The AI Playbook

Mastering the Rare Art of Machine Learning Deployment

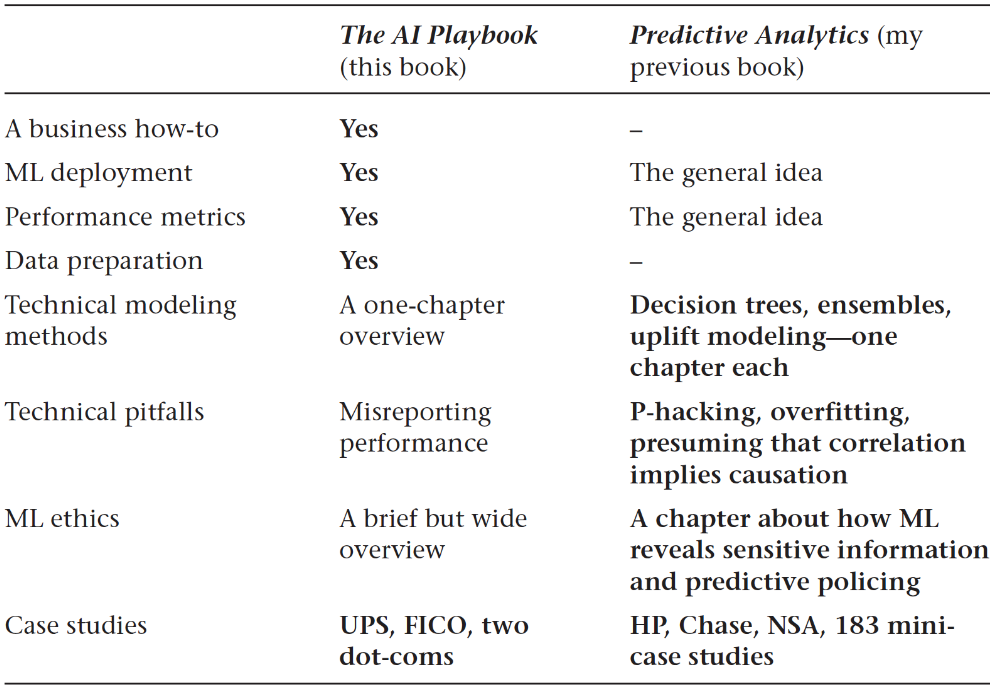
By Eric Siegel

PDF TO ACCOMPANY THE AUDIOBOOK

Click here for more information about the book – including access to its notes: references, plus resources for further learning.

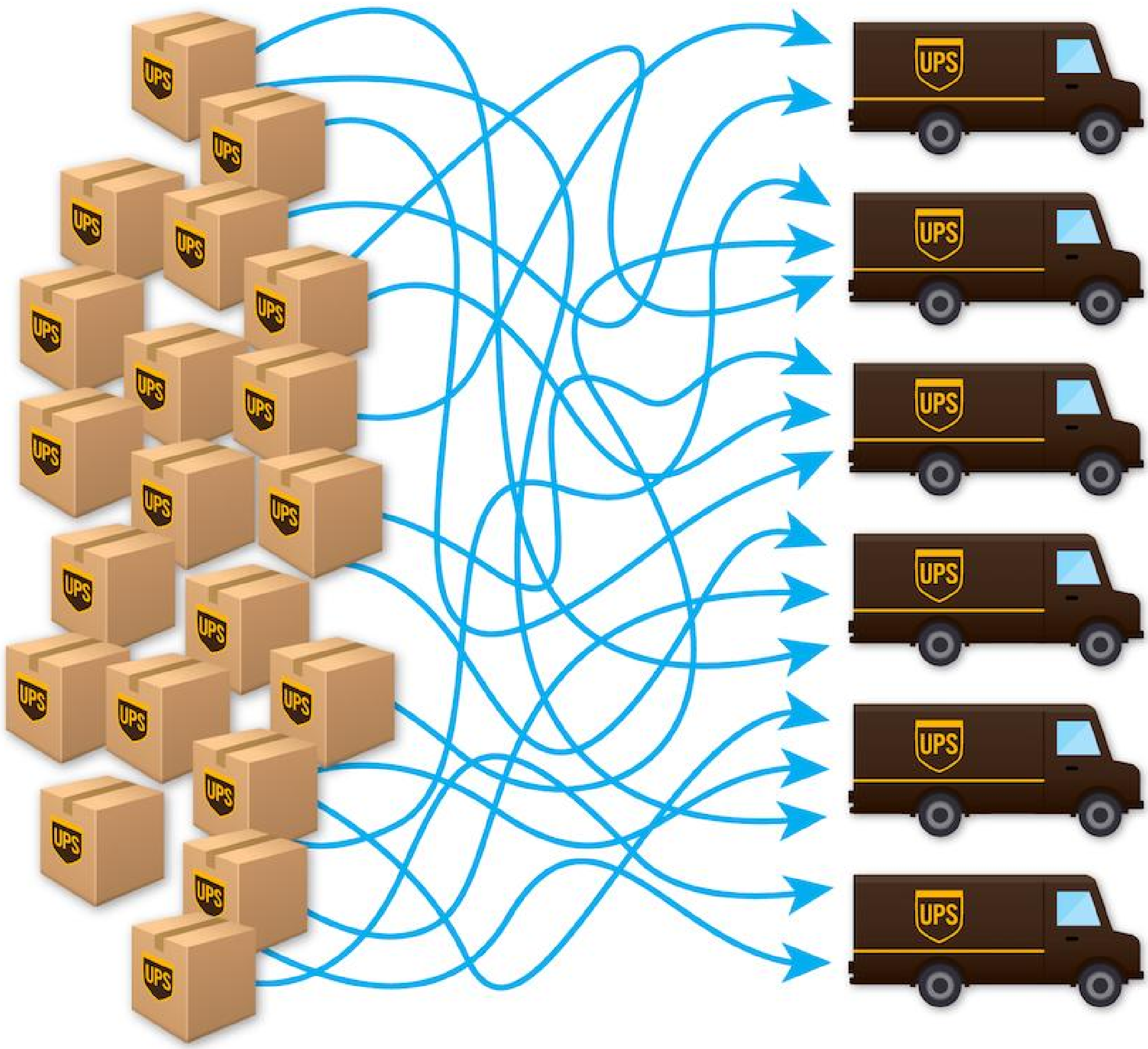
Click here for a tutorial glossary that includes the terms introduced within this book and more.

PREFACE

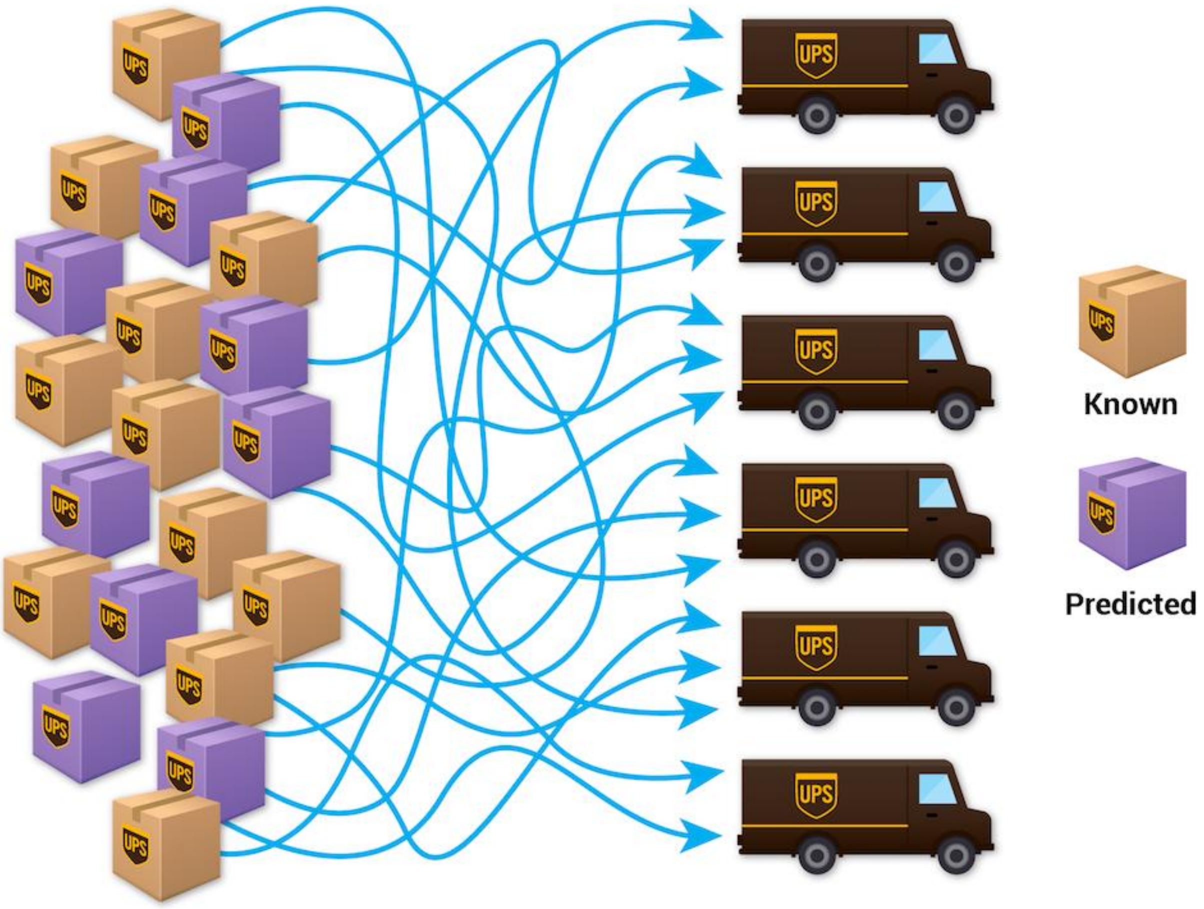


	The AI Playbook (this book)	***Predictive Analytics*** (my previous book)
A business how-to	**Yes**	–
ML deployment	**Yes**	The general idea
Performance metrics	**Yes**	The general idea
Data preparation	**Yes**	–
Technical modeling methods	A one-chapter overview	**Decision trees, ensembles, uplift modeling—one chapter each**
Technical pitfalls	Misreporting performance	**P-hacking, overfitting, presuming that correlation implies causation**
ML ethics	A brief but wide overview	**A chapter about how ML reveals sensitive information and predictive policing**
Case studies	**UPS, FICO, two dot-coms**	**HP, Chase, NSA, 183 mini-case studies**

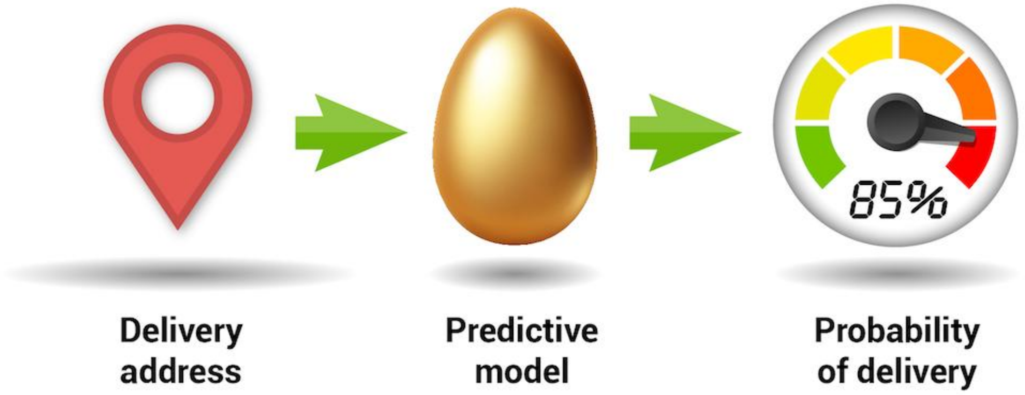
INTRODUCTION

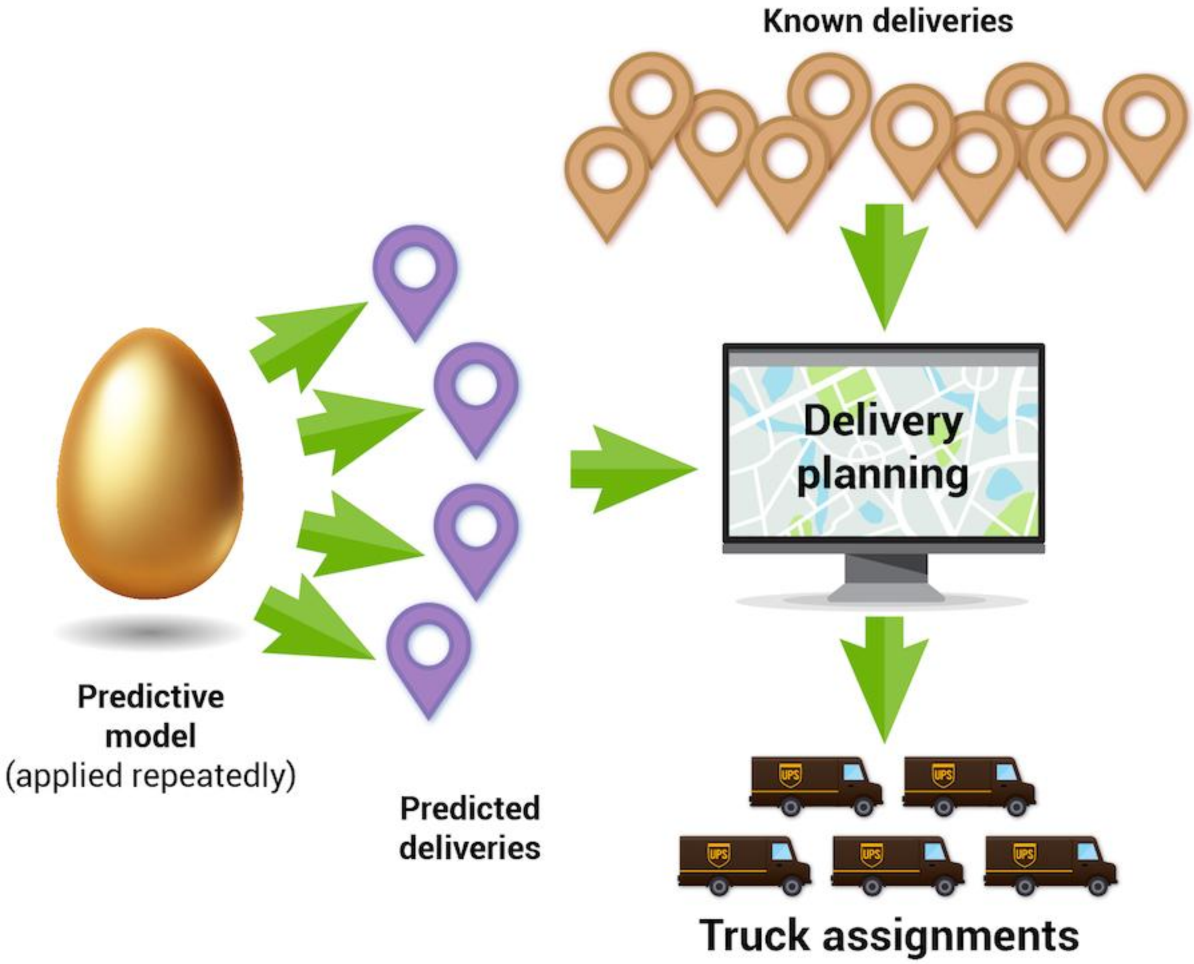


Packages are assigned to delivery trucks at a shipping center.

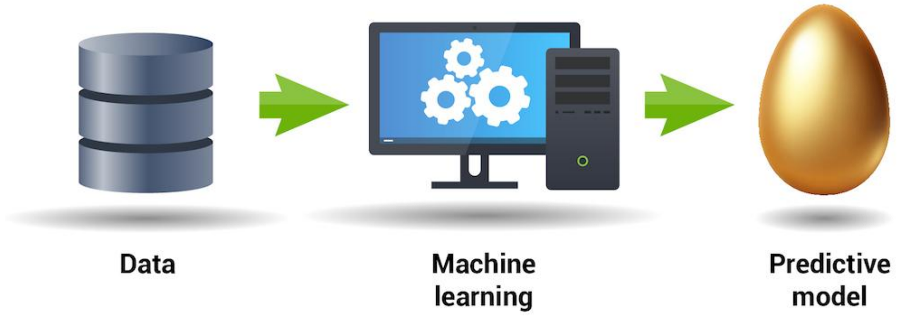


A combination of known and predicted deliveries is assigned to delivery trucks at a shipping center.

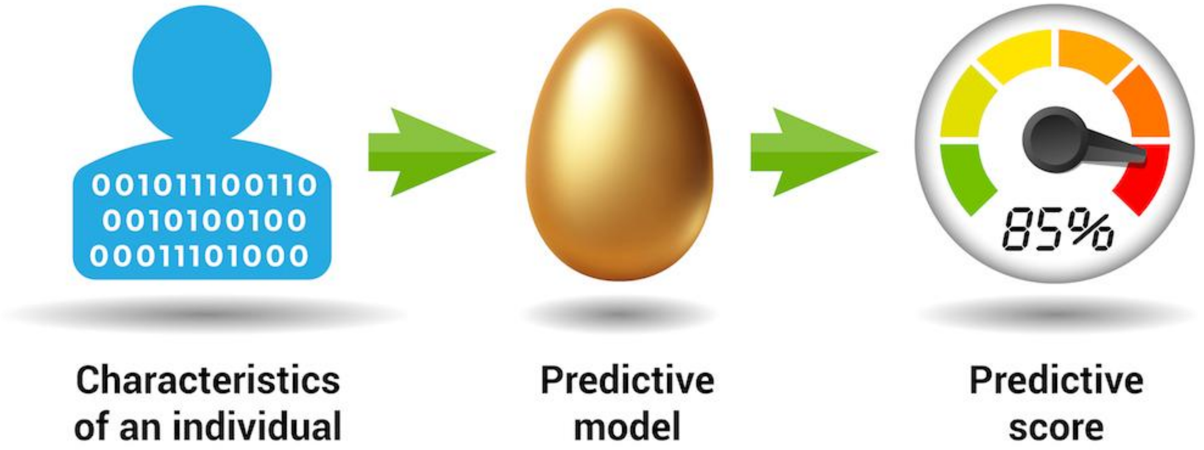




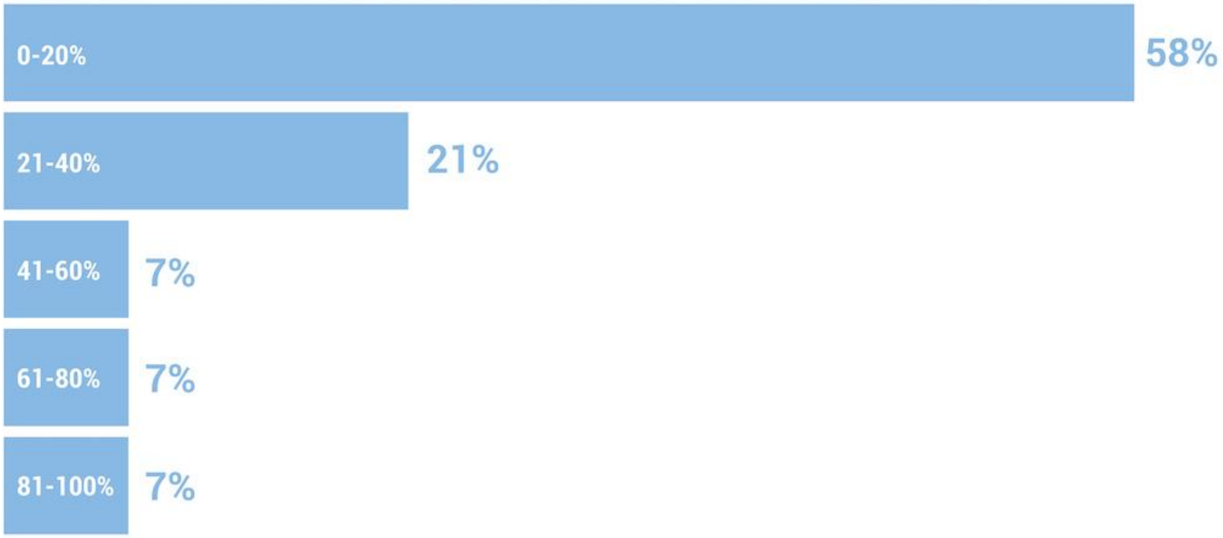
CHAPTER 0



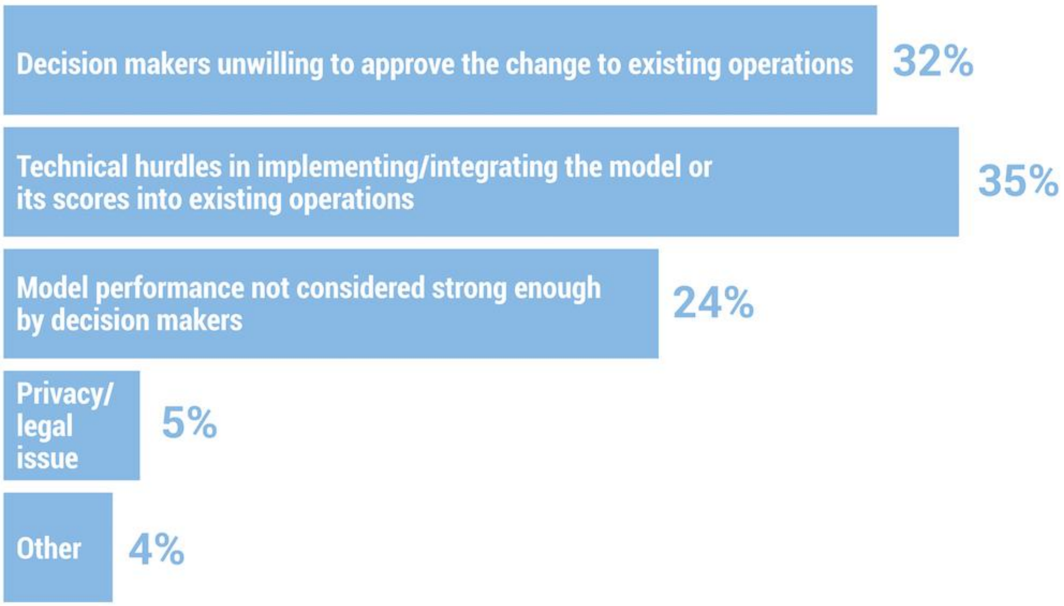
Machine learning generates a predictive model from data.



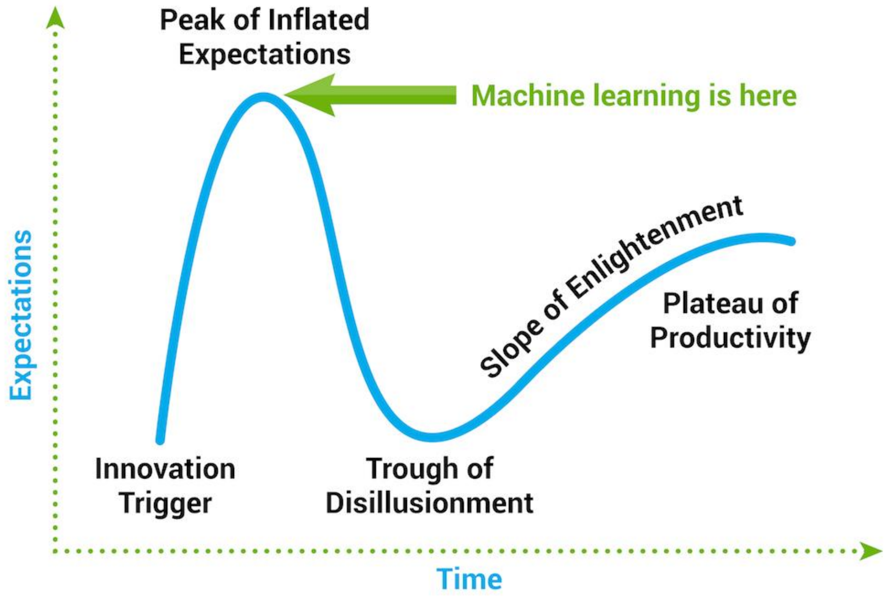
Scoring: A model generates a prediction for an individual.



Survey responses to the question, “What percentage of ML models (created by you or your colleagues with the intention of being deployed) have actually been deployed?” Total respondents: 114.



Survey responses to the question, “What is the main impediment to model deployment?” Total respondents: 114.

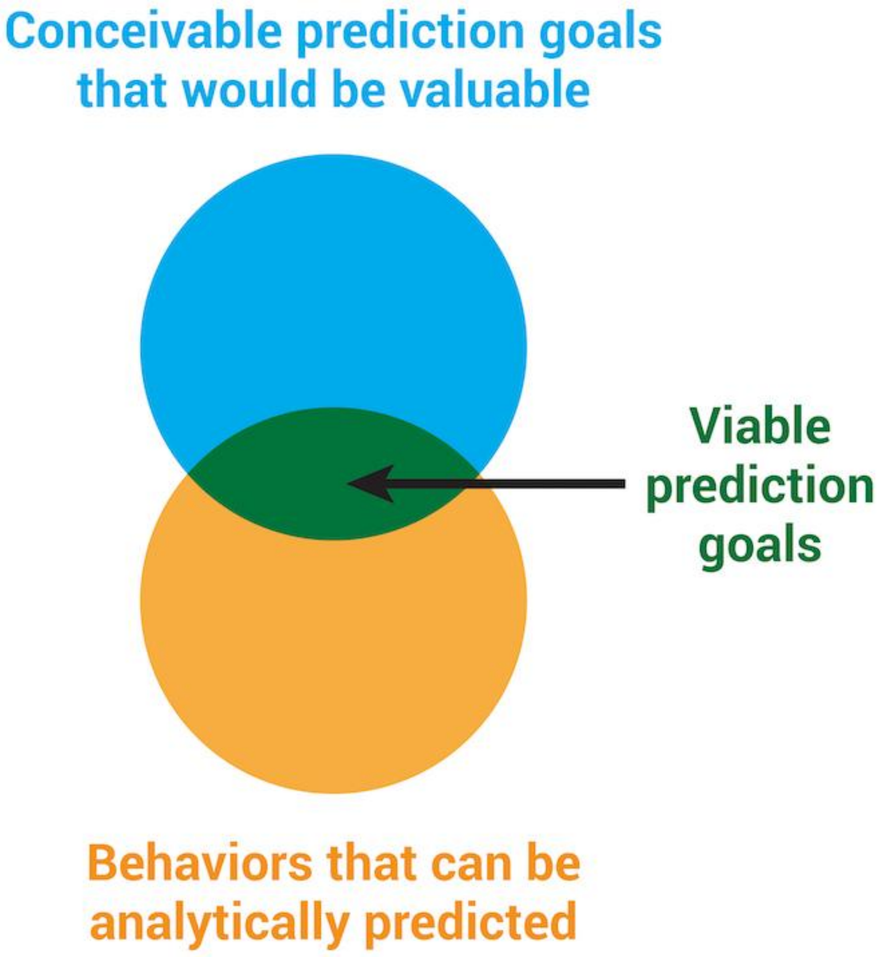


The Gartner hype cycle for technology.

Application and desired business outcome	**What's predicted** (model output)	**What's done about it** (deployment)
Response modeling to increase the marketing response rate	Will the customer buy if contacted?	Mail a brochure to those likely to buy.
Targeting ads to increase clicks	Will the user respond to this ad?	Display the ad to which the user is most likely to respond.
Churn modeling to decrease customer attrition	Will the customer defect if not contacted?	Reach out with a retention offer to those most likely to defect.
Credit scoring to decrease defaults	Will the debtor default on their loan?	Deny risky applications for credit.
Supply chain management to optimize inventory	How much demand will there be for each item?	Maintain stock levels accordingly.
Delivery prediction to plan for more efficient delivery	Will the address receive a package delivery?	Plan the delivery truck assignments of predicted packages alongside known ones.

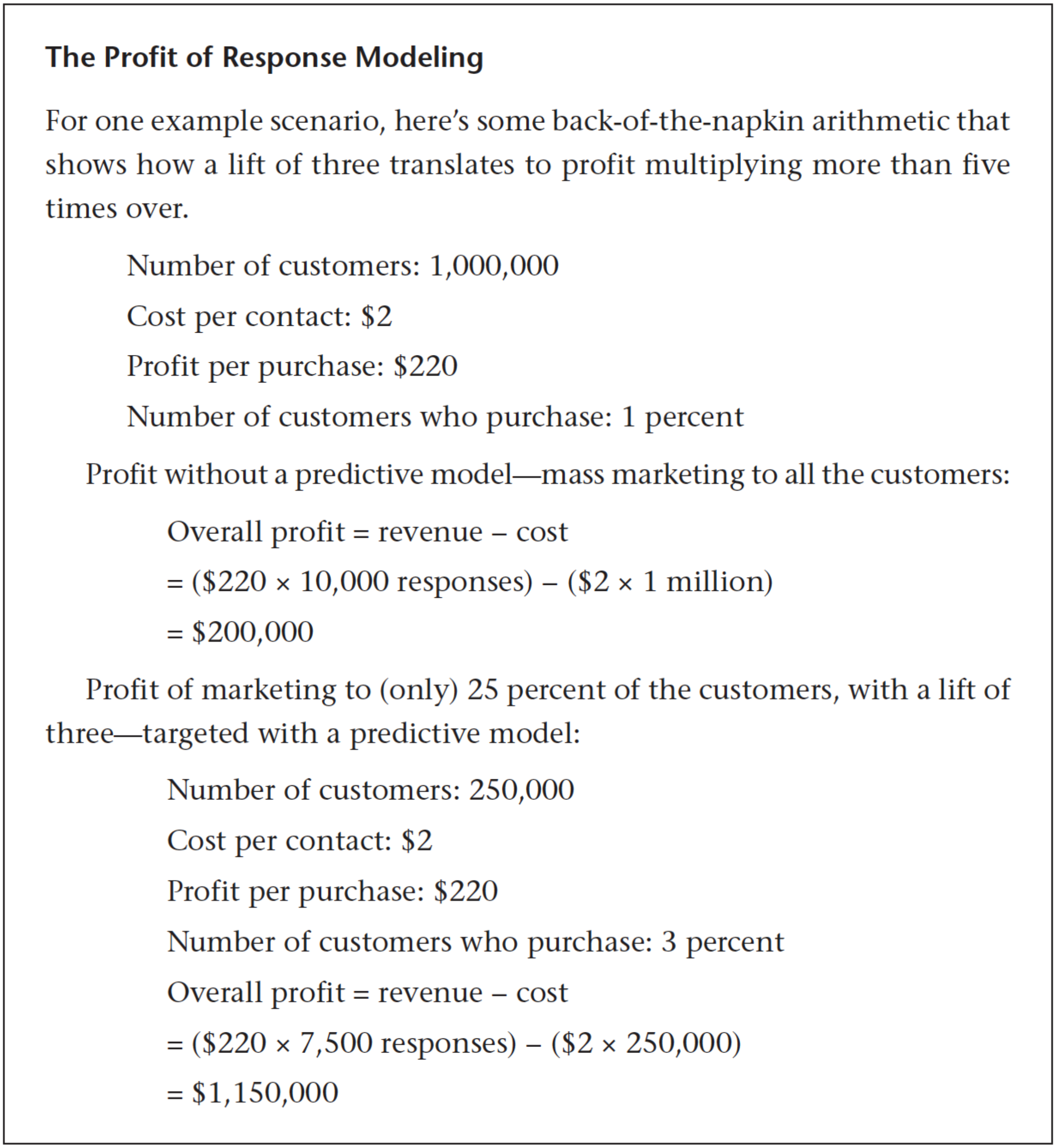
Application and desired business outcome	**What's predicted, i.e., detected** (model output)	**What's done about it** (deployment)
Fraud detection to prevent more fraud	Is the transaction fraudulent?	Place a hold on high-risk transactions and/or send them to human auditors.
Healthcare diagnosis to improve treatment	Does the patient have the condition?	Flag the patient for additional tests to potentially confirm the diagnosis.
Spam filtering so you see less spam	Is the email message spam?	Relegate spam to a separate email folder.
Speech recognition to transcribe the spoken word	Is X the word that corresponds with the audio segment?	Label the segment with the word predicted as most likely.
Fault detection to decrease the number of broken items	Is the item faulty (e.g., as it rolls off a factory assembly line)?	Inspect items predicted as likely to be faulty.
Autonomous driving to lessen human workloads and improve safety	Is there a stop sign in the image?	Bring the vehicle to a stop when a stop sign is detected.

CHAPTER 2


Conceivable prediction goals that would be valuable
Viable prediction goals
Behaviors that can be analytically predicted

CHAPTER 3

Corresponding with the sidebar below, "The Profit of Response Modeling," here is a spreadsheet detailing the calculations, which you may copy and toy with at will. **NOTE – PLEASE DO NOT REQUEST PERMISSION TO EDIT; INSTEAD, MAKE A COPY THAT YOU CAN EDIT.**



The Profit of Response Modeling

For one example scenario, here's some back-of-the-napkin arithmetic that shows how a lift of three translates to profit multiplying more than five times over.

Number of customers: 1,000,000

Cost per contact: $2

Profit per purchase: $220

Number of customers who purchase: 1 percent

Profit without a predictive model—mass marketing to all the customers:

Overall profit = revenue – cost

= ($220 × 10,000 responses) – ($2 × 1 million)

= $200,000

Profit of marketing to (only) 25 percent of the customers, with a lift of three—targeted with a predictive model:

Number of customers: 250,000

Cost per contact: $2

Profit per purchase: $220

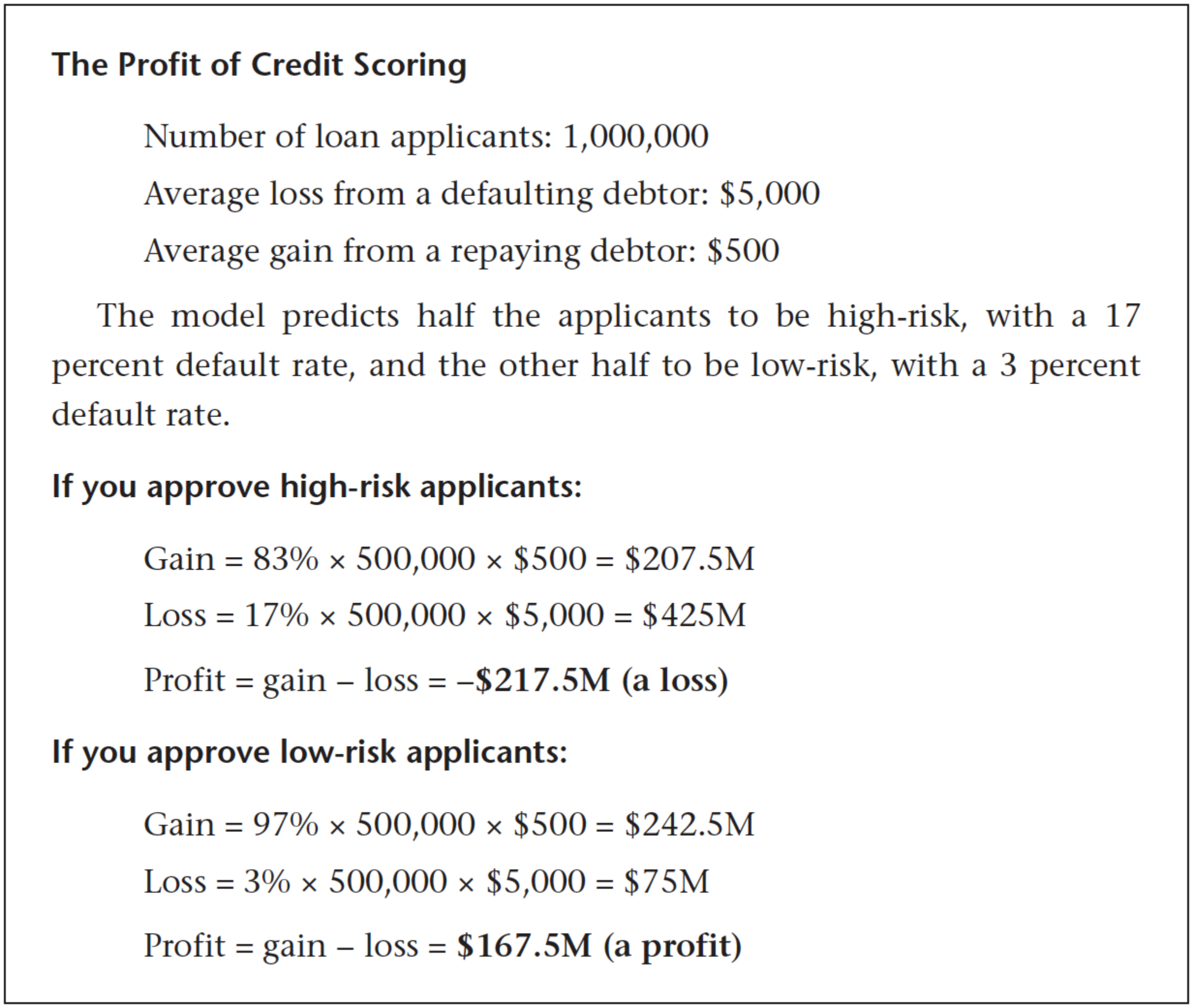
Number of customers who purchase: 3 percent

Overall profit = revenue – cost

= ($220 × 7,500 responses) – ($2 × 250,000)

= $1,150,000

Corresponding with the sidebar below, “The Profit of Credit Scoring,” here is a spreadsheet with the fraud example's calculations so that you can try out changes to the scenario at will. **NOTE – PLEASE DO NOT REQUEST PERMISSION TO EDIT; INSTEAD MAKE A COPY THAT YOU CAN EDIT.**



The Profit of Credit Scoring

Number of loan applicants: 1,000,000

Average loss from a defaulting debtor: $5,000

Average gain from a repaying debtor: $500

The model predicts half the applicants to be high-risk, with a 17 percent default rate, and the other half to be low-risk, with a 3 percent default rate.

If you approve high-risk applicants:

Gain = 83% × 500,000 × $500 = $207.5M

Loss = 17% × 500,000 × $5,000 = $425M

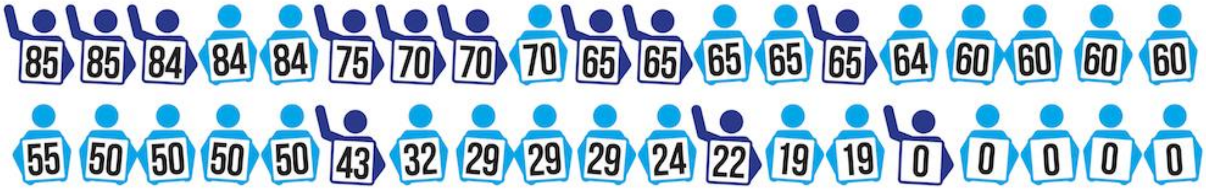
Profit = gain – loss = **–$217.5M (a loss)**

If you approve low-risk applicants:

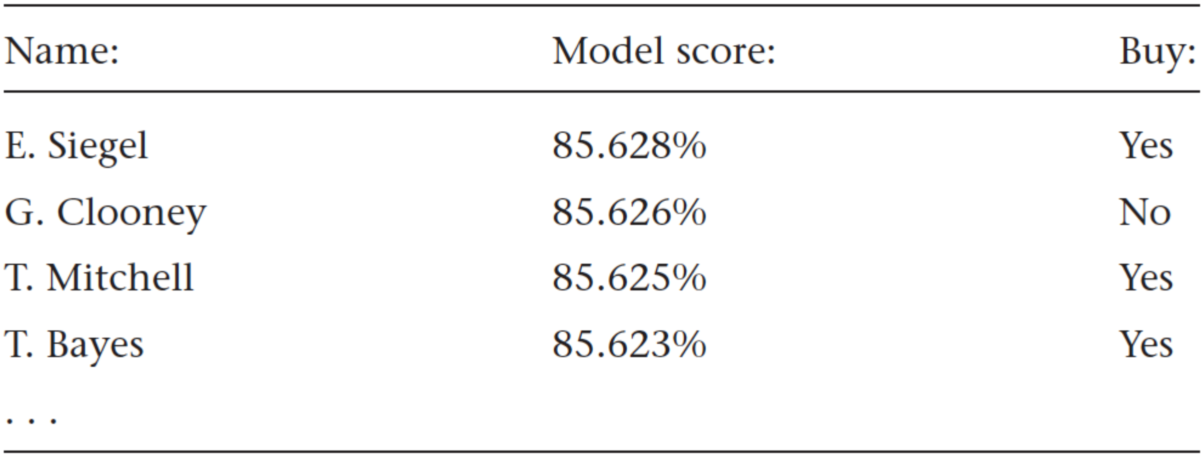
Gain = 97% × 500,000 × $500 = $242.5M

Loss = 3% × 500,000 × $5,000 = $75M

Profit = gain – loss = **$167.5M (a profit)**



People in order of their TV size (a zero means they have no TV). Those with a raised hand are subscribed to HBO.



Name:	Model score:	Buy:
E. Siegel	85.628%	Yes
G. Clooney	85.626%	No
T. Mitchell	85.625%	Yes
T. Bayes	85.623%	Yes
. . .		

First 100 cases:

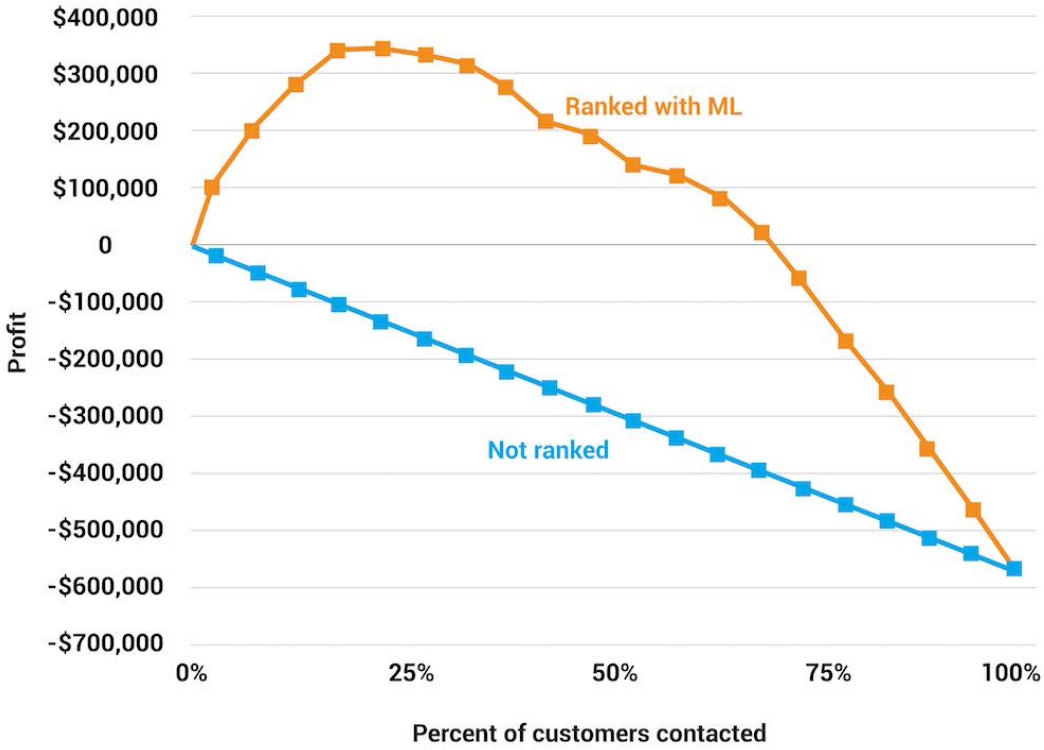
1011111111011110111100111111011111101111111111111011111001111111101
1111101111011111001110111110110111

Middle 100 cases:

10110100010110101110001000101111011100000110101110100110110100
10001011101101000101101110101101110000

Last 100 cases:

110000000010100100001000000010000000001000010000010000011000
0000100100001000011000010000100000100100



A profit curve.

The Cost Savings of Fraud Detection

Consider a bank that has issued 100,000 credit cards and each sees an average of 1,000 transactions per year, with one in 1,000 being fraudulent. To summarize:

Annual transactions: 100 million

Percentage that are fraudulent: 0.1 percent

Annual fraudulent transactions: 100,000

Cost per fraudulent transaction: $500 (the FN cost)

Annual loss from fraud: 100,000 × $500 = $50 million

It looks like crime does pay after all. But before you quit your day job to join the ranks of fraudsters, let's see what fraud detection could do to improve the situation.

If the bank is willing to treat two of every 1,000 attempted transactions as potentially fraudulent—holding the transaction and possibly inconveniencing the customer—then the onus is on a fraud detection model to flag which transactions should be held.

Let's assume the model attains a lift of 300. That's a lot higher than the lift of, say, 3 that we discussed in a previous example. But remember that lift is always relative to the size of the targeted group. In this case, we care about the lift only among the very top, small sliver of transactions scored as most likely to be fraudulent—the top 0.2 percent that will be blocked. We won't block any attempted transactions other than those, so that sliver is all that counts. Given that it's such a small portion, a high lift is feasible—a model's scores can potentially sort transactions well enough so that at least the very top portion includes a high concentration of positive cases.

First, we need to calculate how many errors occur, broken into FPs and FNs—how often the model wrongly blocks a legitimate transaction and how often it lets a fraudulent transaction slip by. Here's the breakdown:

Transactions blocked: 200,000 (two per 1,000)

Percentage blocked that are fraud: 30 percent

(Lift × overall fraud rate = 300 × 0.1 percent)

Fraudulent transactions blocked: 60,000 (30 percent × 200,000)

FPs—legitimate transactions blocked: 140,000 (200,000 – 60,000)

FNs—fraudulent transactions allowed: 40,000 (100,000 – 60,000)

This model is often wrong, but extremely valuable. When it blocks a transaction, it's usually wrong—only 30 percent of the blocked transactions are fraud. This isn't unusual. Since fraud is so infrequent, it would be very difficult to correctly detect some cases without also incorrectly flagging legit transactions even more often. With legitimate transactions—that is, negative cases—so prevalent, even misclassifying a small portion of them means a lot of FPs.

So the best we can hope for from a model is that it provides an advantageous trade-off between FPs (less costly) and FNs (more costly). To calculate the bottom line, we add up the costs. We've already established the cost for individual errors:

Cost of a FP: $100 (inconvenience to a customer)

Cost of a FN: $500 (fraudster gets away with it)

So we need only multiply these costs by how often they're incurred:

Aggregate FP cost: $14 million (140,000 at $100 each)

Aggregate FN cost: $20 million (40,000 at $500 each)

Total cost with fraud detection: $34 million

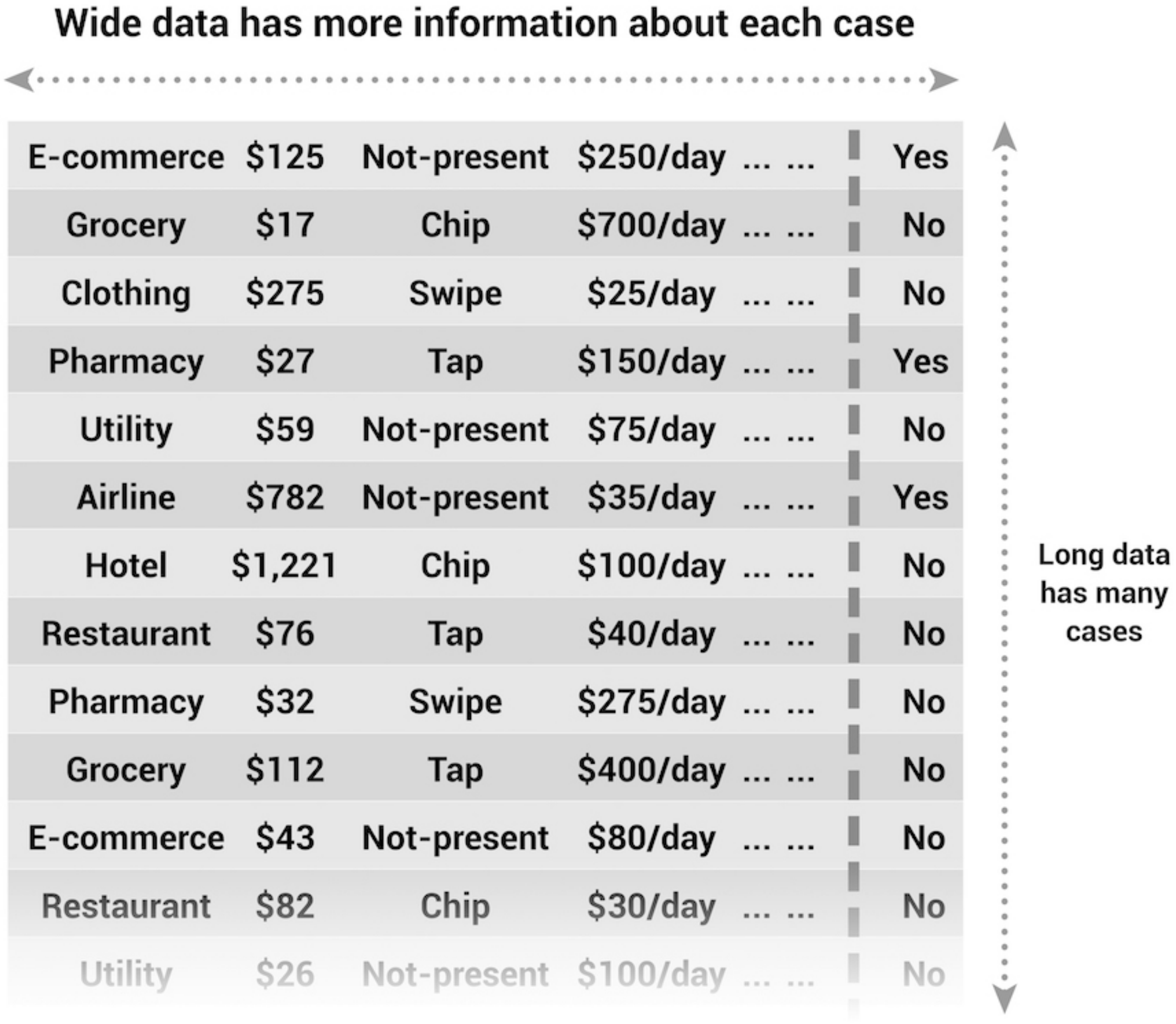
We've cut fraud losses by $30 million (from $50 million to $20 million), but introduced $14 million in new costs resulting from FPs. Clearly, this is a worthy trade-off.

Overall cost savings: $16 million ($50 million – $34 million)

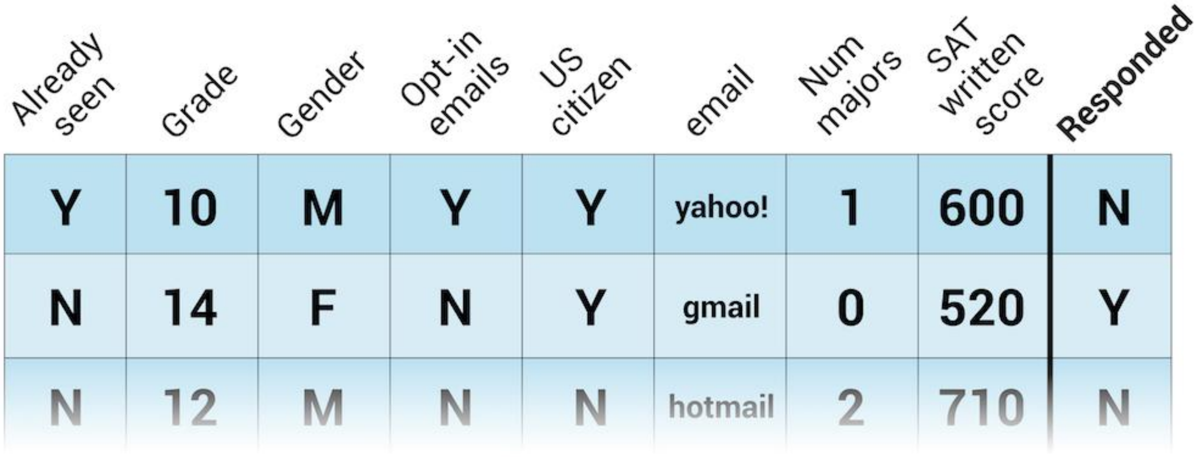
If you would like to access a spreadsheet with these calculations and try out different scenarios—such as varying the model lift, the number of transactions held, or the cost of each FP and FN—see the notes for this chapter at www.bizML.com.

CHAPTER 4

(Note the following table is not directly referred to in the audiobook:)


Wide data has more information about each case
Long data has many cases

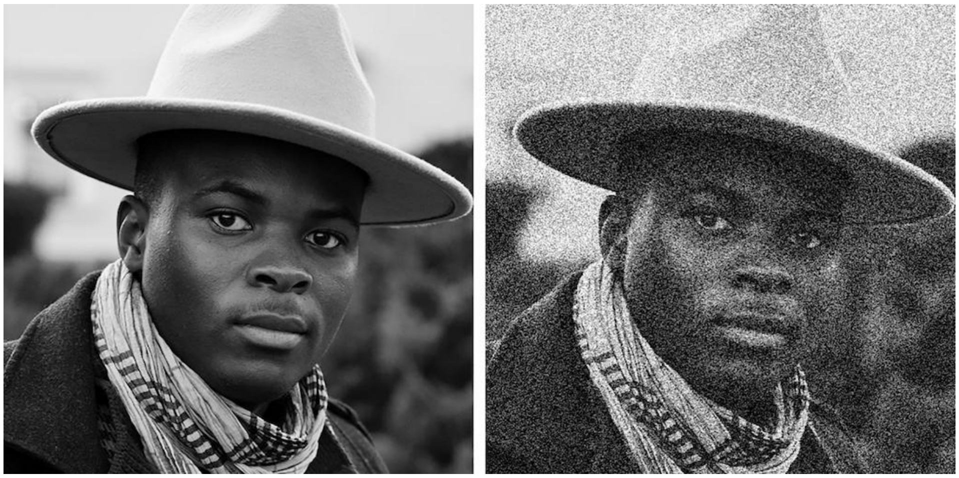
E-commerce	$125	Not-present	$250/day	... ...	Yes
Grocery	$17	Chip	$700/day	... ...	No
Clothing	$275	Swipe	$25/day	... ...	No
Pharmacy	$27	Tap	$150/day	... ...	Yes
Utility	$59	Not-present	$75/day	... ...	No
Airline	$782	Not-present	$35/day	... ...	Yes
Hotel	$1,221	Chip	$100/day	... ...	No
Restaurant	$76	Tap	$40/day	... ...	No
Pharmacy	$32	Swipe	$275/day	... ...	No
Grocery	$112	Tap	$400/day	... ...	No
E-commerce	$43	Not-present	$80/day	... ...	No
Restaurant	$82	Chip	$30/day	... ...	No
Utility	$26	Not-present	$100/day	... ...	No



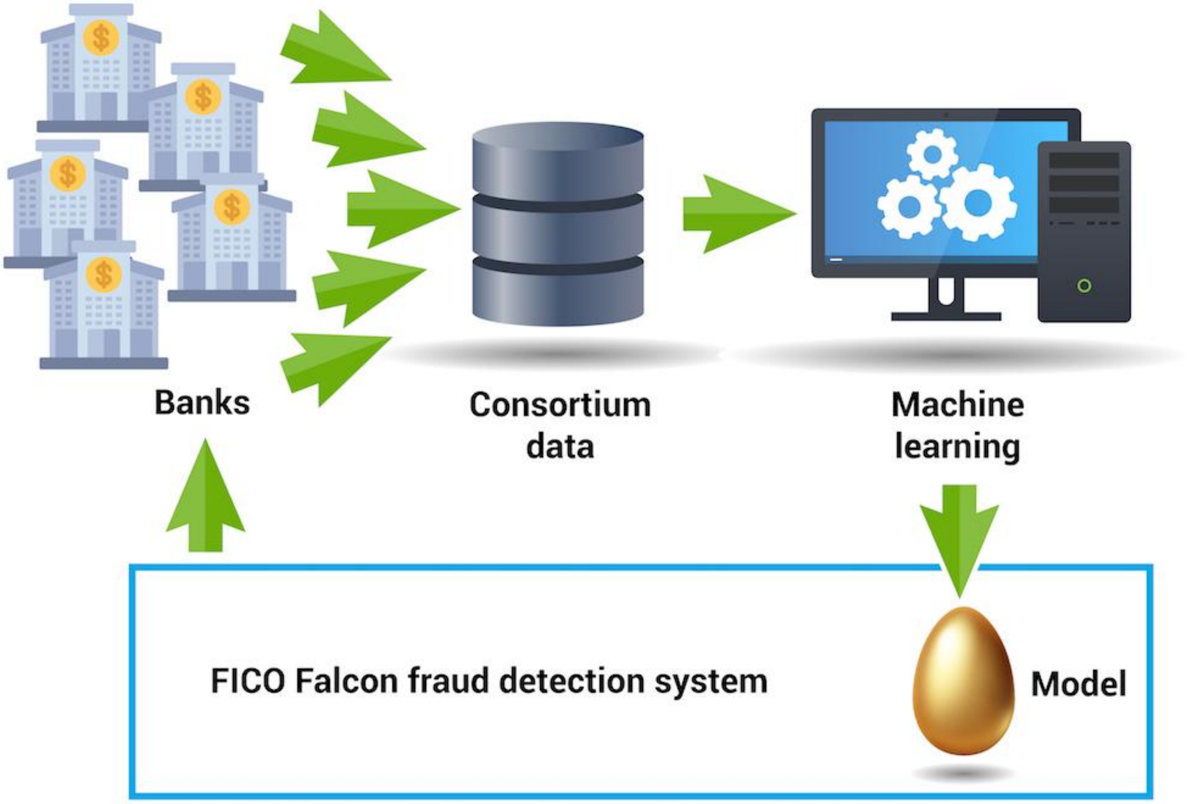
Already seen	Grade	Gender	Opt-in emails	US citizen	email	Num majors	SAT written score	**Responded**
Y	10	M	Y	Y	yahoo!	1	600	N
N	14	F	N	Y	gmail	0	520	Y
N	12	M	N	N	hotmail	2	710	N

Three rows of training data for modeling response to an ad for a university. Only a sample of the input variables (columns) are shown. Grade 14 means the second year of college.

BROCHURE
October 18
December 18
Male, CA, 10 purchases...
Response: Yes

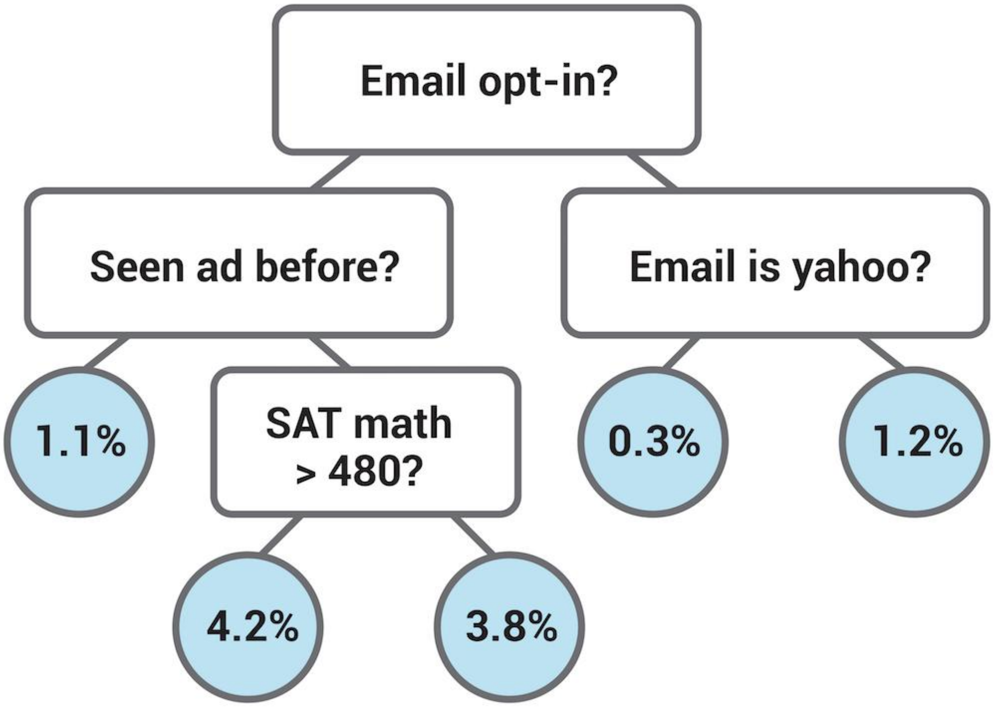


A photograph of a person wearing a hat is still easy to discern with noise added.

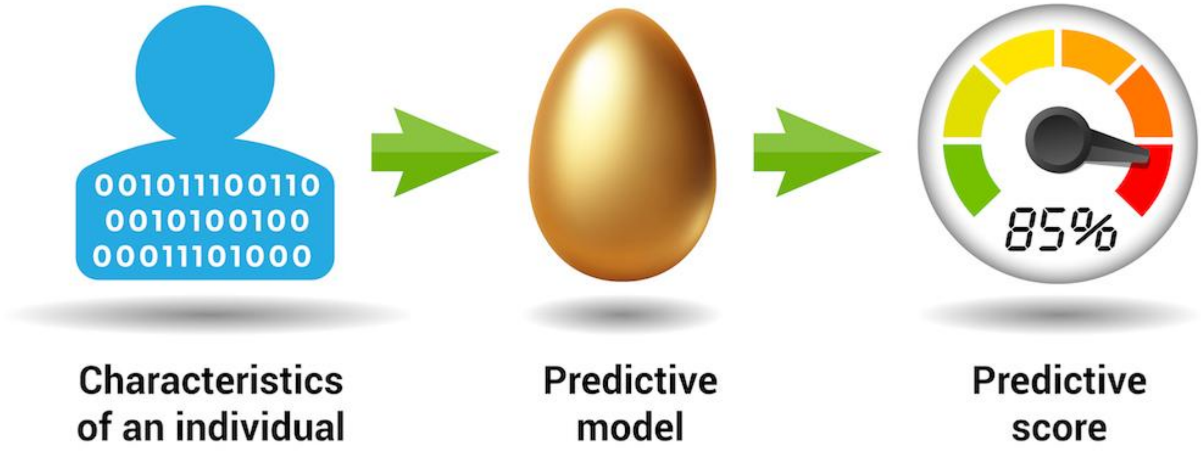


Banks provide data to develop Falcon's fraud detection model and Falcon deploys that model for each bank.

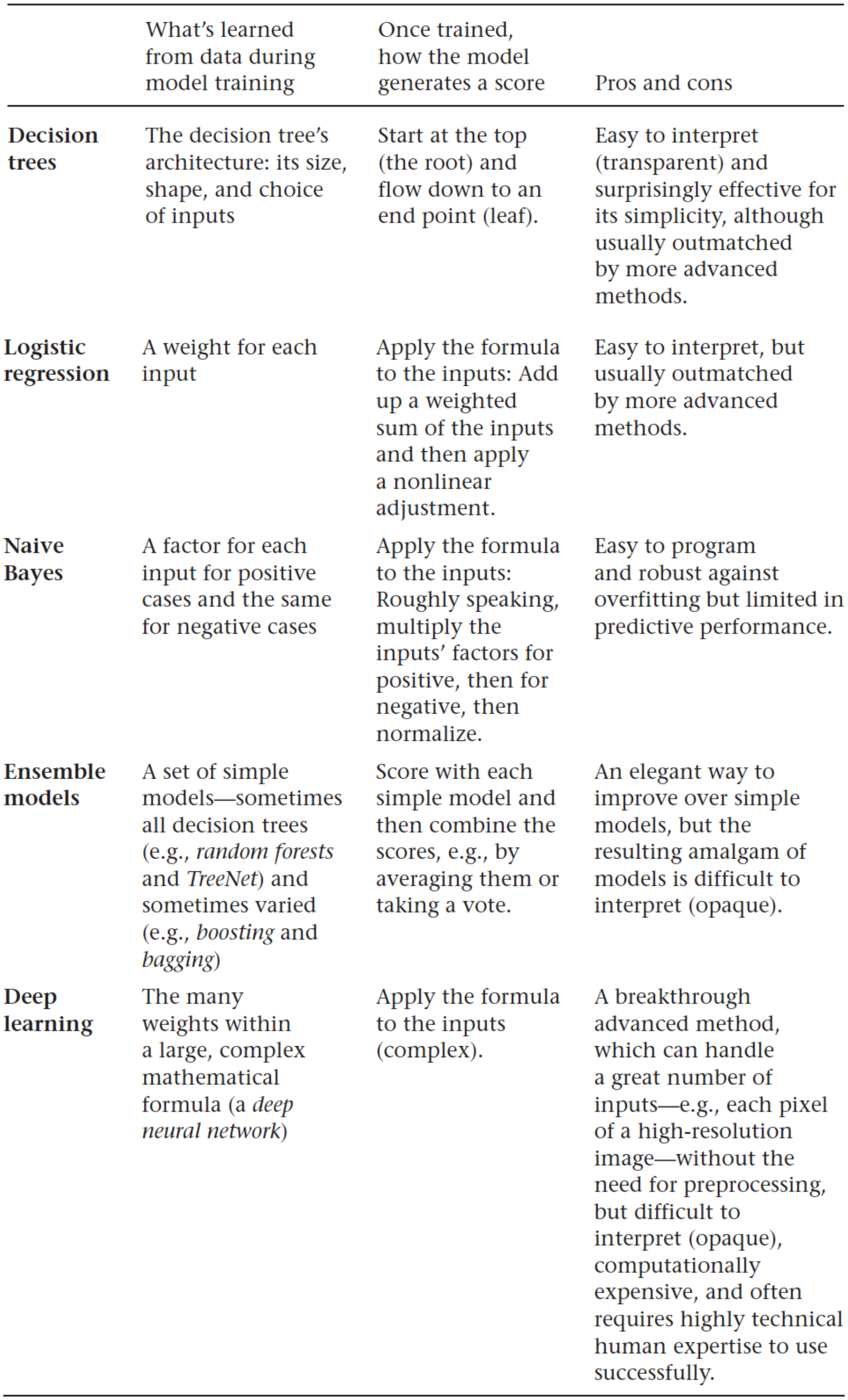
CHAPTER 5



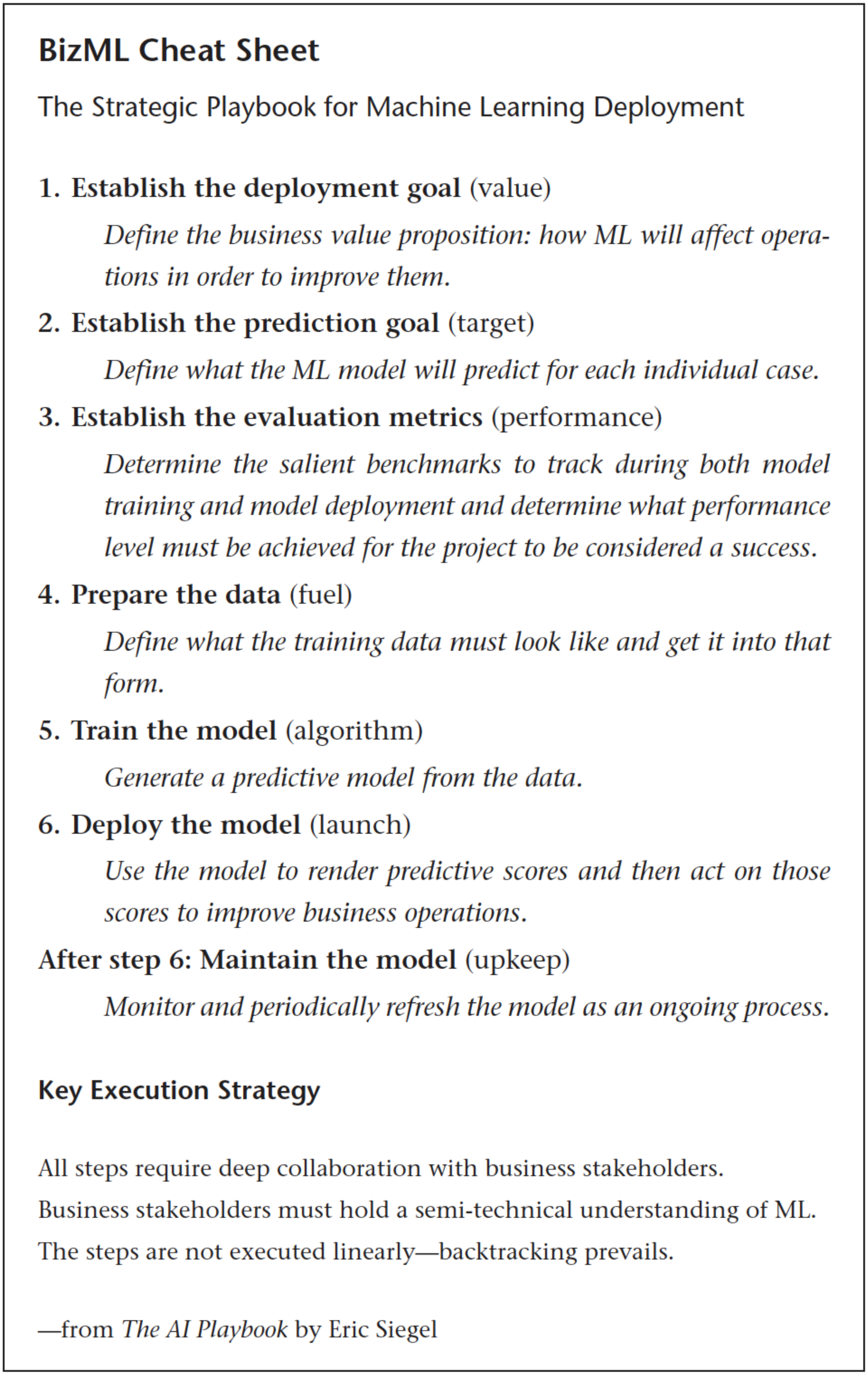
A decision tree to predict ad response. Start at the top. If the answer is yes, go left; otherwise, go right.



Scoring: A model generates a prediction for an individual.



	What's learned from data during model training	Once trained, how the model generates a score	Pros and cons
Decision trees	The decision tree's architecture: its size, shape, and choice of inputs	Start at the top (the root) and flow down to an end point (leaf).	Easy to interpret (transparent) and surprisingly effective for its simplicity, although usually outmatched by more advanced methods.
Logistic regression	A weight for each input	Apply the formula to the inputs: Add up a weighted sum of the inputs and then apply a nonlinear adjustment.	Easy to interpret, but usually outmatched by more advanced methods.
Naive Bayes	A factor for each input for positive cases and the same for negative cases	Apply the formula to the inputs: Roughly speaking, multiply the inputs' factors for positive, then for negative, then normalize.	Easy to program and robust against overfitting but limited in predictive performance.
Ensemble models	A set of simple models—sometimes all decision trees (e.g., *random forests* and *TreeNet*) and sometimes varied (e.g., *boosting* and *bagging*)	Score with each simple model and then combine the scores, e.g., by averaging them or taking a vote.	An elegant way to improve over simple models, but the resulting amalgam of models is difficult to interpret (opaque).
Deep learning	The many weights within a large, complex mathematical formula (a *deep neural network*)	Apply the formula to the inputs (complex).	A breakthrough advanced method, which can handle a great number of inputs—e.g., each pixel of a high-resolution image—without the need for preprocessing, but difficult to interpret (opaque), computationally expensive, and often requires highly technical human expertise to use successfully.



BizML Cheat Sheet

The Strategic Playbook for Machine Learning Deployment

1. **Establish the deployment goal** (value)

   Define the business value proposition: how ML will affect operations in order to improve them.

2. **Establish the prediction goal** (target)

   Define what the ML model will predict for each individual case.

3. **Establish the evaluation metrics** (performance)

   Determine the salient benchmarks to track during both model training and model deployment and determine what performance level must be achieved for the project to be considered a success.

4. **Prepare the data** (fuel)

   Define what the training data must look like and get it into that form.

5. **Train the model** (algorithm)

   Generate a predictive model from the data.

6. **Deploy the model** (launch)

   Use the model to render predictive scores and then act on those scores to improve business operations.

After step 6: Maintain the model (upkeep)

Monitor and periodically refresh the model as an ongoing process.

Key Execution Strategy

All steps require deep collaboration with business stakeholders.

Business stakeholders must hold a semi-technical understanding of ML.

The steps are not executed linearly—backtracking prevails.

—from *The AI Playbook* by Eric Siegel

Acknowledgments

My wife, Luba Gloukhova, contributed immensely to this book. While any author's partner is tasked with cultivating patience and perhaps taking on extra parenting responsibilities, Luba carried the additional burden of being a subject matter expert herself: She's a data scientist, and an insightful, experienced one at that (see her consultancy at www.datagie.com). As such, I endlessly solicited her counsel at meals, on walks, and in playgrounds. She served as a 24/7 sounding board as well as an intensive reviewer. Luba's more formal partnerships with me also provided value for this book, including her work as the founding chair of the Deep Learning World conference series, the editor-in-chief of the *Machine Learning Times*, and the content editor of my online course. I could not have written this book without Luba's enthusiasm and support. I must add, though, that her virtuoso performance as loving wife and mother transcends all of that!

My parents and in-laws also went to extra lengths to support me in this project. Lisa Schamberg (mother), Andrew Siegel (father), Anna Gloukhov (mother-in-law), and Maya Kanyuka (grandmother-in-law) contributed priceless encouragement as well as insights for my writing.

My literary agent, the incomparable Jim Levine of Levine Greenberg Rostan, provided critical corrections to my course during early, formative stages of this project and he followed through in later stages to help me bring the book's vision into a more carefully crafted reality. If not for his keen wisdom and business acumen, this book would have made a lot less sense.

My editor, Catherine Woods of the MIT Press, knows what makes a book readable, relatable, and relevant. Her feedback was critical as I worked through many balancing acts in this book's formulation. And Abbie Lundberg, despite her responsibilities as editor in chief of *MIT Sloan Management Review*, took the time to dig in deep and provide much-needed feedback that improved the book greatly. Earlier on, Emily Taber believed in this project and took me on as an author. Despite being on her way out by the time I finished writing—to another publisher after thirteen years at the MIT Press—she went the extra mile providing in-depth feedback.

Laurie Harper, Alison Jones, David Lamb, Barbara Monteiro, Barbara Oakley, and Myles Thompson served as mentors with a great deal of experience in the publishing industry. They advised me from this project's early conception on navigating the publication process, as well as providing substantive input on the project's formulation.

A career in tech is incomplete without a group of chums with whom you not only collaborate but, more importantly, talk shop on the regular . Each a senior consultant as well as an entrepreneur, Dean Abbott, John Elder, Karl Rexer, and James Taylor have enhanced my knowledge and enjoyment over the years and—more recently—provided extensive reviewing of an earlier draft of this book.

I've also learned tons from a much broader swathe of machine learning professionals: the 17,000 attendees and speakers who have participated at the event series I founded in 2009, Machine Learning Week (www.machinelearningweek.com), which includes the Predictive Analytics World and Deep Learning World conferences. Because of the productive and vibrant community that has attended our dozens of events internationally, I've enjoyed a 360-degree view of the industry that greatly informed this book. I'd like to thank my business partner in its production, Matthew Finlay, and his crackerjack team at Rising Media, who know how to make events excite and unite.

Many at the University of Virginia Darden School of Business—where I held a one-year position as Bodily Bicentennial Professor of Analytics during the writing of this book—contributed significantly to my formulation of ideas as I honed the methodology described herein. I am grateful to Samuel Bodily, the professor emeritus after whom my position was named, who personally provided a great deal of input and guidance on my work, much of which made its way into this book. Various faculty at Darden and other UVA departments also provided such input, including Michael Albert, Alexander Cowan, Renée Cummings, Rupert Freeman, Yael Grushka-Cockayne, Marc Ruggiano, Bill Scherer, Eric Tassone, and Sasa Zorc.

I would like to thank Jack Levis, Scott Zoldi, and Gerhard Pilcher, all innovators in ML deployment who endured my extensive interviews filled with many probing questions.

It takes a village. It was my great fortune that the following people invested time and care reviewing earlier drafts of these chapters: Rich Heimann, Eugene Kirpichov, Sam Koslowsky, Barry Lyons, Matthew Mayo, Glenn McMahan, Gregory Piatetsky-Shapiro, Steven Ramirez, Jonathan Sloan, Graham Southorn, David Stephenson, Martin Szugat, Morgan Vawter (in addition to providing this book's inspirational foreword), and Evan Wimpey.

Thank you, Sean O'Brien and Catherine Truxillo at SAS for their support in producing and hosting my online course series, "Machine Learning Leadership and Practice: End-to-End Mastery" (www.machinelearning.courses), the content of which served as a foundation for this book. Thanks also to William Goodrum for assistance with the course's content.

Many thanks to *Harvard Data Science Review* for originally publishing my writing about UPS's deployment story in an article entitled, "To Deploy Machine Learning, You Must Manage Operational Change—Here Is How UPS Got It Right" (Issue 5.2, Spring 2023). The article tells the part of the UPS story covered mostly by this book's sixth chapter, as well as some aspects covered within the introduction.

Thank you to *Scientific American*, which originally published my article on the *accuracy fallacy*, an earlier version of chapter 3's opening.

Thanks to the supremely gifted designer Matt Kornhaas for the figures throughout these chapters.

Finally, here's a shout-out to the extra-special educators, of whom I had more than my fair share: Thomas McKean (kindergarten), Chip Porter (grades 4–6), Margaret O'Brien (Burlington High School, Vermont), Harry Mairson (Brandeis University), Richard Alterman (Brandeis University), James Pustejovsky (Brandeis University), and Kathleen McKeown (Columbia University).

The aforementioned have molded me and bolstered this book. Nevertheless, I alone take responsibility for errors or failings of any kind in its contents.

About the author

Eric Siegel, Ph.D. is a leading consultant and former Columbia University professor who helps companies deploy machine learning. He is the founder of the long-running Machine Learning Week conference series, the instructor of the acclaimed online course "Machine Learning Leadership and Practice – End-to-End Mastery," executive editor of *The Machine Learning Times*, and a frequent keynote speaker. He wrote the bestselling *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*, which has been used in courses at hundreds of universities. Eric's interdisciplinary work bridges the stubborn technology/business gap. At Columbia, he won the Distinguished Faculty award when teaching the graduate *computer science* courses in ML and AI. Later, he served as a *business school* professor at UVA Darden. Eric also publishes op-eds on analytics and social justice.

Eric has appeared on Bloomberg TV and Radio, BNN (Canada), Israel National Radio, National Geographic Breakthrough, NPR Marketplace, Radio National (Australia), and TheStreet. Eric and his previous book have been featured in *Big Think, Businessweek, CBS MoneyWatch, Contagious Magazine, The European Business Review, The Financial Times, Forbes, Fortune, GQ, Harvard Business Review, The Huffington Post, The New York Review of Books, The New York Times, Newsweek, Quartz, Salon, The San Francisco Chronicle, Scientific American, The Seattle Post-Intelligencer, The Wall Street Journal, The Washington Post,* and *WSJ MarketWatch*.

- ➢ *Eric Siegel is available for select lectures. To inquire:* www.MachineLearningKeynote.com

- ➢ *Attend the conference founded by the author:* www.MachineLearningWeek.com

- ➢ *Access the author's online course:* www.MachineLearning.courses

- ➢ *Follow the author:* @predictanalytic *or* www.linkedin.com/in/predictiveanalytics