Expected Value Framework

For Data Science Projects

CS5056 Data Analytics

Francisco J. Cantú, Héctor Ceballos
Tecnológico de Monterrey

April 7, 2021

Februrary-June, 2021

Expected Value

- The expected value method provides a framework that is extremely useful in organizing thinking about data-analytic problems.
- Specifically, it decomposes data-analytic thinking into:

  (i) The structure of the problem

  (ii) The elements of the analysis that can be extracted from the data, and

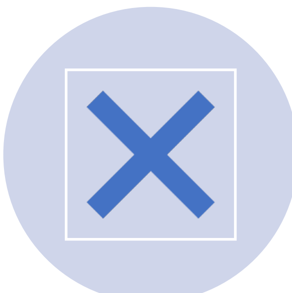
  (iii) The elements of the analysis that need to be acquired from other sources (e.g., business knowledge of subject matter experts)

Expected Value Framework

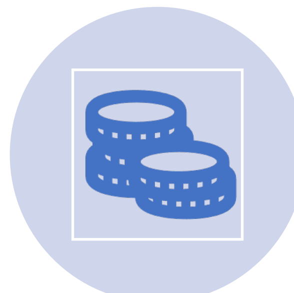
One of the most difficult and most critical parts of implementing data science in business is **quantifying the return-on-investment or ROI**

The Expected Value Framework, is a method that connects the machine learning classification model to ROI

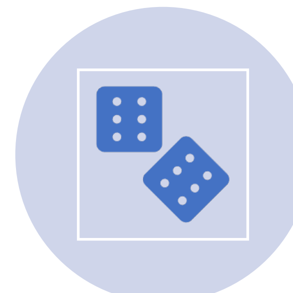
3 REASONS YOU NEED TO LEARN THE EXPECTED VALUE FRAMEWORK



REASON 1: CLASSIFICATION MACHINE LEARNING ALGORITHMS OFTEN MAXIMIZE THE WRONG METRIC



REASON 2: THE SOLUTION IS MAXIMIZING FOR EXPECTED VALUE



REASON 3: EXPECTED VALUE CAN TEST FOR VARIABILITY IN ASSUMPTIONS (ANALYSIS OF SCENARIOS)

https://www.kdnuggets.com/2018/07/data-science-business-expected-value-framework.html

Expected Value

- Suppose that X is a discrete random variable with Probability Mass Function MF p_X (x) and g : R → R is an arbitrary function
- g(X) is a random variable, and we define the expectation or **expected value** of g(X) as:

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x)$$

Expected Value

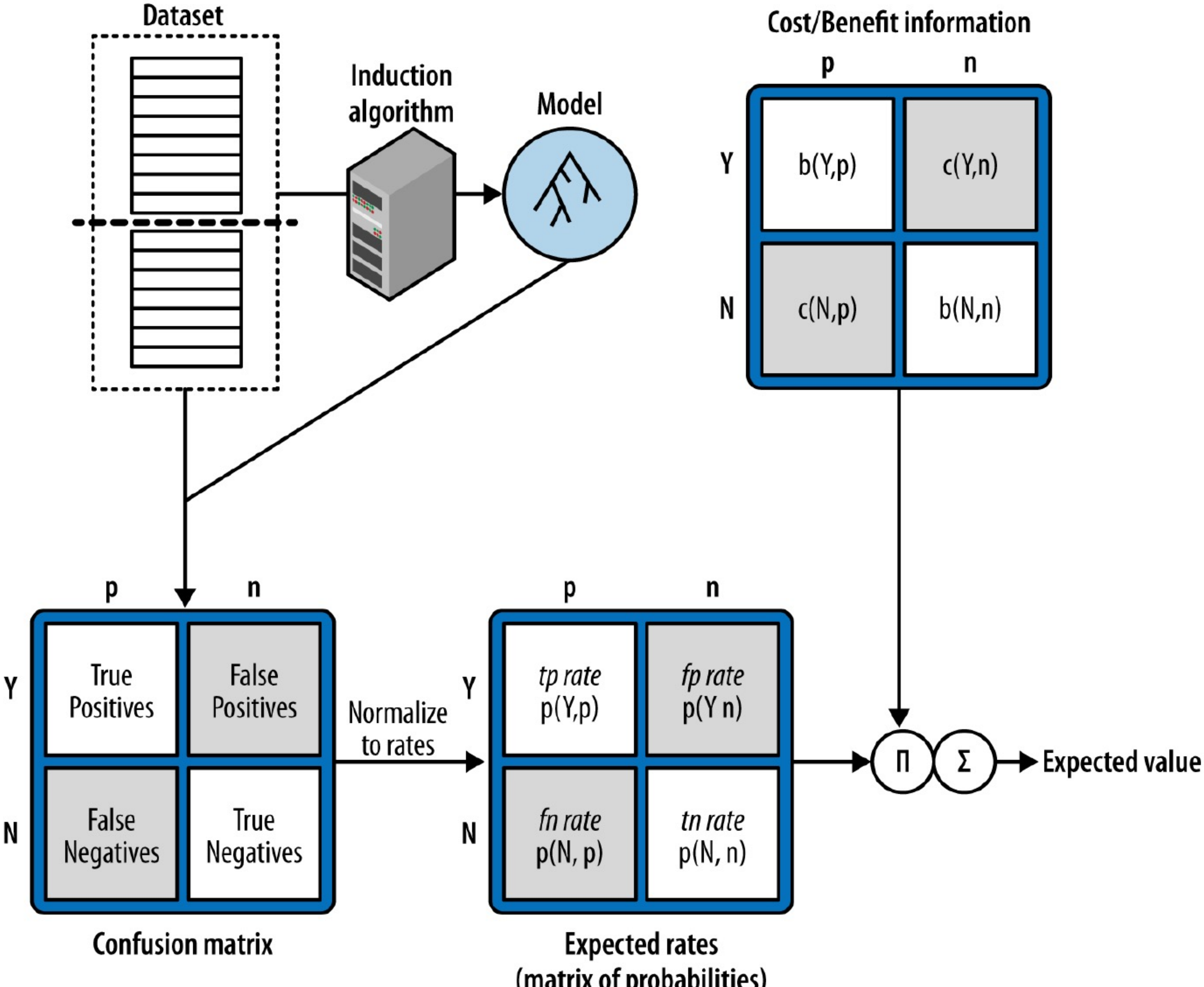
- If X is a continuous random variable with PDF $f_X(x)$, then the **expected value** of g(X) is defined as:

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- Intuitively, the expectation of g(X) can be thought of as a "weighted average" of the values that g(x) can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$
- As a special case of the above, note that the expectation, E[X] of a random variable itself is found by letting g(x) = x; this is also known as the mean of the random variable X.

Expected Value Framework

- Example 1: Provost, F. & Fawcett, T. (2013). *Data Science for Business*, pp 194-202
- Example 2: Explain the "So What?" Behind Machine Learning Models with the Expected Value Framework (Part 2 of 3). https://blogs.oracle.com/ai-and-datascience/post/explain-the-quotso-whatquot-behind-machine-learning-models-with-the-expected-value-framework-part-2-of-3
- Case Study 1: Bancomer
- Case Study 2: TV Azteca
- Case Study 3: Metalsa


Dataset
Induction algorithm
Model
Cost/Benefit information
p
n
Y
N
b(Y,p)
c(Y,n)
c(N,p)
b(N,n)
p
n
Y
True Positives
False Positives
N
False Negatives
True Negatives
Normalize to rates
p
n
Y
tp rate p(Y,p)
fp rate p(Y n)
N
fn rate p(N, p)
tn rate p(N, n)
Π
Σ
Expected value
Confusion matrix
Expected rates (matrix of probabilities)

	p	**n**
Y	56	7
N	5	42

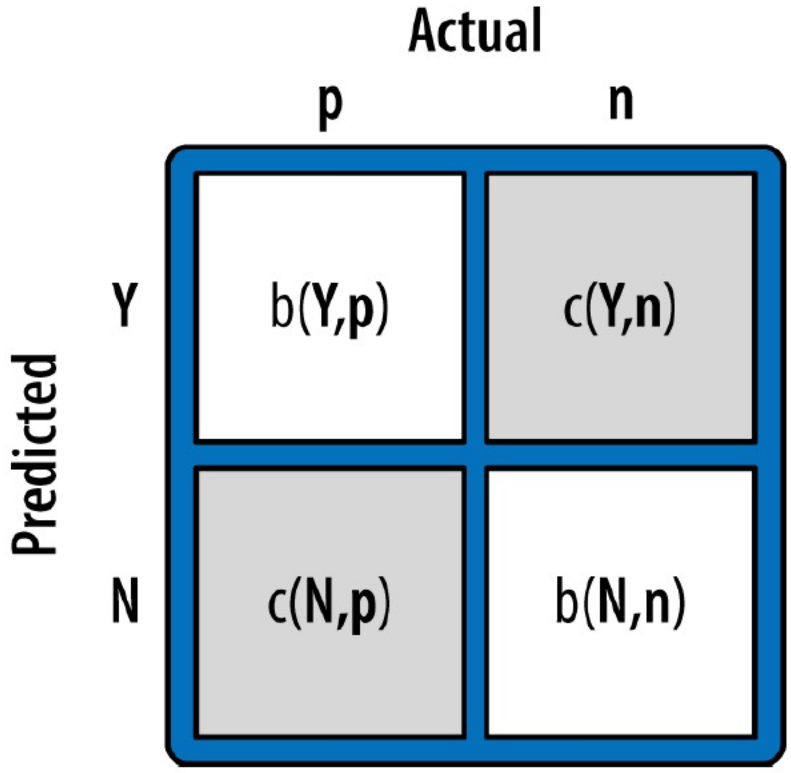
$$p(h\,,\,a) = count(h\,,\,a)\,/\,T$$

T = 110

$p(\mathbf{Y,p}) = 56/110 = 0.51$ $p(\mathbf{Y,n}) = 7/110 = 0.06$

$p(\mathbf{N,p}) = 5/110 = 0.05$ $p(\mathbf{N,n}) = 42/110 = 0.38$

Costs and Benefits



Costs and Benefits

- A false positive occurs when we classify a consumer as a likely responder and therefore target her, but she does not respond. We've said that the cost of preparing and mailing the marketing materials is a fixed cost of \$1 per consumer. The benefit in this case is negative: $b(Y, n) = -1$.

- A false negative is a consumer who was predicted not to be a likely responder (so was not offered the product), but would have bought it if offered. In this case, no money was spent and nothing was gained, so $b(N, p) = 0$.

- A true positive is a consumer who is offered the product and buys it. The benefit in this case is the profit from the revenue (\$200) minus the product-related costs (\$100) and the mailing costs (\$1), so $b(Y, p) = 99$.

- A true negative is a consumer who was not offered a deal and who would not have bought it even if it had been offered. The benefit in this case is zero (no profit but no cost), so $b(N, n) = 0$.

$$
\begin{array}{cc}
 & \text{Actual} \\
\text{Predicted} & \begin{array}{c} \\ \mathbf{Y} \\ \mathbf{N} \end{array}
\left(\begin{array}{cc} \mathbf{p} & \mathbf{n} \\ 99 & -1 \\ 0 & 0 \end{array}\right)
\end{array}
$$

$$\text{Expected profit} = p(Y, p) \cdot b(Y, p) + p(N, p) \cdot b(N, p)$$
$$+$$
$$p(N, n) \cdot b(N, n) + p(Y, n) \cdot b(Y, n)$$

$$p(x, y) = p(y) \cdot p(x \mid y)$$

$$\text{Expected profit} = p(Y \mid p) \cdot p(p) \cdot b(Y, p) + p(N \mid p) \cdot p(p) \cdot b(N, p)$$
$$+$$
$$p(N \mid n) \cdot p(n) \cdot b(N, n) + p(Y \mid n) \cdot p(n) \cdot b(Y, n)$$

$$\text{Expected profit} = p(p) \cdot p(Y \mid p) \cdot b(Y, p) + p(N \mid p) \cdot c(N, p)$$
$$+$$
$$p(n) \cdot p(N \mid n) \cdot b(N, n) + p(Y \mid n) \cdot c(Y, n)$$

T = 110

P = 61

N = 49

p(p) = 0.55 p(n) = 0.45

tp rate = 56/61 = 0.92

fp rate = 7/49 = 0.14

fn rate = 5/61 = 0.08

tn rate = 42/49 = 0.86

$$\text{Expected profit} = p(p) \cdot p(Y \mid p) \cdot b(Y, p) + p(N \mid p) \cdot c(N, p) + p(n) \cdot p(N \mid n) \cdot b(N, n) + p(Y \mid p) \cdot c(Y, n)$$

$$= 0.55 \cdot 0.92 \cdot b(Y, p) + 0.08 \cdot b(N, p) + 0.45 \cdot 0.86 \cdot b(N, n) + 0.14 \cdot p(Y, n)$$

$$= 0.55 \cdot 0.92 \cdot 99 + 0.08 \cdot 0 + 0.45 \cdot 0.86 \cdot 0 + 0.14 \cdot -1$$

$$= 50.1 - 0.063$$

$$\approx \$50.04$$

This expected value means that if we apply this model to a population of prospective customers and mail offers to those it classifies as positive, we can expect to make an average of about $50 profit per consumer.

Other Example with Expected Value Framework

	Actual Purchase	Actual Non-purchase
Predicted Purchase	True Positive (TP)	False Positive (FP)
Predicted Non-purchase	False Negative (FN)	True Negative (TN)

https://blogs.oracle.com/ai-and-datascience/post/explain-the-quotso-whatquot-behind-machine-learning-models-with-the-expected-value-framework-part-2-of-3

Other Example with Expected Value Framework

$$E[X] = P(TP,p) * V(TP,p) + P(FN,p) * V(FN,p) + P(FP,n) * V(FP,n) + P(TN,n) * V(TN,n)$$

$$P(TP,p) = P(TP|p) * P(p)$$

$$E[X] = P(p) * [P(TP|p) * V(TP,p) + P(FN|p) * V(FN,p)] + P(n) * [P(FP|n) * V(FP,n) + P(TN|n) * V(TN,n)]$$

	Actual Purchase	Actual Non-purchase	TOTAL
Predicted Purchase	1000	1500	**2500**
Predicted Non-purchase	500	8000	**8500**
TOTAL	**1500**	**9500**	**11,000**

	Actual Purchase	Actual Non-purchase
Predicted Purchase	1000/1500 = **.67**	1500/9500 = **.16**
Predicted Non-purchase	500/1500 = **.33**	8000/9500 = **.84**
TOTAL	**1500**	**9500**

	Actual Purchase	Actual Non-purchase
Predicted Purchase	$305	-$15
Predicted Non-purchase	$0	$0

Other Example with Expected Value Framework

Let's plug all of this into our Expected Value equation:

$$E[X] = .14 * [.67 * 305 + .33 * 0] + .86 * [.16 * -15 + .84 * 0]$$

Solving for Expected Value, we get **$26.55**.

Let's circle back to the original statement we wanted to make about our model:

"If I apply this model to a new set of data on prospective customers and target my marketing efforts towards only those prospects predicted as purchasers, I can expect to make, on average, $26.55 profit per customer."

Francisco J. Cantú-Ortiz, PhD

Professor of Computer Science and Artificial Intelligence
Tecnológico de Monterrey
Enago-Academy Advisor for Strategic Alliances

E-mail: fcantu@itesm.mx, fjcantor@gmail.com
Cel: +52 81 1050 8294, SNI-2 CVU: 9804
Personal Page: http://semtech.mty.itesm.mx/fcantu/
Facebook: fcantu; Twitter: @fjcantor; Skype: fjcantor
Orcid: 0000-0002-2015-0562
Scopus ID:6701563520
Researcher ID: B-8457-2009
https://www.researchgate.net/profile/Francisco_Cantu-Ortiz
https://scholar.google.com.mx/citations?hl=es&user=45-uuK4AAAAJ
https://itesm.academia.edu/FranciscoJavierCantuOrtiz
Ave. Eugenio Garza Sada No. 2501, Monterrey N.L., C.P. 64849, México