Introduction to Conditional Probability for Data Science Professionals

As the name suggests, Conditional Probability is the probability of an event under some given condition. And based on the condition our sample space reduces to the conditional element.

For example, find the probability of a person subscribing for the insurance given that he has opted for the house loan. Here sample space is restricted to the persons who have taken house loan.

To understand Conditional probability, it is recommended to have an understanding of probability basics like Mutually Exclusive and Independent Events, Joint, Union and Marginal Probabilities and Probability vs Statistics etc. In case you want to revise those concepts, you can refer those here Probability Basics for Data Science.

Wiki Definition:

In probability theory, conditional probability is a measure of the probability of an event occurring given that another event has occurred. If the event of interest is A and the event B is known or assumed to have occurred, "the conditional probability of A given B", or "the probability of A under the condition B", is usually written as P(A | B), or sometimes PB(A) or P(A / B)

Now the question may come like why use conditional probability and what is its significance in Data Science?

Let's take a real-life example. Probability of selling a TV on a given normal day may be only 30%. But if we consider that given day is Diwali, then there are much more chances of selling a TV. The conditional Probability of selling a TV on a day given that Day is Diwali might be 70%. Which can be written as P(TV sell on a random day) = 30%. P(TV sell given that today is Diwali) = 70%.

So Conditional Probability helps Data Scientists to get better results from the given data set and for Machine Learning Engineer helps in building more accurate models for predictions.

Let's deep dive into it more:

Following table contains the different age group peoples who have defaulted and not defaulted on Loans.

		Age			
		Young	Middle-Aged	Senior Citizens	Total
Loan Default	No	**10503**	27368	259	38130
	Yes	3,586	4,851	**120**	8557
	Total	14089	32219	379	46687

Table - 1

Converting the above table into probabilities

		Age			
		Young	Middle-Aged	Senior Citizens	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0003	0.184
	Total	0.302	0.690	0.008	1.000

Table - 2

So if we have tabular data then in case of conditional probabilities sample space get reduced to either the full column or a complete row and rest of the sample space becomes irrelevant.

What is the probability that a person will not default on the loan given he/she is middle-aged?

P(No | Middle-Aged) = 0.586/0.690 = 0.85 [referring table – 2, probability form data]

P(No|Middle Aged) = 27368/32219 = 0.85 [referring table -1, normal numbered data]

If you notice, it is very clear that in the numerator it is the Joint Probability that is the Probability of a person not defaulting on the loan and also the person is middle-aged.

And in the denominator, it is the Marginal probability that is the Probability of a Person being middle-aged.

Hence we can also define the Conditional probability as the ratio of Joint probability to the Marginal probability.

$$P(A|B) = P(A \text{ and } B)/P(B)$$

Again let's ask the question little differently by changing the order as below.

What is the probability that a person is middle-aged given he/she has not defaulted on the loan?

Now see, sample space has changed to the colored row that is the not defaulters row.

		Age			
		Young	Middle-Aged	Senior Citizens	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0003	0.184
	Total	0.302	0.690	0.008	1.000

Table - 3

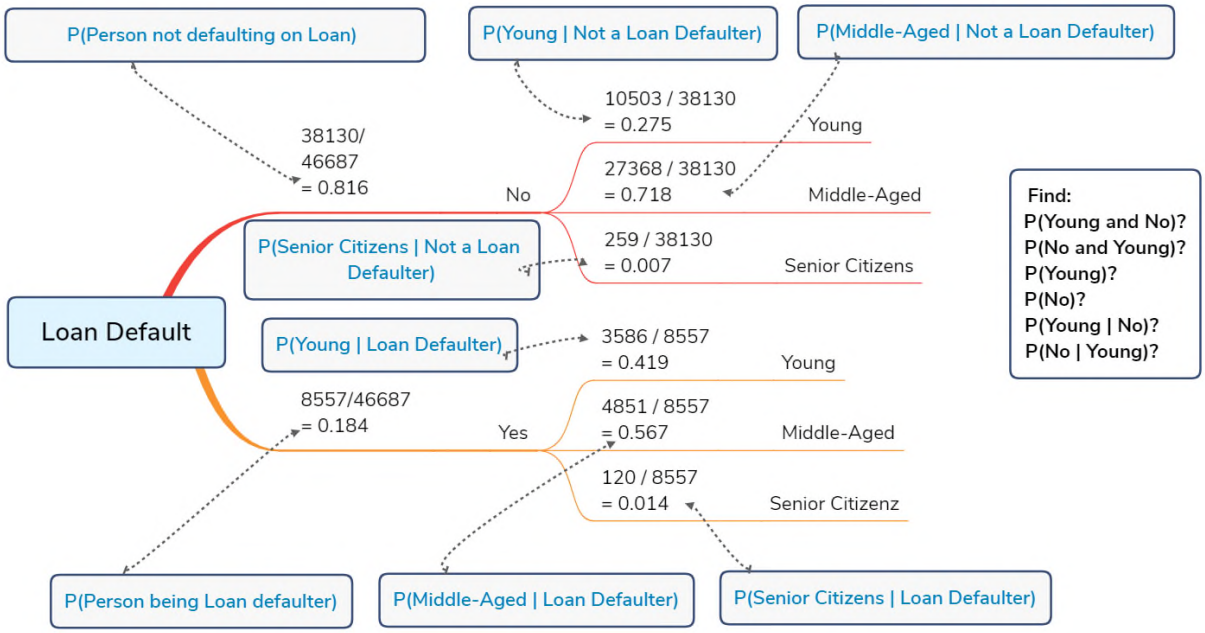
P(Middle-Aged | No) = 0.586/0.816 = 0.72 (Order Matters)

Now did you notice something again, probability is changed by changing the order of the events.

Hence in Conditional probability order matters.

Conditional Probability Visualization using Probability Tree

		Age			
		Young	Middle-Aged	Senior Citizens	Total
Loan Default	No	10503	27368	259	38130
	Yes	3,586	4,851	120	8557
	Total	14089	32219	379	46687



www.ashutoshtripathi.com

Explanation:

I have tried to explain each branch logic within the tree itself. Now let's dive into the questions which will explain the importance of probability tree in calculating the conditional probabilities.

P(Young and No)?

- ⇨ Use standard conditional probability formula:
- ⇨ P(Young | No) = P(Young and No)/P(No)
- ⇨ By Probability tree, we know the probability of P(Young | No) = 0.275.
- ⇨ P(Young and No) = P(Young | No) * P(No)

- ⇨ Now see right side all probabilities values are known, hence put them in above equation and we will get the desired probability.
- ⇨ P(Young and No) = 0.275 * 0.816 = 0.2244 = ~0.225

P(No and Young)? (Order is changed)

- ⇨ P(No and Young) = P(Young and No) = 0.225 [same as above]
- ⇨ In Joint probability order does not matter

P(Young)?

- ⇨ Look at all the branches associated with Young (ending with Young) and take Sum of Products of probability values within branch
- ⇨ Which means
- ⇨ P(Young) = 0.816 * 0.275 + 0.184 * 0.419 = 0.301496 = ~ 0.302

P(No)?

- ⇨ P(No) = 0.816 (Directly from the tree)

P(Young | No)?

- ⇨ P(Young | No) = p(Young | Not a loan defaulter) = 0.275 [see the tree]

P(No | Young)? [Order changed]

- ⇨ P(No | Young) = P(Young and No)/P(Young) [we have already calculate right side probabilities in above calculation]
- ⇨ P(No | Young) = 0.225/0.302 = 0.745

Now let's explore the standard conditional probability formula.

From conditional probability we know that

- ⇨ P(A|B) = P(A and B)/P(B)
- ⇨ P(A and B) = P(B) * P(A|B) ----------[1]

Similarly

- ⇨ P(B|A) = P(B and A)/P(A) = P(A and B)/P(A) [I n Joint Probability order does not matter]
- ⇨ P(A and B) = P(A) * P(B|A) ---------[2]

From equation [1] and [2],

- ⇨ P(B) * P(A|B) = P(A) * P(B|A)
- ⇨ **P(A|B) = P(A) * P(B|A) / P(B)**

Now if we want to find the P(No | Young). Then we can use the above derived formula directly. Because P(Young | No) as well as P(Young) values will get from probability tree and putting in above formula will give the result.

Examples 1:

Three persons A, B and C are competing for the post of CEO of a company. The chances of they becoming CEO are 0.2, 0.3 and 0.4 respectively.

The chances of they taking employees beneficial decisions are 0.50, 0.45 and 0.6 respectively

What are the chances of having employee's beneficial decisions after having new CEO?

Solution:

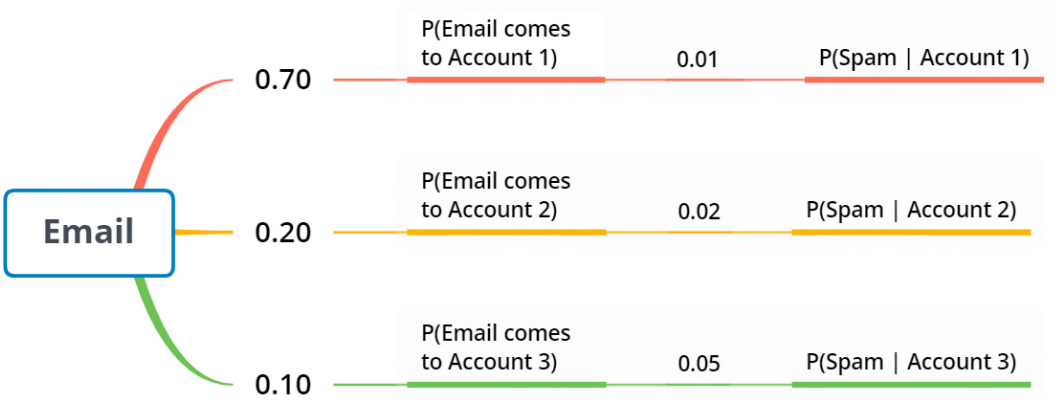
Example 2:

An individual has 3 different email accounts. Most of her messages, in fact 70% come into account #1, whereas 20% come into account#2 and the remaining 10% into account #3.

Of the messages into account #1, only 1% are spam, whereas the corresponding percentages for accounts #2 and #3 are 2% and 5% respectively.

What is the probability that a randomly selected message is a spam is from account#2?

Solution:



P(Account 2 | Spam) = P(Account 2)*P(Spam | Account 2)/P(Spam)
P(Account 2 | Spam) = 0.20*0.02 / (0.70*.01 + 0.20*0.02 + 0.10*0.05)
P(Account 2 | Spam) = 0.25