there could be something else that will explain this correlation between two variables.

So in order to address the problem of causality, the analysis that I did is not just enough. There should be correlation, but I cannot show or I cannot conclude that online customers became more profitable because they became online. I need to look at what was their status before and after profitability. Suppose their profitability before becoming online and after becoming online, suppose it is P1 and P2. And if there is a significant relation, if there is a significant difference between the two, if P2 is higher than P1 significantly, then I am more comfortable to infer that online status have made them more profitable.

Having become online, they have become more of more profitable. But without having this data about before and after, I cannot conclude that online makes customers more profitable. And that is a fallacy in thinking process or thought process. If you follow a process, if you follow a model of regression analysis to conclude that online customers are more profitable than offline customers, and therefore, you know, give rebate to convert offline customers to online customers. That will be a fallacy. So as analyst, and as researchers, it is very important to be careful about your arguments, about your thought process.

Is it causality or correlation that you are testing? Are you inferring causal inferences from correlational analysis, then your entire analysis or thought process is wrong or is misleading. Therefore, that is another story that is a learning along with this case. So that is what I wanted to actually conclude. So all this data analysis is good to some extent, it gives you some sense of how the bank is doing.

But in order to conclude what should be done, this analysis is not enough. As I said, one may have to look for before and after and then do the analysis. And that is the conclusion of the case. Thank you very much and see you in the next class.