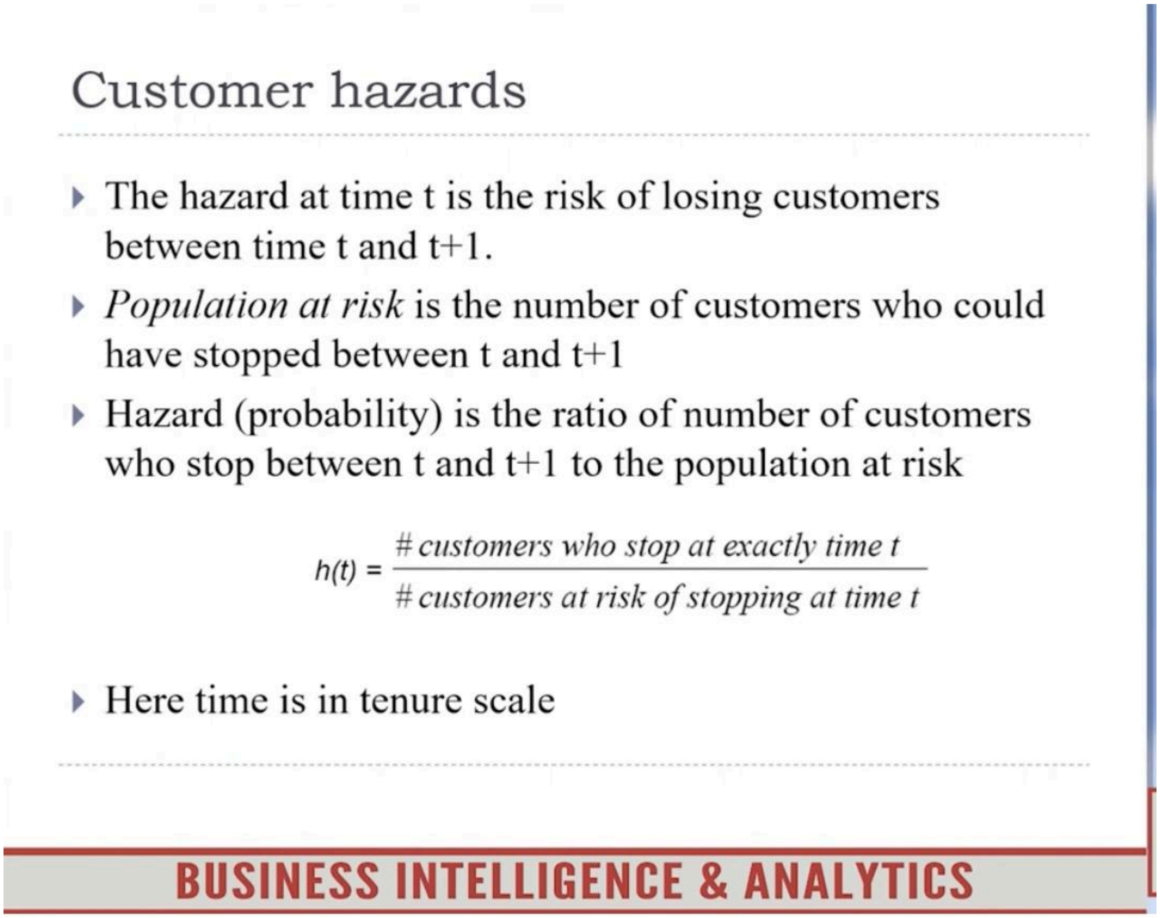
Course Name:Business Intelligence and Analytics
Professor Name:Prof. Saji.K.Mathew
Department Name:Department of Management Studies
Institute Name:Indian Institute of Technology Madras
Week:04
Lecture:14

CHURN ANALYSIS | Business Intelligence & Analytics

Welcome back. We will continue to discuss customer analytics. Particularly, churn analysis for which we laid some foundation in the previous module. And now we are going to look more closely at churn analysis and especially certain measures that are derived from tenure. So, the basic measure is tenure as I explained to you, but from tenure you can derive further measures or further values which are useful in estimating a customer's or a customer segment's survival, retention etc. That is what we are going to see now.



Customer hazards

- The hazard at time t is the risk of losing customers between time t and t+1.
- *Population at risk* is the number of customers who could have stopped between t and t+1
- Hazard (probability) is the ratio of number of customers who stop between t and t+1 to the population at risk

$$h(t) = \frac{\#\,customers\ who\ stop\ at\ exactly\ time\ t}{\#\,customers\ at\ risk\ of\ stopping\ at\ time\ t}$$

- Here time is in tenure scale

BUSINESS INTELLIGENCE & ANALYTICS

Now, yeah, so I explained to you what is a hazard of course, in relation to human life, but here hazard is in relation to customer. So, look at the definition, the hazard at time t is the risk of losing customers between time t and t + 1. So, what does that mean? There is a timeline, we know that customer joined at some point in time, and we can divide your time into tenure points. And now you are standing at time t and here is t +1 and here is t - 1.

And here of course, there is a, let us try discontinuity,say continuity. So, a t - 1, t, t + 1. A customer has survived from t -1 and reach the point t. So, in hazard and survival analysis, we analyze customers at discrete points in time, discrete points in time, it is not continuous timeline that we use, it is discrete timelines that we use. So, therefore, what it means is that customer defect at discrete points in time, customer defects at t point 1 t, t - 1, t or t + 1.

And we do not look at what happens in between. So, so, so tenure is measured like that also, you know that. So, the hazard at time t, is the risk of losing customers between t and t + 1. So, a customer is there at t, but the customer is no there when it comes to t + 1. There is a hazard, that is a hazard.

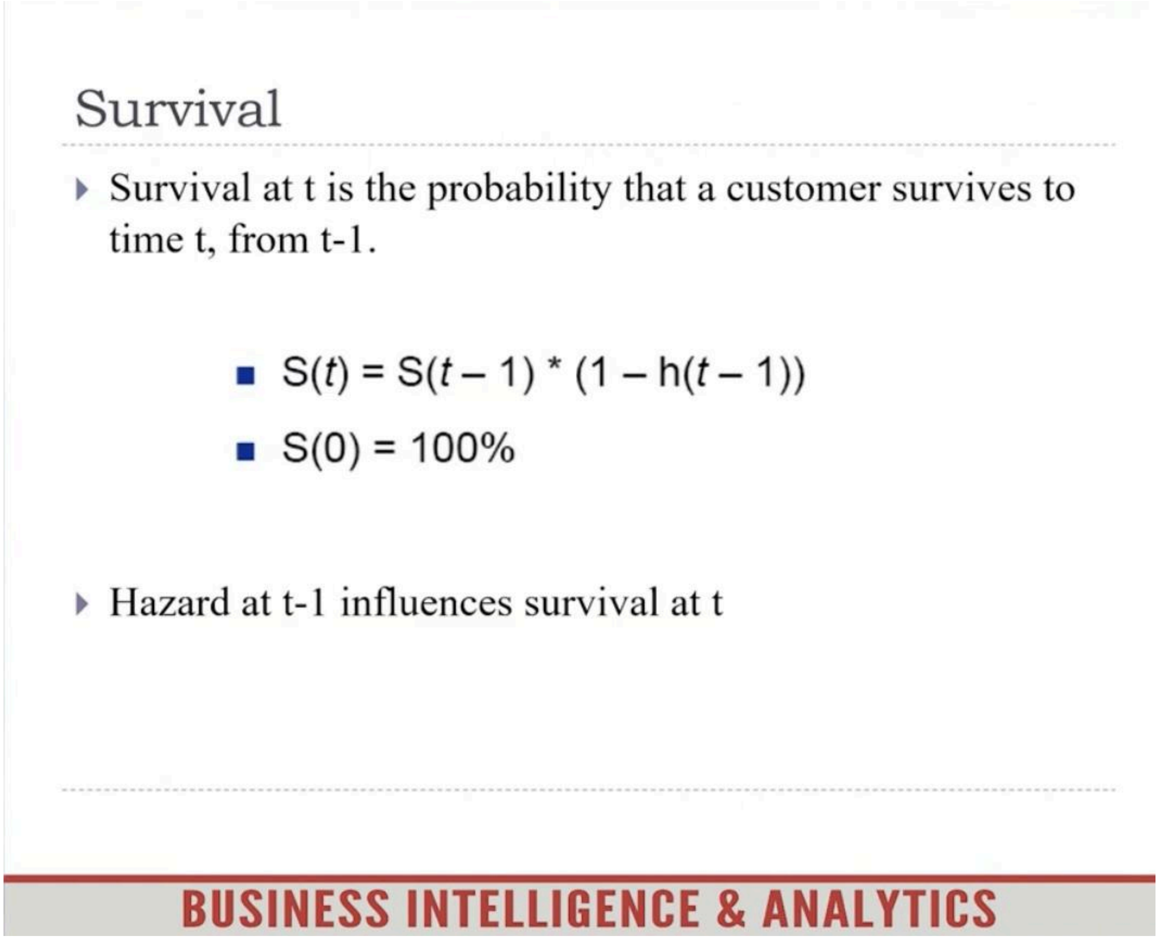
And you, if you lose a customer, you know that actually affects the survival of the customer. That is what we are going to analyze. So, there is another concept called population at risk. Population at risk is the number of customers who could have stopped between t and t + 1. So actually, a given number of customers defect, but that is, that given number is not the only number who could have defected.

More customers could have defected, that did not happen. And therefore, we calculate defection, sorry, hazard at point in time t as a ratio, not as a absolute count or a number, but it is a ratio. And what is that ratio? Hazard at t is defined as number of customers who stop at exactly time t. They stop at time t, as I said, they stop at discrete intervals. They stop at t, that is why they are not there at t + 1.

That is the numerator, we have not given any symbol to there, but it is a textual description of the variable. So, percent, sorry, not percent, number of customers who stop at exactly at t divided by number of customers at risk of stopping at t, who could have stopped, that is a population at risk. Suppose n is the population, you all want symbols, n is the total population who could have stopped or who survived up till point t. And if it is n who actually do not survive up till t + 1, then hazard is n/N

. And the time is here in the tenure scale.

And I hope this is not difficult for you to understand. So, customer hazard is the number of customers who stop at time t, divided by number of customers who could have stopped, the total number of customers who could have stopped. So I am going to, I am going to further expand this and also give you a sense of how you can analyze customer transaction data or a transaction table typically available in a database to arrive at measures like hazard and survival. That is our next step.



Survival

- Survival at t is the probability that a customer survives to time t, from t-1.
  - $S(t) = S(t-1) * (1 - h(t-1))$
  - $S(0) = 100\%$
- Hazard at t-1 influences survival at t

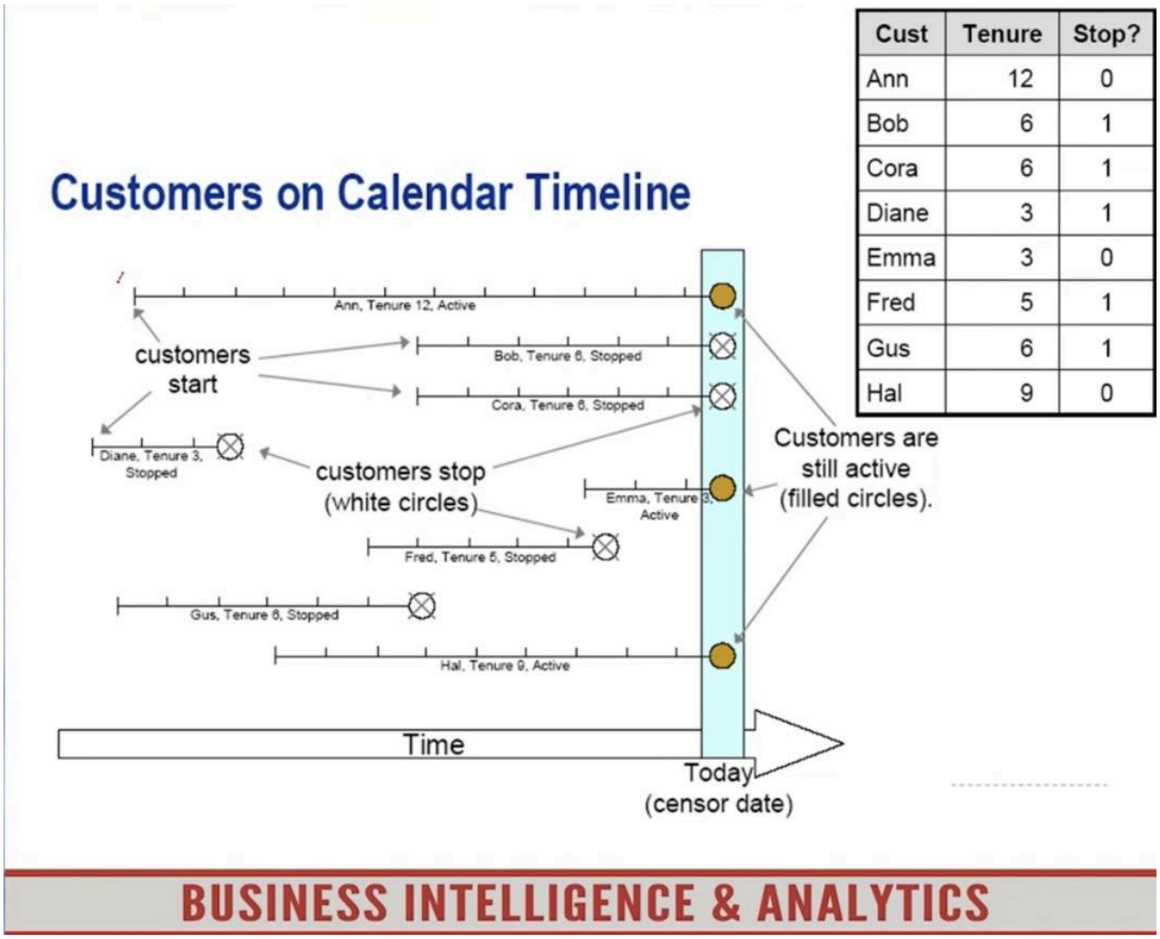
BUSINESS INTELLIGENCE & ANALYTICS

Now, hazard is defined in a particular way and we saw how it is defined. It is defined based on t to t + 1. So, let us again draw those timelines. 0 is somewhere here and you have tenures, you have tenures and continuity here and it has reached some other point in time. And we say this is t -1, this is t, this is t + 1. Now, survival at t is the probability that a customer survives to time t from t - 1.

So, survival is defined with respect to the previous instance or the previous tenure .Survival is at t is the probability that a customer survives to time t. Now, how this survival t can be estimated? Survival t at t is s (t - 1), s(t - 1) means what is the survival of customers at the previous instance that is s(t - 1) × 1 - h(t - 1). So, then another incident happens. One incident here is that customers survived up till t - 1.

There is a set of customers who survived till t - 1. Then between t - 1 and t or at t - 1, this period we did not consider, t - 1 customers defect, customers leave and that is h(t - 1). We saw the formula for h(t - 1) already. Number of customers who defect at t - 1 divided by total number of customers who could have defected. 1 - this is the gain, the, it gives you the opposite of hazard, the opposite of hazard, which is how many customers retained after the defection, after the defection happens.

So, this is a measure of retention. So, survival is the combined probability of survival at t -1 × hazard, 1 - hazard at t - 1. It is the combined probability. P a x Pb, probability a × probability b, P a here is survival at t - 1 and probability b is hazard at t - 1, number of customers who leave to, the number of customers who have left at t - 1. Of course, in order to calculate this, you need a starting point.



Customers on Calendar Timeline

Cust	Tenure	Stop?
Ann	12	0
Bob	6	1
Cora	6	1
Diane	3	1
Emma	3	0
Fred	5	1
Gus	6	1
Hal	9	0

BUSINESS INTELLIGENCE & ANALYTICS

So, survival at 0 or starting point is of course 100 percent. So, we can obviously see that s(t) is a function of s(t - 1). So, survival at the previous instance influences survival at t. So, now you got two concepts related to customer retention or customer churn. One is hazard, other is survival, the formal definitions.

Now if you are given a data set, how would you compute survival and hazard or hazard first and then survival? That is what we are going to see now. Now we have a challenge here because if you look at a transaction data, we have a problem. It is not a problem, you have a challenge. The challenge is that the data that is available is about when did a customer join. You know the start time of a customer, that you can run an SQL query and get the start, the lowest time or the starting point of a customer.

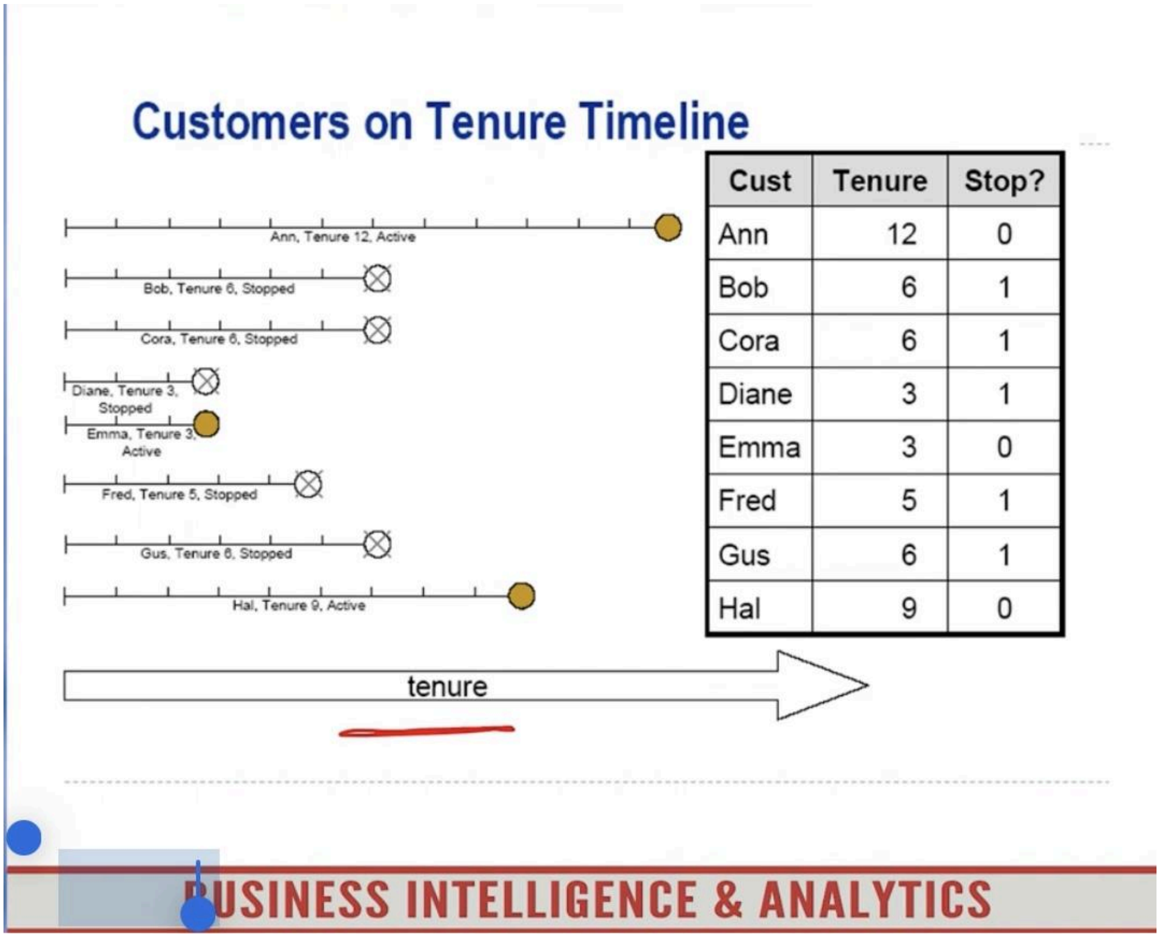
And you can also look at the most recent transaction of the customer and find out how many tenures a customer has actually stayed with the business. So, that is something that you can calculate or directly query and find out. But this can be done for, you know, a group of customers. So here we are taking, say 8 customers as you can see in this table. There are 8 customers, and Bob, Cora, Diane and so on.

And you can count their tenures So, Ann has 12 tenure and the brown circle shows that Ann has 10 tenures, 12 tenures and she is still active. You see, the customer is still active. And Emma is another customer who has how many tenures, 3 tenures, 1, 2, 3 and still active.

And so is Hal is active. But look at some other customers like Bob and Cora. This picture shows that the cross sign shows that they have stopped. They have stopped subscription. This customer, Bob for example, has 6 tenures at the end of 6 tenures, the customer stopped.

There is a stop flag here. Ann has a 0 stop flag meaning the customer is still active, still alive, still alive. So, this is something that you will observe in your data that customers, their number of tenures and whether they are active or terminate, active or stopped, not terminated. Termination is a different term. We talked about it already. And you can see that this is calendar time.

What you see here is the calendar time. We are actually going with our regular flow of time, which is a calendar time. But we have a challenge if you use the calendar time to calculate the hazard and survival. Because what is interesting to us is not the absolute calendar months or calendar weeks or calendar years. What is more interesting to us is the number of tenures, irrespective of when the customer stopped, started.



Customers on Tenure Timeline

Cust	Tenure	Stop?
Ann	12	0
Bob	6	1
Cora	6	1
Diane	3	1
Emma	3	0
Fred	5	1
Gus	6	1
Hal	9	0

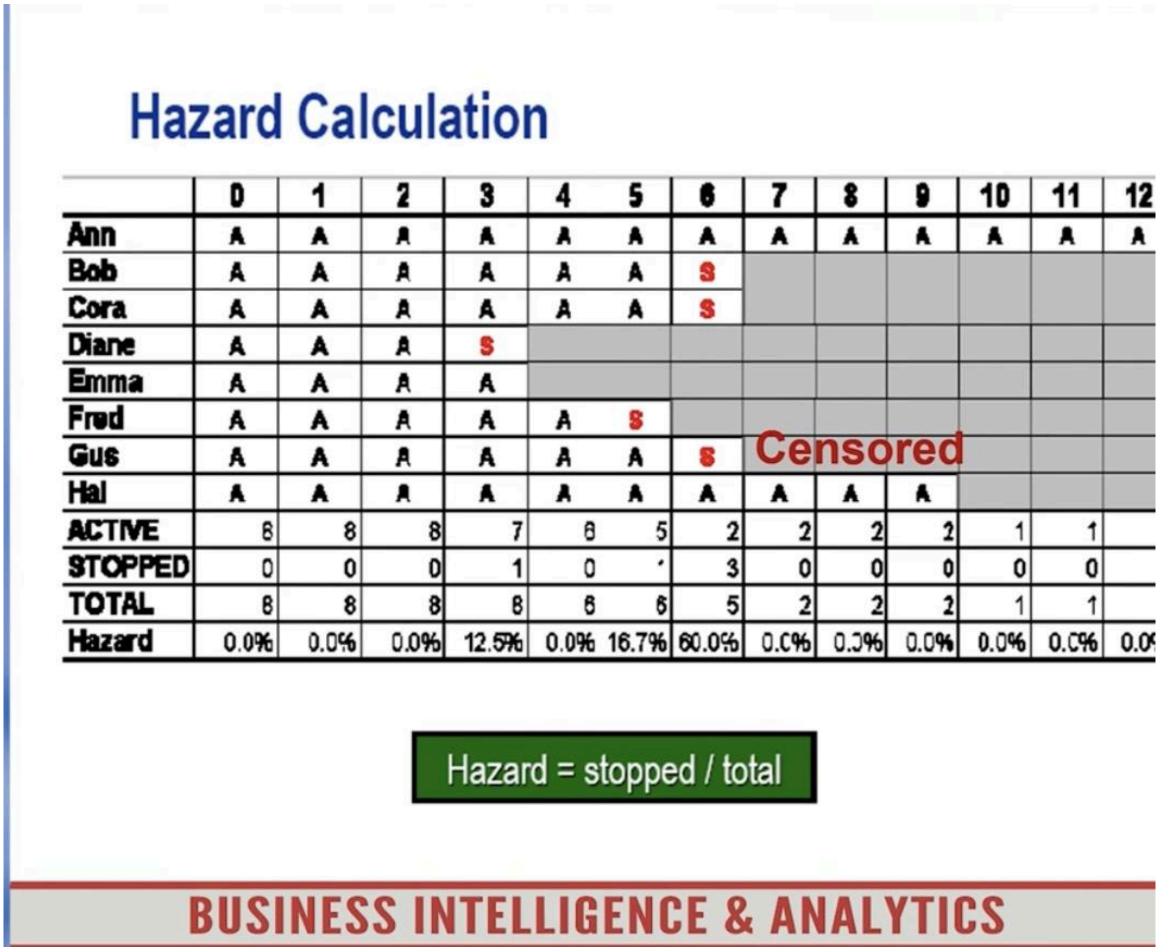
It is the tenure months, not just the calendar months that interest us. We just count the number of tenures and that is what is interesting in a churn analysis. And therefore, we have to change this calendar timeline to tenure timeline. You see, the x axis has changed in terms of the time measure that we use. Time is measured here in tenure counts, not in calendar tenures or not in the measure of absolute calendar times.

But we are only concerned about number of tenures. That is another way of putting it. When we do that, our starting time is not an year or a particular month in a calendar, but we are only, all customers are starting at the same time. That is because we are starting from number of tenures. Number of tenures could be 0, 1, 2, 3, 4, etc.

So, this is of course the starting point, we can call it 0, 1, 2, 3, 4, 5, 6 etc. Of course, this is 1 tenure, this is 2 tenure, this is 3 tenure, you can understand that easily. This is in terms of the count. So, maybe that is precisely what we are going to use instead of the graduations there. So, here the data becomes independent of the calendar timeline, but we are going to use the tenure timeline.

And therefore, the data is reconfigured or the graph is reconfigured here in that way. And look at, let us look at a customer like Emma. Emma has, is still active and she has 1 tenure, 2 tenure, 1, 2, 3, she has 3 tenures. And she has not stopped.

Emma has not stopped. So, she at point in time t, she is active and she is moving forward. Just keep that in mind because this is useful to us when we look at the calculation of hazard and survival. Whereas a customer like Fred is at t th, at a particular tenure , which is 5 tenures, but is not going to the next tenure. And that is important to note here.



Hazard Calculation

	0	1	2	3	4	5	6	7	8	9	10	11	12
Ann	A	A	A	A	A	A	A	A	A	A	A	A	A
Bob	A	A	A	A	A	A	S						
Cora	A	A	A	A	A	A	S						
Diane	A	A	A	S									
Emma	A	A	A	A									
Fred	A	A	A	A	A	S							
Gus	A	A	A	A	A	A	S	Censored					
Hal	A	A	A	A	A	A	A	A	A	A			
ACTIVE	8	8	8	7	6	5	2	2	2	2	1	1	
STOPPED	0	0	0	1	0	1	3	0	0	0	0	0	
TOTAL	8	8	8	8	6	6	5	2	2	2	1	1	
Hazard	0.0%	0.0%	0.0%	12.5%	0.0%	16.7%	60.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0

Hazard = stopped / total

BUSINESS INTELLIGENCE & ANALYTICS

All right, keep this in mind. Let us move on. Now, this sheet provides a basis for calculating hazards. Hazard is the first thing that we are going to calculate. And we have seen the formula also in simple terms, the formula is reproduced here. It is number of customers who stop at t, divided by number of customers who could have stopped at t.

So, those who stopped do not pass on to the t + 1 and those who do not stop, they move on to t + 1. Now, let us look at actually the references to the previous data. And let us look at each customer here. So, Ann is present in the 0th tenure , 1st tenure , 2nd tenure , 3rd tenure . And when it comes to the 3rd tenure , we can see that Diane, one of the customers has stopped.

What happened to Diane? Let us look at the previous chart. Diane is here. She had 1, 2, 3, 3 tenure . She had 3 tenure and she is, she stopped business with the, in terms of calendar time we have already seen, they have stopped except for 3 persons, they are not active, others have all stopped. So, Diane had 3 tenures and she is no more with the business.

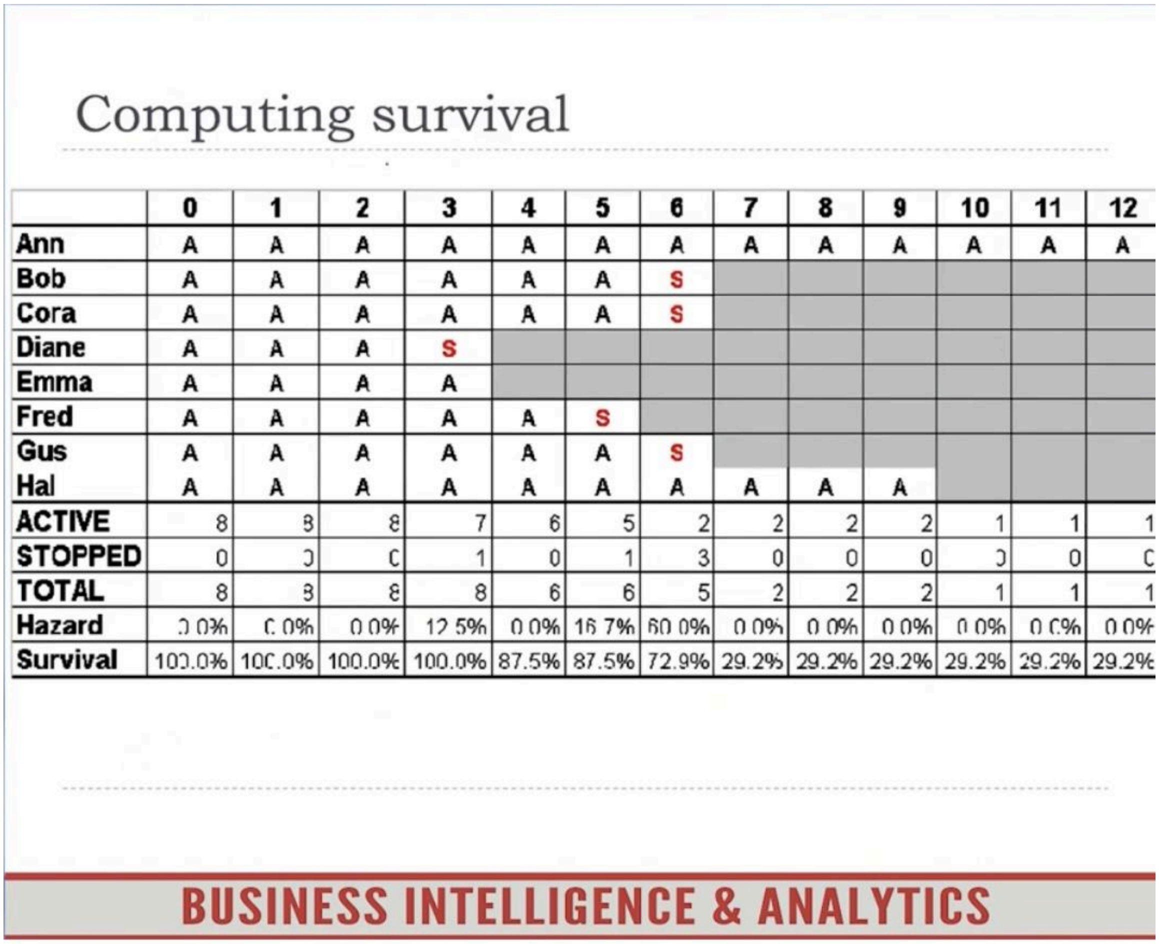
And therefore, there at the end of the 3rd tenure , this is the 3rd tenure , 0, 1, 2. Suppose, let us count, if you count this as t. Now, at this point, at this t, there is a defection or there is a churn. And how do you actually calculate defection at t or hazard at t? Stopped divided by total. So, what is the number of customers who stopped at t? This is, this is just 1, 1 divided by, what is the number of customers who could have stopped? Who could have stopped is the total number of customers who are there.

So, that is 1, 2, 3, 4, 5, 6, 7, 8. 8 customers who could have stopped, but only 1 stopped. And that is hazard at time t. So, this is hazard here. So, a hazard happened at t, affects the survival in t, - t.

So, this is t, this is t + 1. Or let us for the, for just retaining the previous notations, let us call this as t - 1. And let us call this t. So, something happened at t - 1, which is a hazard. And this is t and let us call this as t + 1. So, what happened at t - 1 affects the survival in t.

So, that is what you see. That is what we are going to see in the next graph, next chart. So, similarly, you can see the number, the number of customers keep reducing as you increase the number of tenures. Look at this, after Diane has stopped, we are censoring the data, we are not using Diane's data. So, that affects the number of customers who are present. 7 became 6 here, because that is censored and here one customer Fred stopped and therefore, number of customers has become 5, because Fred has stopped and that is censored.

So, you can see that a censoring has to happen, once a customer has stopped in the calculation of hazard. And this is pretty intuitive and it is just the application of the formula. But in order to calculate survival, as I said, we take this instance as t - 1, we take this as t and we take this as t + 1. And we know that an event in at t point 1 affects the survival at t. So, you can see that here, a survival at t - 1 is 100 percent, 100 percent of the customers are present at t - 1.



Computing survival

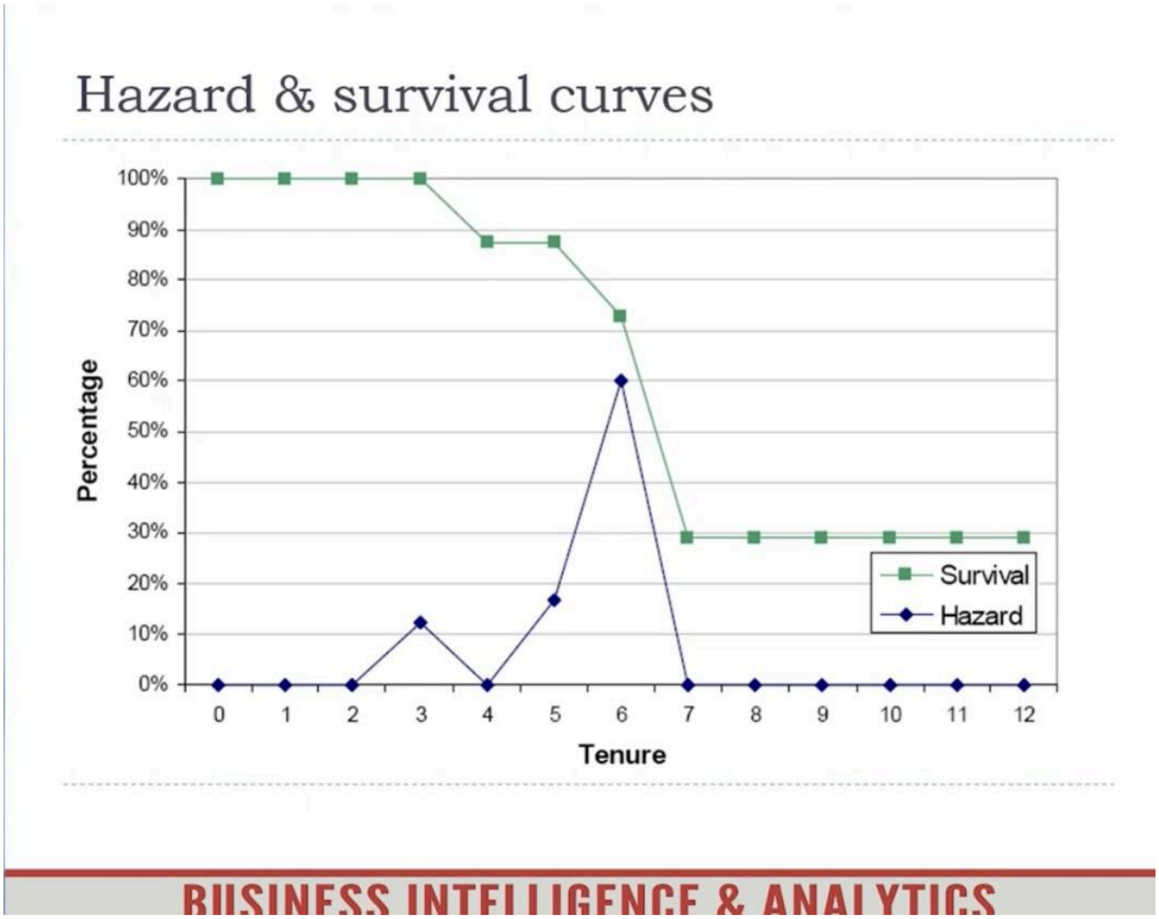
	0	1	2	3	4	5	6	7	8	9	10	11	12
Ann	A	A	A	A	A	A	A	A	A	A	A	A	A
Bob	A	A	A	A	A	A	S						
Cora	A	A	A	A	A	A	S						
Diane	A	A	A	S									
Emma	A	A	A	A									
Fred	A	A	A	A	A	S							
Gus	A	A	A	A	A	A	S						
Hal	A	A	A	A	A	A	A	A	A	A			
ACTIVE	8	8	8	7	6	5	2	2	2	2	1	1	1
STOPPED	0	0	0	1	0	1	3	0	0	0	0	0	0
TOTAL	8	8	8	8	6	6	5	2	2	2	1	1	1
Hazard	0.0%	0.0%	0.0%	12.5%	0.0%	16.7%	60.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Survival	100.0%	100.0%	100.0%	100.0%	87.5%	87.5%	72.9%	29.2%	29.2%	29.2%	29.2%	29.2%	29.2%

BUSINESS INTELLIGENCE & ANALYTICS

When they reach t - 1, all the customers are present, 100 percent. But here a defection is happening, someone is leaving, that is Diane. And therefore, the survival at t is affected by what happens in t - 1. That is the idea of survival probability. So, this is st and we know that st is s(t - 1) × 1 - h(t - 1).

And what is that? st - 1 is 100 percent here. And 1 - h(t - 1) is 1 - 0.125, is not it? And that is what is shown as 87.5 percent or 0.875 in terms of probability. And that is the calculation that you do in order to compute survival.

So, hazard happens in the previous instance, survival is affected in the next instance. And this calculation can continue for all tenures. And you can see the idea of censoring here. And you can notice that survival is something that is dropping from 100 percent to 29.2 percent. But you can also see that churn is, churn is not a downward sloping curve, it is rather these are spikes. These are spikes that happen, these are events that happen in different instances. And that is the way to understand survival and churn. And now let us set up graphs for this. Here you can see hazard and survival curves, the same data that we computed in the previous slide is plotted here.



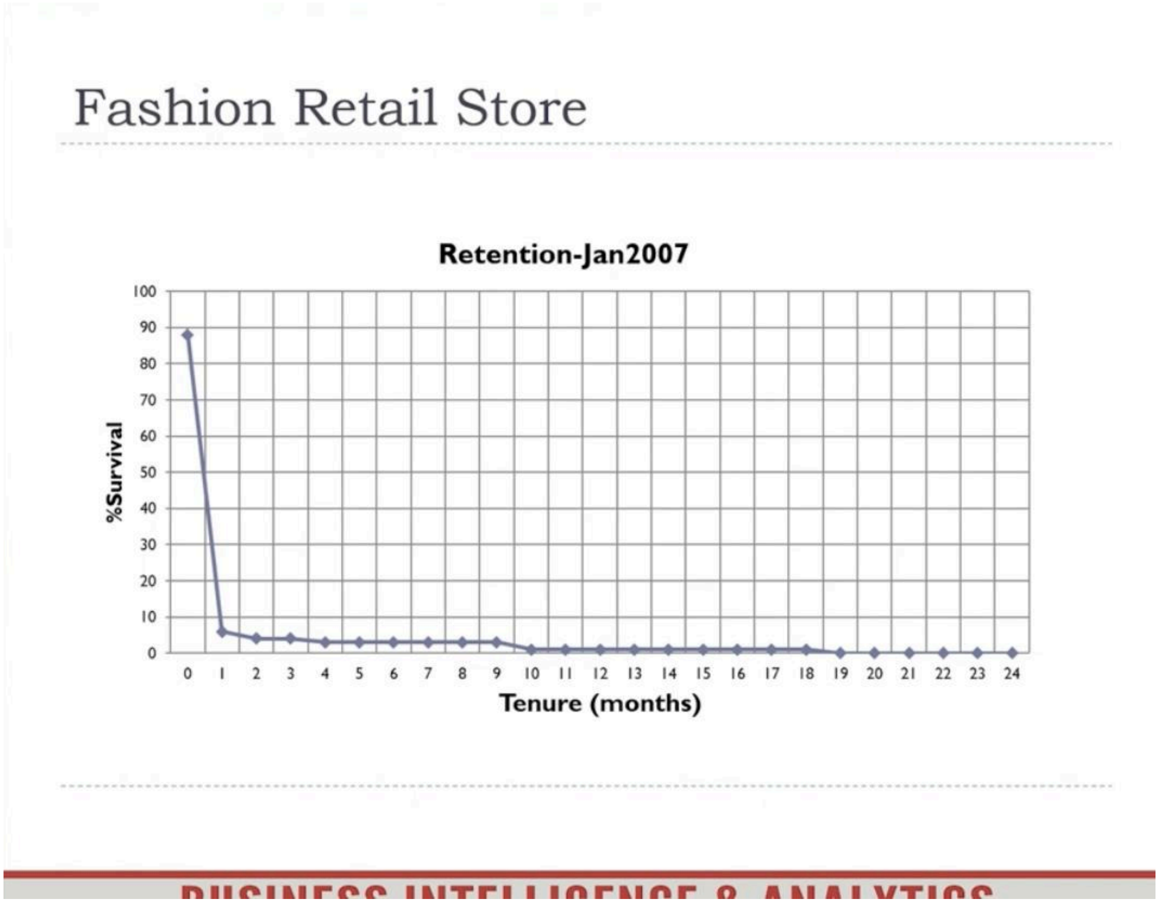
Hazard & survival curves

BUSINESS INTELLIGENCE & ANALYTICS

So, of course, the x axis is tenure and y axis is percentage or probability. Now, you can see that up till the fourth, sorry, up till the third tenure , the survival is 100 percent. And you can see that at this third tenure , there is a defection. And this particular defection affects the survival in the next instance.

So, the survival keeps falling based on churn or based on hazard. And you also notice that, you know, there is no churn after some time. And therefore, the survival actually becomes almost constant at 30 percent. These are what? Who are those customers? They are the real loyal customers. These are customers who actually stay on with you. The 30 percent, this 30 percent is a very important group of customers for this business.

And you also as a manager, if you look at or eyeball this data, you see that well, there is 60 percent defection or a major incident that is happening at the sixth tenure . Well, this should be of interest to a manager. Why are many people leaving after six tenures? Is that a contract period? What needs to be done so that this defection can be reduced? You can see this is a lot of actionable insight. One has major spikes that is happening and at how many tenures that is happening and who are the group of loyal customers, the sort of 30 percent and a company's interest should be to boost this numbers etc. So, hazard and survival curves are insightful information as far as business is concerned, because you get to know customers' behavior in a very insightful or informative way, in an actionable way, when you look at this graphs.

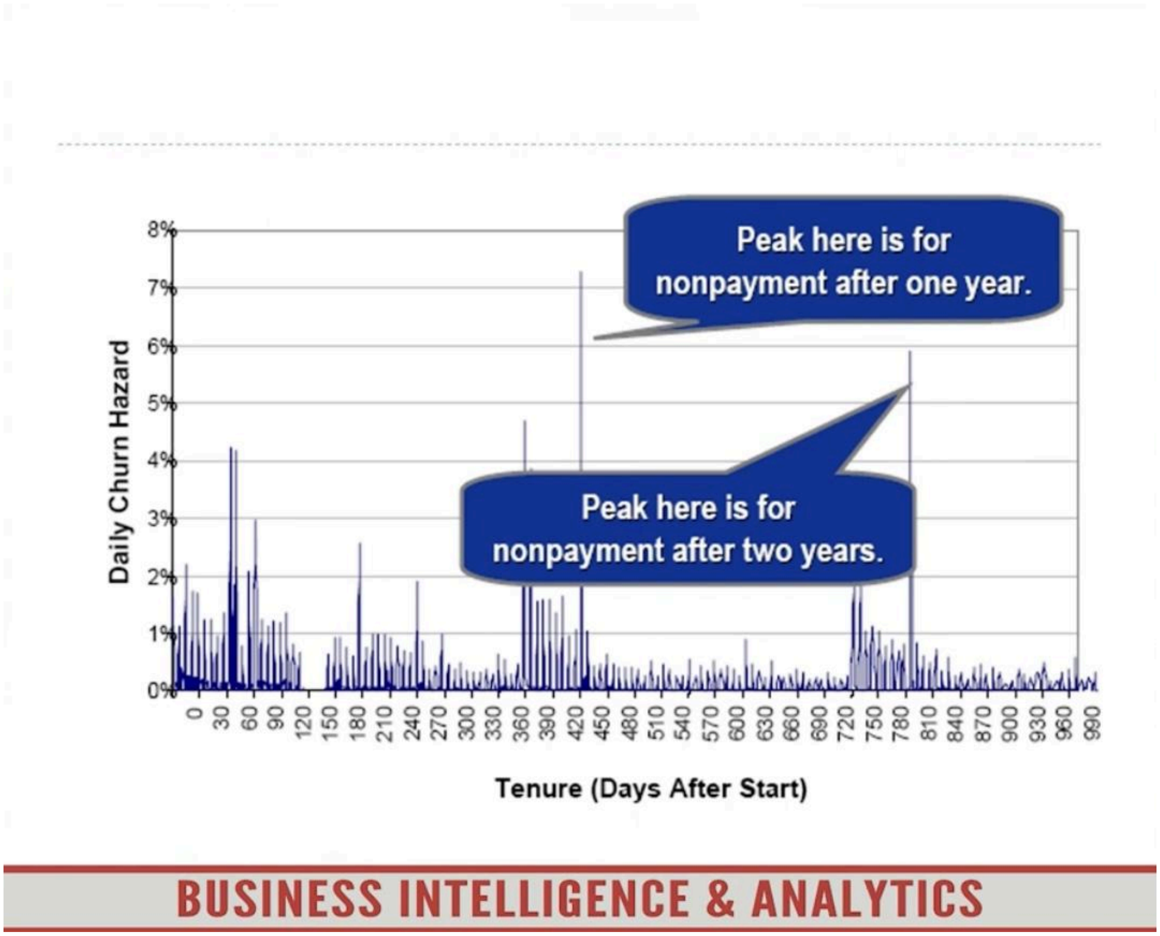


Fashion Retail Store

BUSINESS INTELLIGENCE & ANALYTICS

 Here I am showing a result from a data analysis, I did for a fashion retail. I will show the more details about this work in a spreadsheet very soon. But you can see that this retail store, since it is a retail store and not a subscription business, there are certain limitations in terms of understanding start time   and stop time. But within those limitations, you can see the behavior of customers as sort of, very different from the graph which we just noticed. You can see that here the graph became steady at 30 percent and here unfortunately it is not a 30 percent, it is a much smaller percent.

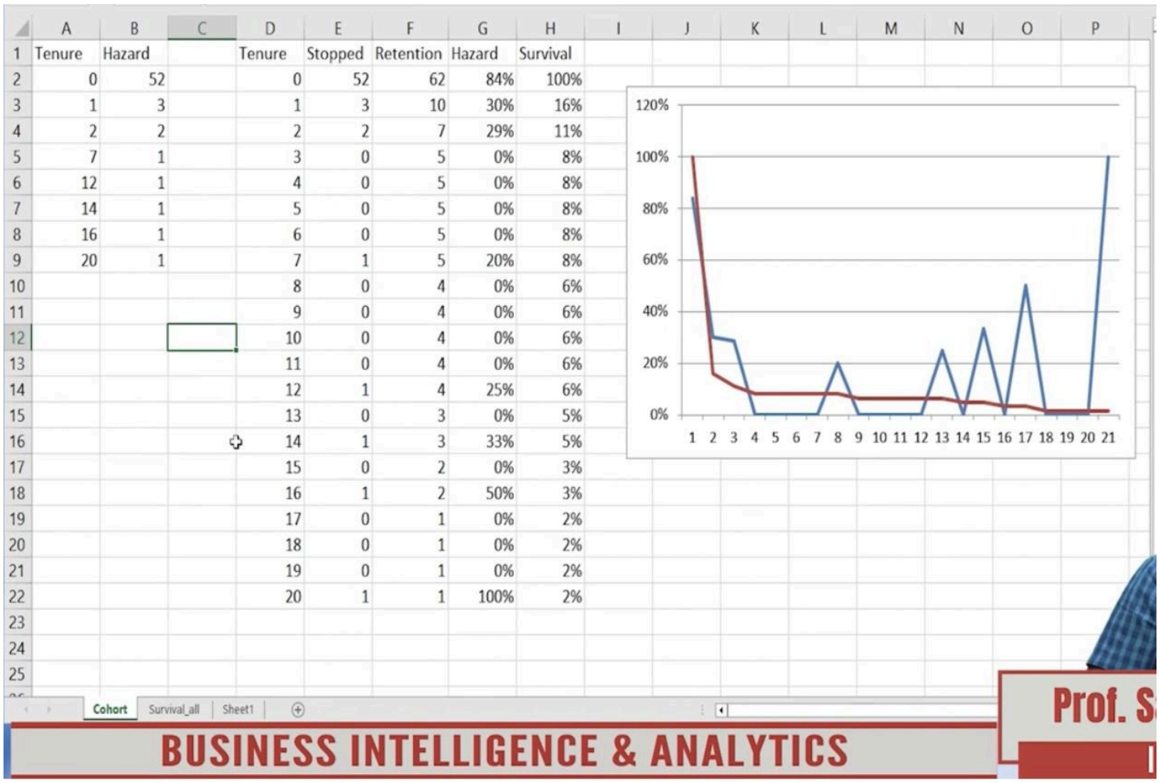
  And well, that is an insight when you compare the survival graphs. And of course, this is plotted for a year and the tenure is in months and not in years as we saw in the previous one. And whether you use tenure in months or years also depends on the type of product or type of service. For automobiles you may use tenure as years, whereas for a magazine subscription, you may actually or a journal subscription or something like that you may look at a monthly contract. So, this could differ. And this is just further illustration of the concepts that we discussed how survival and churn, here it is titled as churn hazard, same as hazard.



BUSINESS INTELLIGENCE & ANALYTICS

How when you plot hazard can be informative to a manager, you know you see the spikes. These spikes are useful for managers to look at customer behavior. Why is there a spike at this number of tenures and why is there a spike here and why is it steady here. So, we say this is very useful in diagnostics. So, once you know what is going on or how, what is going on in terms of customers behavior, then you can further investigate through other means as to why customers are leaving at a given, after a given number of tenures.

There could be other you know, other environmental factors that could be affecting the customers or it could be contracts or it could be quality of service, it could be quality of products, it could be competition. There could be several reasons there, but it is for a manager to analyze this and take actions accordingly. So, now in order to illustrate this further, I would use a spreadsheet. Now, let us continue to look at real life data. As I indicated, this is data that I derived from, a transaction data that pertains to a retail business, a fashion retail store, which sells fashion products in one store and they actually started investing in customer loyalty.



Tenure	Hazard		Tenure	Stopped	Retention	Hazard	Survival
0	52		0	52	62	84%	100%
1	3		1	3	10	30%	16%
2	2		2	2	7	29%	11%
7	1		3	0	5	0%	8%
12	1		4	0	5	0%	8%
14	1		5	0	5	0%	8%
16	1		6	0	5	0%	8%
20	1		7	1	5	20%	8%
			8	0	4	0%	6%
			9	0	4	0%	6%
			10	0	4	0%	6%
			11	0	4	0%	6%
			12	1	4	25%	6%
			13	0	3	0%	5%
			14	1	3	33%	5%
			15	0	2	0%	3%
			16	1	2	50%	3%
			17	0	1	0%	2%
			18	0	1	0%	2%
			19	0	1	0%	2%
			20	1	1	100%	2%

So, what this company did was, that they started issuing loyalty cards and of course with a loyalty program and they offered special discounts for customers who subscribe or enroll for their loyalty program and then at every purchase of course, you know you have to date it back to, say ten years back. So, instead of asking your mobile number today, they will ask for your loyalty card. So, every purchase gets accounted for a loyal customer. So, essentially the idea was to sort of retain customers or you know by investing in customer loyalty program. So, they started a loyalty program and then based on that information, they started actually tracking customers and this is the data, the raw data and the actual data analysis for calculating hazards or number of, here hazard only means number of customers who left at a given point in time or at a given tenure.

Zero tenure is the first month. So, we just used SQL queries to count the number of customers who abstended after the first purchase and we found that there were 52 customers who subscribed or who joined the loyalty program, but never turned up after the first purchase that is 52. So, we counted this as hazard as 52 in the zero tenure. When it comes to the next tenure, next month, three customers stopped. After second tenure, two customers stopped and between two and seven, no customers stopped.

So, we are actually, this is SQL query data. So, seventh tenure you see one customer and 12, 14, 16 and 20, one customer each defecting. So, this is result from an SQL query. And now, our task is to plot hazard and survival graphs from this data because simply looking at this data is not very insightful, but if you visualize it, it becomes more insightful. So, that is the purpose. So, the simple thing that we did was to convert this data by including all tenures so that you know your x axis becomes continuous. So, we have all 0 to 20th and filled in 0 wherever required, 0 defections. So, it is the same data, but expanded for all graduations. And then we started looking at survival, this is same as survival in terms of count, I have titled it as retention, but essentially showing how many customers were there at this beginning. And that is nothing but the total of all this. If you count all the customers together, it actually is 62. So, when you started, they were in the first tenure or 0 tenure, they were 62 customers, out of which 52 left.

And therefore, how many survived to the next instance, it is 10 customers, 62 - 52, you can see I have just put that formula there. If 62 - 52. So, again, 3 customers left. So, therefore, 7 is retained.

Now, and here 2 left and therefore, 5 is retained and that 5 continues. So, in this way, we calculated the absolute counts of survival from this 0th tenure till the 20th tenure. And now we started calculating the hazard. So, since 52 customers out of 62 left in the first tenure itself, hazard at the first tenure is 52 divided by 62, E2/ F2, we can see the formula here, that is actually 84, 84%, 0.84 or 84%. That is a big spike, 84% of the customers left after the first tenure. And then we just apply the hazard formula. So, how many are left actually 10. And 3 out of 10, again leave in the second tenure and therefore, that becomes 30% and so on. So, we calculate the hazard spikes, as we saw in the slides in this particular column, the hazard percentage is calculated.

And once we know the hazard, we can calculate the survival. So, survival as a percentage is calculated here. So, you know already that in the first tenure, survival will always be 100%. Survival is going to reduce if there is a hazard in the previous instance.

So, now we know that 84% is left and we know the formula 1 - 84 1, that is 0.16 or 16%. And here 11% is this, 16% or 0.16 x 1 - 0.3, that is 11%. We know the formula. So in that way, we continue to calculate the survival for the subsequent tenures and we complete this column, which is the survival.

So, we have the survival in percentage, we have the hazard in percentage. And therefore, that is plotted here. The red curve denotes survival and the blue curve denotes hazard. Now, this is interesting graph. Number 1, if you are looking at a group of customers and over a period of time, over certain tenures, which is 20 tenures, the interesting facts that you can observe here is number 1, the survival after the first tenure is just 16%, meaning 84% of the customers left in the first tenure itself. What is the message what is the message to the management for action? You expect you started a loyalty program, but 84% of the customers left after the first tenure or 84% of the customers, in other words, agreed to subscribe to loyalty program.

But, of course, 100% joined, but 84% of them joined the loyalty program, but made only one purchase, meaning that well, you know, it is something that is freely given, they subscribe, but they did not care about the loyalty program. Or in other words, the loyalty program is run not very thoughtfully, it does not have any impact in retaining customers. Just by giving plastic cards, you do not, you should not expect that customers will stay with you , just because you gave a card. So, loyalty program should be well designed, otherwise there is no point in giving loyalty cards.

That is a very important message by looking at this graph. And then you also see, well, the, so the spike is really in the first instance or the first tenure showing that people do not come back. And those who stayed on, they stayed on for some time and you can see that towards the 17th tenure, there is a big spike. You know, of course, it is for the store to find out why it is so. And unfortunately, you can see that most customers are actually leaving.

So, you can see 100%, hazardous 100%, all customers have left at the end of 20 months. So, the loyalty program of this business is a complete disaster. Of course, I am not suggesting that the company is a disaster, but I am suggesting that the loyalty program, of course, it is in an infant stage and it is not well thought through.

And although it is a small example, it has a couple of takeaways. Number one, in terms of how you can use raw data, which is nothing but customer transaction data from a database table and run certain SQL queries and calculate hazards or number of defections per tenure and then organize that data such that you can plot the survival and hazard and visualize how customers are behaving. And it obviously brings the picture that customers do not survive and they quit after they all close their business after 20 months.

So, or 20 tenures . So, this is actionable information as far as business is concerned. It is very insightful as far as business is concerned. So, this is the first part or the first aspect of survival analysis or churn modeling. And subsequently, we would very soon look at the same concept of survival and we see how survival as a probability can be factored into a bigger model, which is customer lifetime value analysis. So, we will now discuss in the next module, CLV or customer lifetime value analysis where survival becomes a part of the model. So, survival analysis in itself is insightful to understand customer behavior, their tenures, their churns etc.

And subsequently, that information which is derived from the data can be used in another model for customer loyalty programs or deciding customer segmentation and customer loyalty programs and so on. That will be our next discussion. Thank you very much.