

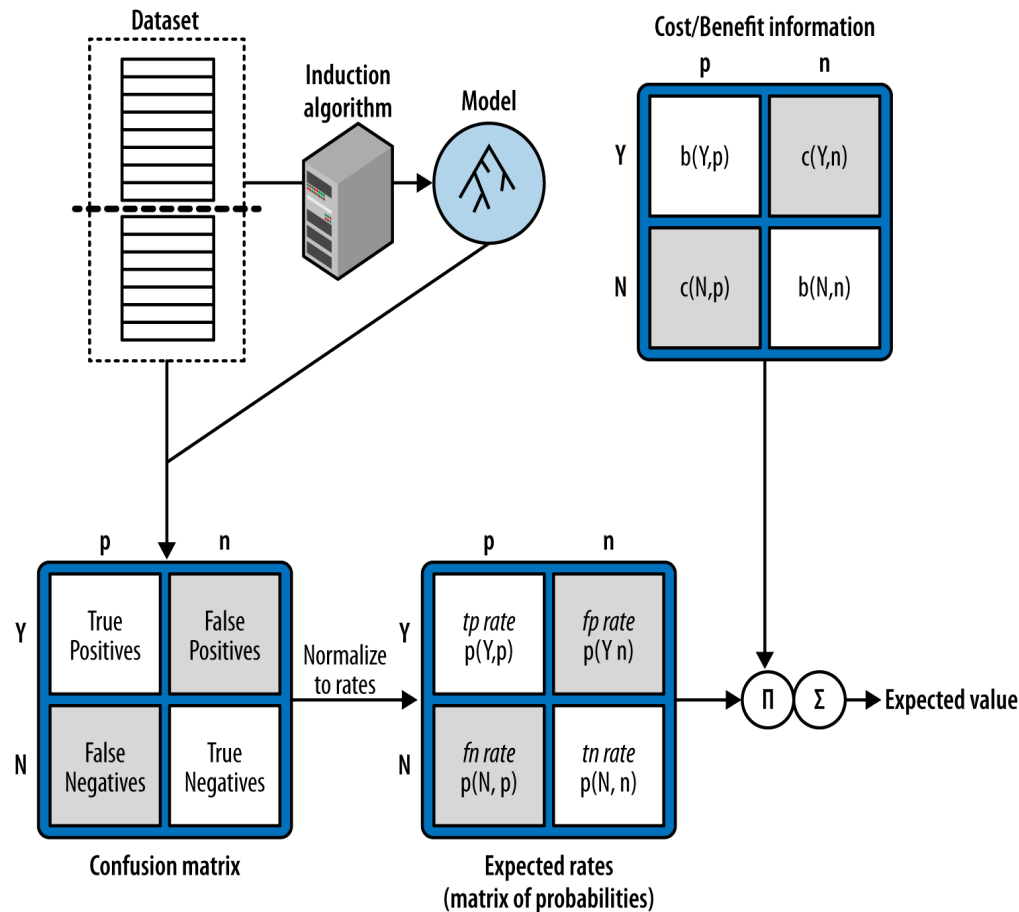
Introduction to Data Science

DECISION PROCESSES

BRIAN D'ALESSANDRO

Fine Print: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute, except for as needed as a pedagogical tool in the subject of Data Science.

DESIGNING A DECISION PROCESS



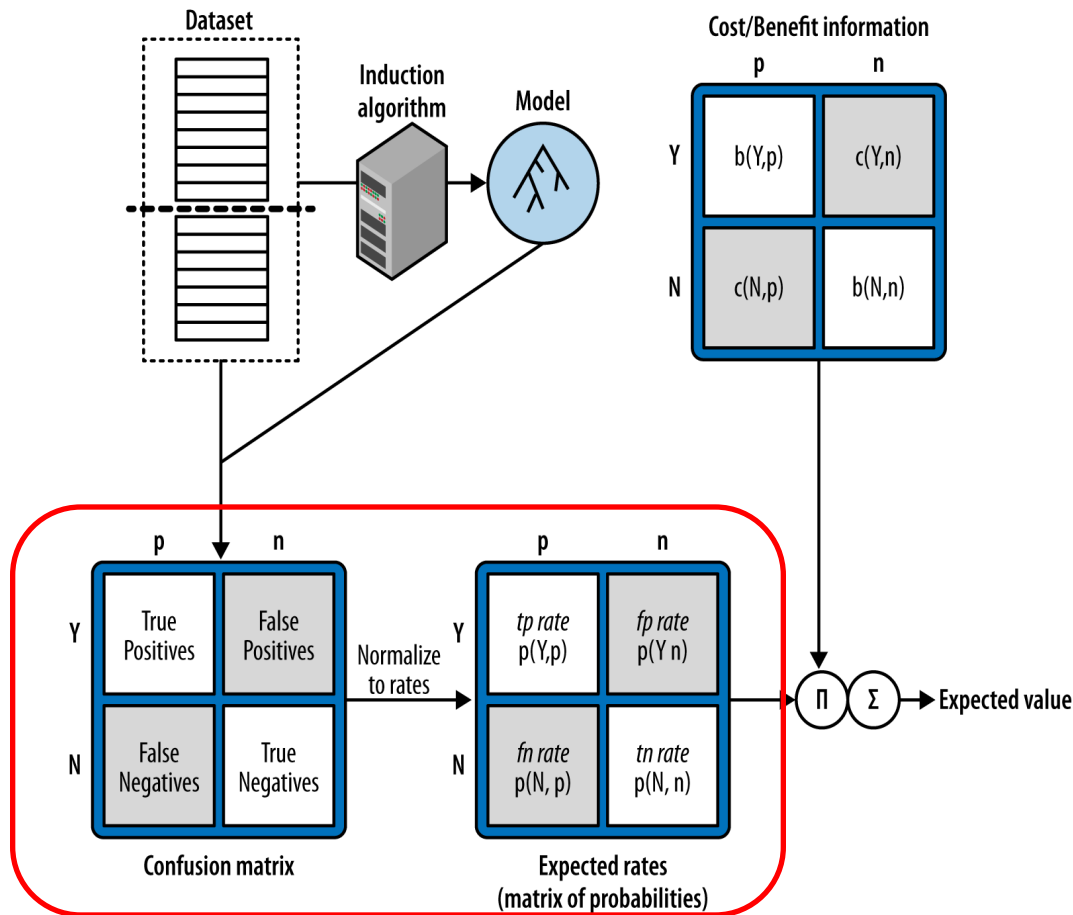
Your classifier is usually not the end state.

The classifier provides critical information for a decision system.

Once the classifier is trained, you still need to determine how to use it drive system decisions.

How do you design the system to make decisions that optimizes expected values?

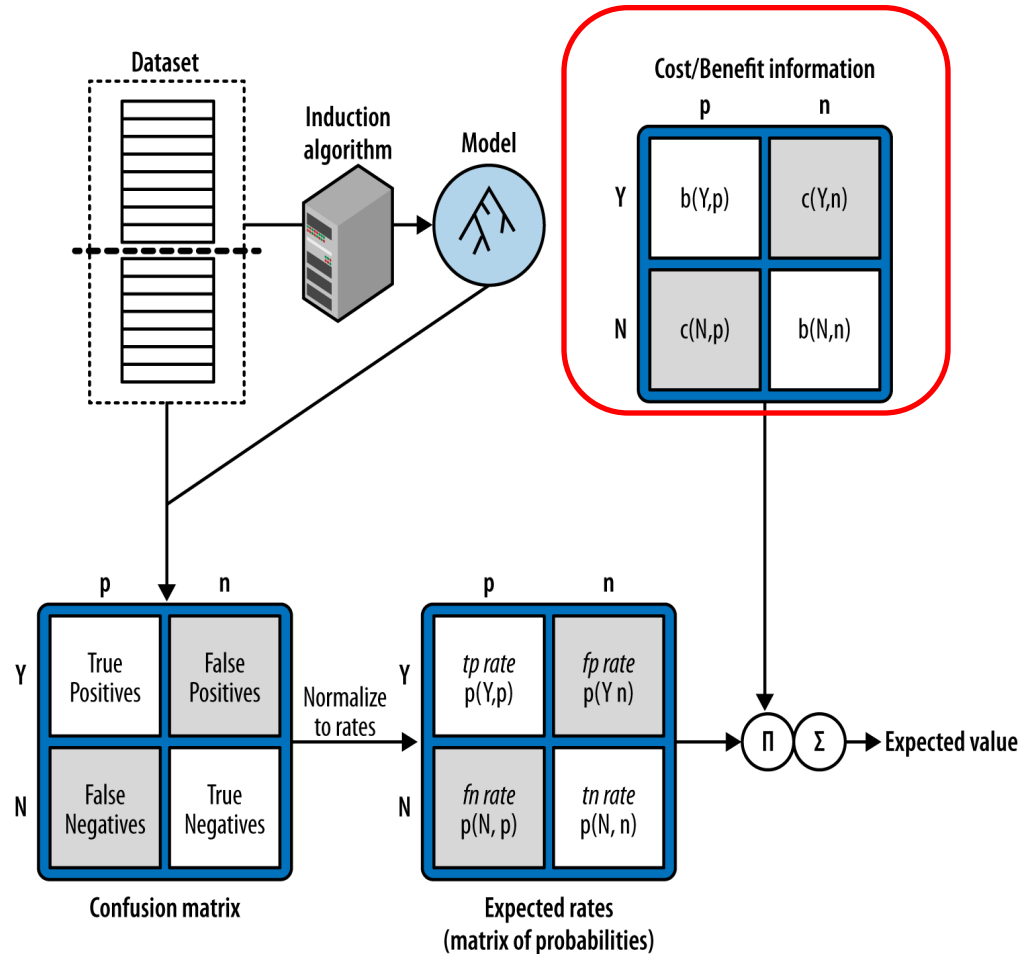
DESIGNING A DECISION PROCESS



The confusion matrix should be calculated from your true out-of-sample data.

The matrix and associated rates become your best working estimate for how the classifier will perform "in the wild"

DESIGNING A DECISION PROCESS

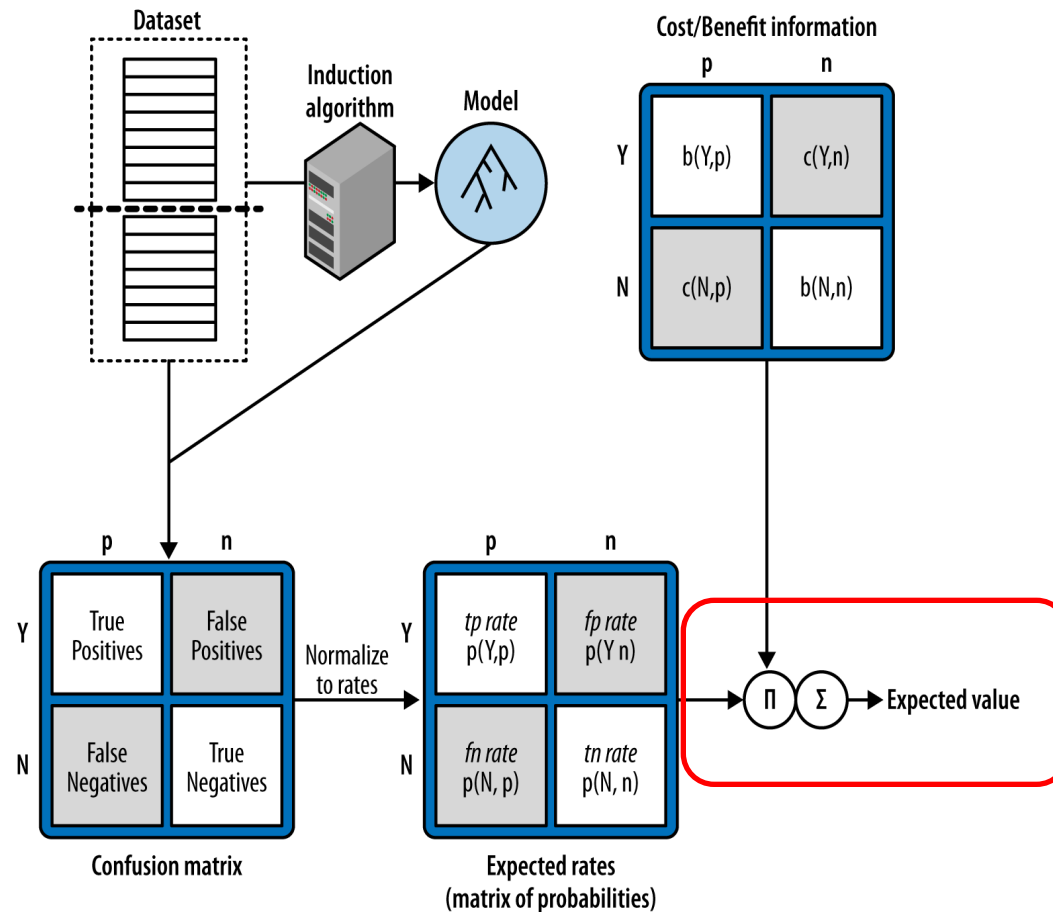


The cost-confusion matrix is designed from understanding the core business processes involved.

Designing this usually involves consulting with various stakeholders.

This should be part of every modeling project, as it informs the choice of metric, as well as the parameters of the decision process.

DESIGNING A DECISION PROCESS



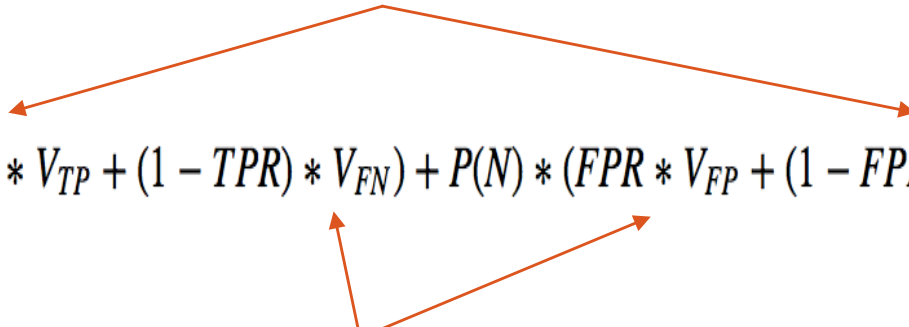
With the modeling and prep-work complete, the scientists can then proceed to design the rules of the decision process.

One good framework is to maximize the expected value of decision making.

COST SENSITIVE CLASSIFICATION

After learning a classifier's TPR vs FPR curve, and filling in the cost-confusion matrix, we can then compute different expected values to find thresholds that optimize expected values.

When we are right we generally incur some positive benefit

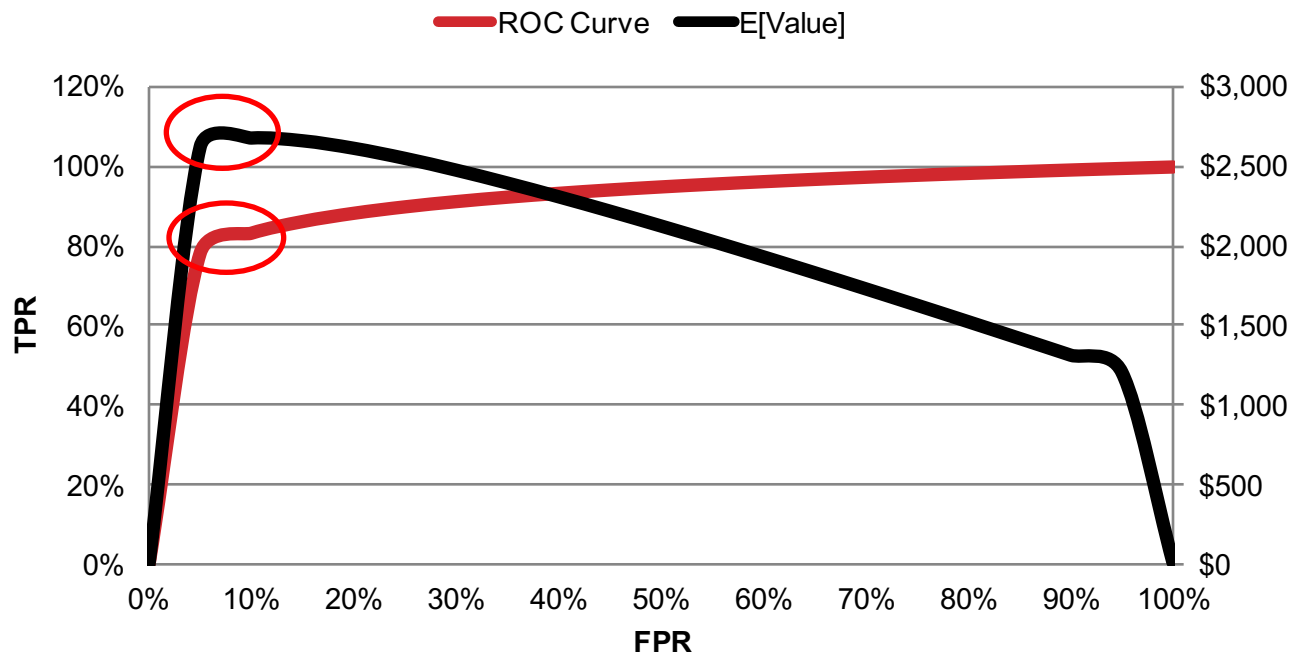


The diagram consists of two orange arrows. One arrow originates from the text 'When we are right we generally incur some positive benefit' and points to the V_{TP} term in the equation. The other arrow originates from the text 'When wrong we generally incur a negative value (loss)' and points to the V_{FN} term in the equation.

$$EV = P(Y) * (TPR * V_{TP} + (1 - TPR) * V_{FN}) + P(N) * (FPR * V_{FP} + (1 - FPR) * V_{TN})$$

When wrong we generally incur a negative value (loss)

COST SENSITIVE CLASSIFICATION



Using the EV formula on the previous slide with $[VTP, VFN, VFP, VTN] = [5000, 0, -8000, 0]$, we can see that an expected value optimizing threshold is one that produces a FPR of 30% and TPR of 74%.

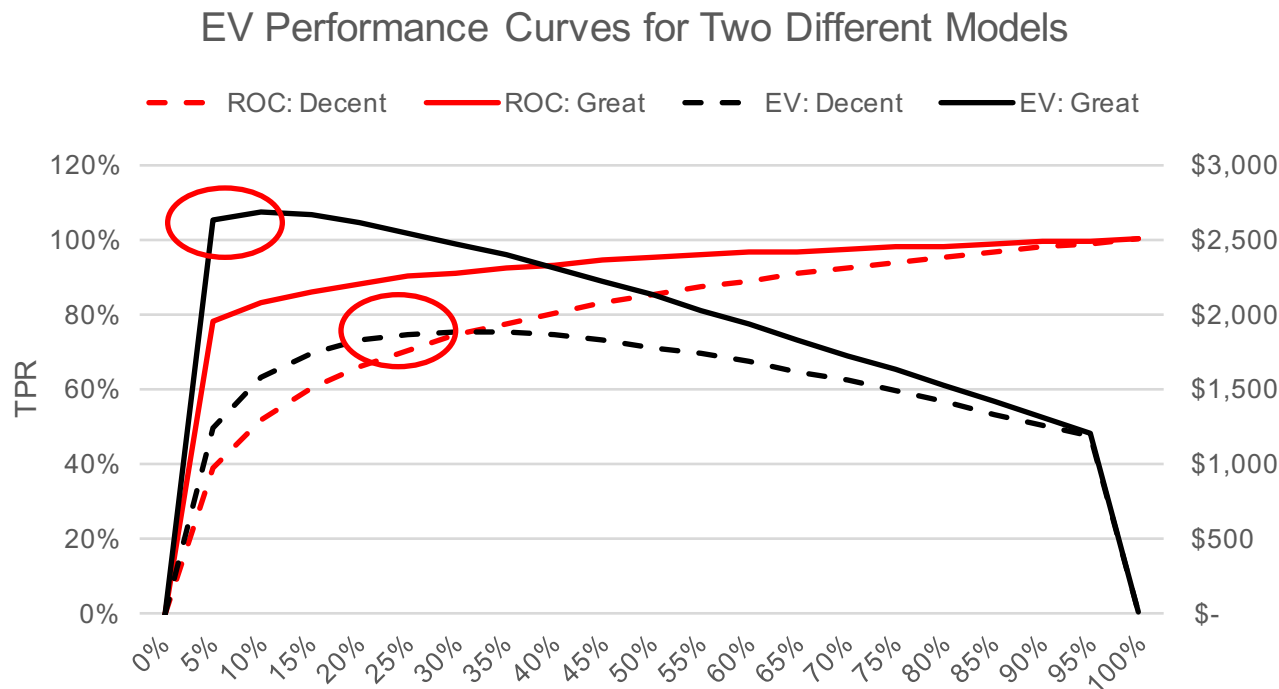
Each possible threshold gives us a (FPR, TPR) pair.

With each (FPR, TPR) pair we can compute the expected value of the classifier at that threshold.

With the expected value curve we can choose the threshold that maximizes expected performance.

COST SENSITIVE CLASSIFICATION

With a better model we can get more true positives per false positive, and our max expected value per classification goes up.



Improving the model can improve the expected value of our decision system.

It is important to remember that with a better model we need to find a new threshold in order to capture the increase in expected value

SOME THOUGHT STARTERS

For each of the following predictive modeling scenarios, answer the following:

1. Give a qualitative description, in terms relevant to the application domain, for a false positive and a false negative.
2. Make an assessment on the relative costs of a FP and a FN
3. If you were in charge of deploying the model (assume the model is fixed), how would you design the deployment system to minimize expected misclassification costs?

Modeling Scenarios:

- A medical screening test that classifies the presence of a brain tumor given fMRI images.
- A fraud detection system that automatically freezes an account if it suspects suspicious activity.
- A credit scoring system that automatically decides whether or not an applicant should receive a credit line.
- An automatic face tagging system for images uploaded to a social network.