# AI and Machine Learning

## Day 2: From a fitted line to a decision you can price

Day 2 of 4

Fatih Kansoy
Oxford Summer School

# Outline

# Five questions reopen yesterday's 90% claim

Keep your written answers in front of you; today's material tests every one.

1. What baseline did the 90% model have to beat, and what was its value?
2. A new column makes your validation score jump. What do you check first?
3. Why may the test set be used only once?
4. Accuracy 90.2% and recall 23.3% describe the same model. How can both be true?
5. Why does a high response probability alone not show that calling causes subscriptions?

Yesterday we judged predictions from the outside. Today we open the models themselves.

# Two cases will test every model we open

**A completed taxi trip**

- distance: 1.70 miles
- duration: 14.40 minutes
- recorded fare: $14.20

We predict this fare with a fitted line, and take the model's miss seriously.

**A bank client scored at 0.30**

- one client-campaign record
- fitted response probability: 0.30

What does that number claim, how do we check it, and does it justify a phone call?

Both cases stay open until the last twenty minutes. Everything we build today exists to close them.

# Today's four outcomes move from fit to action

By the end of today you will be able to:

1. read a regression coefficient with its units, its comparison, and its scope
2. explain what a fitted probability of 0.30 claims, and audit that claim with log loss and calibration
3. derive a decision threshold from the costs of the two possible errors
4. say when even a well-audited probability is the wrong decision quantity, and define the uplift that replaces it

Each outcome gets checked against the two opening cases before you leave.

# Today we open the middle of the evidence chain

Yesterday built the outer links: the question, the data, and the discipline of splits and baselines. Today we live in the middle.

> question → data → **loss** → **model** → **evaluation** → **decision** → monitoring

Part I **Predicting a number.** The taxi fare: loss, model, and evaluation with simple and multiple linear regression.

Part II **Predicting a probability.** The bank client: the same additive machinery, rebuilt as logistic regression.

Part III **From probability to decision.** Costs, thresholds, and the point where prediction stops being enough.

# First fix the taxi task: predict a completed fare

You met the prediction contract yesterday. Here it is, completed for the fare task.

| Field | Entry |
|---|---|
| Unit | one completed New York yellow-taxi trip |
| Population | completed city trips like those in the file |
| Target and horizon | the recorded fare, in dollars, for that finished trip |
| Information cutoff | after the trip ends: we predict a fare that already exists |
| Features | recorded trip information, including realised distance and duration |
| Output | a number, in dollars |
| Evaluation | MAE and RMSE on held-out trips, against the training-mean baseline |

The sample: 50,000 completed trips I selected from the January 2024 records.

A pre-trip quotation would be a different contract: realised duration would not exist yet. That distinction returns later today.

## The meter makes a linear model unusually plausible

New York yellow-taxi fares come from a meter:

- an initial amount when the trip starts
- a set amount per mile
- a set amount per minute when traffic is slow

So the true rule behind our target is close to a line in distance and duration. A linear model is the right first choice here because the rule behind the data, for once, is linear.

> Most prediction problems give you no meter. Treat today's fit as an unusually friendly case, and expect nothing like it later in the week.

# With no features, squared loss chooses the training mean

Suppose I hide every feature and you must quote one number for every trip. Which number should you pick? Under squared error, the answer:

$$\widehat{c} = \arg\min_{c} \sum_{i=1}^{n} (y_i - c)^2, \qquad -2\sum_{i}(y_i - \widehat{c}) = 0 \implies \widehat{c} = \bar{y}.$$

$c$: one constant guess quoted for every trip · $y_i$: trip $i$'s recorded fare · $\bar{y}$: the training average

In plain words: at the mean, moving your guess up or down can only add penalty. For our training trips, $\bar{y} = \$18.3521$.

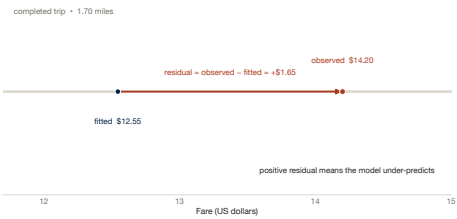
This is the **mean baseline**, the numerical twin of yesterday's training response-rate baseline. Any model with features must beat it on new trips.

# The line misses the opening trip by $1.65

Here is the fitted line; how it was fitted comes in a moment. First, read its miss on the opening trip:

$$\widehat{\text{fare}} = 6.3357 + 3.6529 \times \text{distance}$$

$$\hat{y} = 6.3357 + 3.6529(1.70) \approx \$12.55$$




Residual: $14.20 - $12.55 = +$1.65.

6.3357: fitted fare at zero miles · 3.6529: dollars per recorded mile

A positive residual means the model guessed low.

**Was the model wrong, or was this trip unusual? Hold your answer.**

# Squaring makes large misses dominate the fit

Before fitting anything more, we must decide how to score a miss.

| error | absolute penalty | squared penalty |
|---|---|---|
| \$1 | 1 | 1 |
| \$5 | 5 | 25 |

- under total absolute error, a \$5 miss costs five times a \$1 miss; under total squared error, twenty-five times
- squaring stops positive and negative misses cancelling
- squaring gives the computer a smooth objective to minimise

This is a choice with consequences: different penalties can prefer different predictions, as yesterday's mean-versus-median example showed.

# Squared loss makes the conditional mean the target

Everything in Part I is machinery for estimating one object. If squared error is the penalty, the best possible prediction for a trip with features $\mathbf{x}$ is:

$$m(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

$m(\mathbf{x})$: the average outcome among cases with features $\mathbf{x}$ · $Y$: the fare of a new trip · $\mathbf{X}$: its features

- $m(1.70 \text{ miles})$ is the average fare over all 1.70-mile trips
- no prediction rule beats this conditional mean on expected squared error; Appendix K gives the proof
- a linear model estimates its best straight-line approximation; here the meter makes that approximation unusually close

# At the least-squares minimum, two imbalances vanish

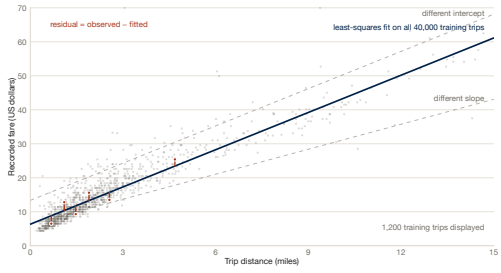
To "fit" is to search over an intercept and a slope:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\beta_0$: intercept · $\beta_1$: slope · $x_i$: trip $i$'s distance

At the minimum, two conditions hold (Appendix A):

- $\sum_i e_i = 0$: the misses balance
- $\sum_i x_i e_i = 0$: no co-movement left with distance
- individual residuals can still be large




Vertical squared residuals select the line.

# The slope: co-movement over variation

Solving the two conditions gives both coefficients:

$$\widehat{\beta}_1 = \frac{\widehat{\mathrm{Cov}}(X, Y)}{\widehat{\mathrm{Var}}(X)}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

numerator: how distance and fare move together (miles × dollars) · denominator: how much distance varies (miles$^2$) · ratio: dollars per mile

- the units check themselves: miles-dollars over miles-squared is dollars per mile, exactly what a slope should be
- the slope measures how fare moves with distance, scaled by how much distance varies

This formula covers one predictor. With several predictors the computer solves a system of such conditions; Appendix A shows it.

## A coefficient needs units, comparison, and scope

The fitted slope supports exactly one sentence:

> In these completed trips, one extra recorded mile goes with about $3.65 more recorded fare, along the fitted line.

You must always supply the three parts of that sentence:

- **units**: dollars per recorded mile
- **comparison**: trips that differ in recorded distance, inside this sample
- **scope**: an association, in this population, within the observed distance range

The intercept 6.3357 is the fitted fare at zero miles; read it gently, because zero-mile trips do not appear in the data. And none of this says what would happen if a driver chose to drive one extra mile: that is an intervention question.

# Adding duration nearly halves this trip's residual

$$\widehat{\text{fare}} = 4.1202 + 3.0426 \times \text{distance} + 0.2822 \times \text{duration}$$

3.0426: dollars per extra mile at the same duration · 0.2822: dollars per extra minute at the same distance

- each coefficient now compares trips while holding the other feature fixed
- the meter charges per mile and per minute in slow traffic, so the fitted pair reads like that tariff, split into its two parts
- for the opening trip: $\widehat{y} = 4.1202 + 3.0426(1.70) + 0.2822(14.40) \approx \$13.36$, and the residual shrinks from +\$1.65 to about +\$0.84

One trip proves nothing by itself. The held-out trips will judge whether duration helps in general.

# MAE and RMSE put held-out error back in dollars

The training trips estimate the coefficients; held-out trips judge the predictions. For $m$ held-out trips:

$$\text{MAE} = \frac{1}{m}\sum_i |y_i - \widehat{y}_i|, \qquad \text{RMSE} = \sqrt{\frac{1}{m}\sum_i (y_i - \widehat{y}_i)^2}.$$

$m$: number of held-out trips · $y_i - \widehat{y}_i$: trip $i$'s miss · both metrics are in dollars

- MAE is the average size of a miss
- RMSE squares the misses first, so large misses weigh more
- both compare directly with an actual fare

The baseline plays by the same rules: compute the mean on training data, then score it on the held-out trips.

# Test $R^2$ is a comparison, not an error size

$$R^2_{\text{test}} = 1 - \frac{\sum_{i \in \text{test}} (y_i - \widehat{y}_i)^2}{\sum_{i \in \text{test}} (y_i - \bar{y}_{\text{test}})^2}$$

numerator: your model's squared error · denominator: the squared error of guessing the test mean

- $R^2 = 1$: perfect prediction
- $R^2 = 0$: no better than guessing the test mean
- $R^2 < 0$: worse than that

$R^2$ has no units and no causal content, and it never tells you the dollar size of a miss. Keep MAE and RMSE beside it.

One footnote for the next table: the deployable baseline uses the training mean; the $R^2$ benchmark uses the test mean. That gap puts the baseline's own $R^2$ at $-0.0003$.

# Distance cuts held-out RMSE from $16.12 to $4.16

| model | test MAE | test RMSE | test $R^2$ |
|---|---|---|---|
| training-mean baseline | \$10.65 | \$16.12 | −0.0003 |
| distance only | \$2.49 | \$4.16 | 0.9336 |
| distance + duration | \$1.55 | \$3.70 | 0.9474 |

- distance alone cuts the typical miss to about a quarter of the baseline's
- duration trims a further \$0.46 off the RMSE
- an $R^2$ of 0.9336 from a single feature is the meter at work, and the reason I warned you not to expect this elsewhere

> **With an RMSE of \$4.16, a +\$1.65 miss is smaller than a typical miss. The trip was ordinary, and the model was doing its job.**

# Good predictions leave two uncertainty questions open

**Established on held-out trips**

- $18.3521 is the no-feature baseline
- the slope is about $3.65 per recorded mile, with a stated comparison and scope
- RMSE falls from $16.12 to $3.70

**Still unanswered**

1. How much would the fitted coefficient move in a new sample?
2. How wide should an interval be for one new fare?

The first is coefficient uncertainty. The second adds natural trip-to-trip variation.

# A new sample would move the fitted slope

My slope of 3.6529 came from one sample. A fresh sample of January trips would give a slightly different number; the standard error estimates that sampling wobble. A large-sample approximation gives the familiar interval:

$$\widehat{\beta}_j \;\pm\; 1.96\,\widehat{\mathrm{SE}}(\widehat{\beta}_j)$$

$\widehat{\beta}_j$: the fitted coefficient · $\widehat{\mathrm{SE}}$: its estimated sampling variability · 1.96: the 95% multiplier

- read it as a range of slopes that stay consistent with the data, under the method's assumptions
- the SE formula must match how the data were produced (Appendix D)

More data shrinks coefficient uncertainty, not the trip-to-trip spread of fares: that spread is next.

# One new fare needs a wider interval than an average fare

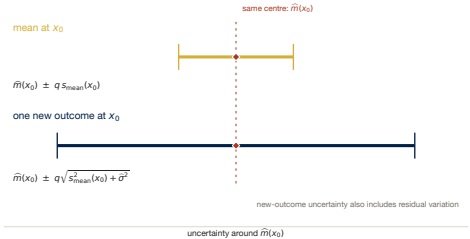
Two questions hide behind one fitted line:

- the **average** fare for trips like this one
- the fare of **your own next** such trip

The second adds trip-to-trip spread on top of uncertainty about the average, so its interval is wider:

$$\widehat{m}(x_0) \pm q\, s_{\text{mean}}(x_0)$$

$$\widehat{m}(x_0) \pm q\sqrt{s^2_{\text{mean}}(x_0) + \widehat{\sigma}^2}$$




A new outcome adds residual variation.

A confidence interval for the average; a prediction interval for one outcome. Appendix I builds one from held-out trips.

# Checkpoint: interpret, compare, and qualify the fare model

Answer each in one sentence:

1. Why is \$18.3521 the correct no-feature baseline under squared loss?
2. Read 3.6529 with its units, comparison, and scope.
3. Which metric tells a passenger the approximate dollar size of a miss, and why is $R^2$ insufficient?

Your answers rely on two conditions: the line approximates the rule, and future trips resemble the training trips. Appendix L separates those assumptions. Next, the output becomes a probability.

# For a 0/1 target, the conditional mean is a probability

- the bank file holds 41,188 client-campaign records from May 2008 to November 2010
- one row is one phone call to one client; the same client can appear more than once
- $Y = 1$ marks a term-deposit subscription: 4,640 records, a rate of 0.1127

Now watch what the conditional mean does with a 0/1 target:

$$\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] = 0 \cdot \mathbb{P}(Y = 0 \mid \mathbf{X} = \mathbf{x}) + 1 \cdot \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = p(\mathbf{x})$$

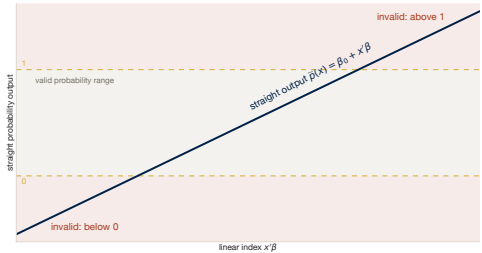
$p(\mathbf{x})$: the share of yes among records with features $\mathbf{x}$

For a yes/no target, the average among comparable records is the share of yes. Predicting the conditional mean now means predicting a probability.

# Probabilities need a bounded output, not a straight line

With no features, we predict the **training response-rate baseline**, 0.1127, for every client. Now add features, and ask what goes wrong if the prediction stays a straight line $\widetilde{p}(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \beta$:

- a line can output probabilities below 0 or above 1
- it moves the probability by the same amount everywhere, near 0.5 and near 0.99 alike; near 0.99 the same push would pass 1




A line can predict below zero or above one.

The linear probability model still earns use in applied work, with care; Appendix E gives its terms. For prediction, we want an output that respects the 0 to 1 range.

# Odds remove the ceiling; log-odds remove the floor

To fix the range problem we need a scale with no ceiling. The first step is odds:

$$\text{odds} = \frac{p}{1-p}, \qquad \text{odds}(0.30) = \frac{0.30}{0.70} = 0.429.$$

$p$: the probability of yes · $1-p$: the probability of no · 0.429: chances for, per chance against

- a probability of 0.5 is odds of 1
- as the probability climbs towards 1, the odds climb without limit: same information, stretched scale
- odds fix the ceiling but not the floor; taking the logarithm fixes that too

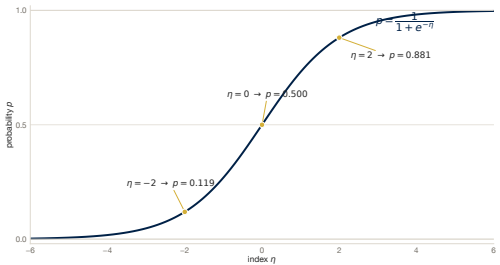
Log-odds run over the whole number line, symmetric around zero. That is the scale we can hand to a line.

# The sigmoid turns a linear index into a valid probability

Put the additive index on the log-odds scale, then solve back:

$$\log \frac{p_\beta(\mathbf{x})}{1 - p_\beta(\mathbf{x})} = \eta(\mathbf{x})$$

$$p_\beta(\mathbf{x}) = \frac{1}{1 + e^{-\eta(\mathbf{x})}}$$




The sigmoid maps $-2$, 0, and 2 to 0.119, 0.500, and 0.881.

$\eta(\mathbf{x}) = \beta_0 + \mathbf{x}^\top \beta$: the additive index · $p_\beta$: the model probability

Every output sits strictly between 0 and 1, and order is preserved: a larger index always means a larger probability.

# A fitted 0.30 is a frequency claim

Take the opening case: the model outputs $\widehat{p}(\mathbf{x}) = 0.30$ for one client-campaign record. The model is saying:

> Among clients whose records look like this one, about 30 in 100 subscribed.

- it does not say which 30
- it does not name this client's outcome
- it is checkable: collect validation records scored near 0.30 and count the yes outcomes

That check has a name, calibration, and we run it before Part II ends.

# Logistic coefficients multiply odds, not probabilities

Hold the other included features fixed. A one-unit change in feature $x_j$ adds $\beta_j$ to the log-odds, which multiplies the odds:

$$\Delta \log \frac{p_\beta}{1 - p_\beta} = \beta_j, \qquad \frac{\text{odds}(x_j + 1)}{\text{odds}(x_j)} = e^{\beta_j}.$$

$\beta_j$: the shift in log-odds per unit of $x_j$ · $e^{\beta_j}$: the multiplier applied to the odds

- for a categorical feature, one level serves as the reference group, and each coefficient compares a level against that reference
- logistic coefficients speak in odds multiples, and odds multiples are not probability changes

How far apart the two can sit is what we look at next.

# The 19.27 odds multiple is descriptive, not causal

In a descriptive univariate fit on all 41,188 records, outside the modelling split: clients whose previous campaign ended in success carry

$$e^{\widehat{\beta}} \approx 19.27$$

times the response odds of the reference group, clients with no previous campaign outcome.

Read it with the same discipline as the taxi slope:

- an odds comparison, in that stated sample and specification
- not a 19-fold probability change
- not the effect of running a successful campaign

What an odds multiple does to a probability comes next.

# One odds multiplier moves different probabilities differently

Suppose a coefficient gives $e^{\beta_j} = 2$, so the odds double. Watch the probability:

| starting probability | new probability | change |
|---:|---:|---:|
| 0.10 | 0.182 | +0.082 |
| 0.50 | 0.667 | +0.167 |

Same coefficient, different probability movement: the sigmoid is flat near the ends and steep in the middle, so where you start decides how far you move. The derivative behind this table is in Appendix G.

> **This is why 19.27 must never be read as "19 times more likely".**

# Interpreting 0.30 leaves fitting and validation unresolved

**We can now decode the output**

- 0.30 means about 30 responses per 100 comparable records
- $e^{\beta_j}$ is an odds multiplier
- its probability change depends on the starting probability

**The evidence is still missing**

1. Which loss fits the coefficients?
2. Which candidate wins on validation?
3. Does 30% behave like 30%?

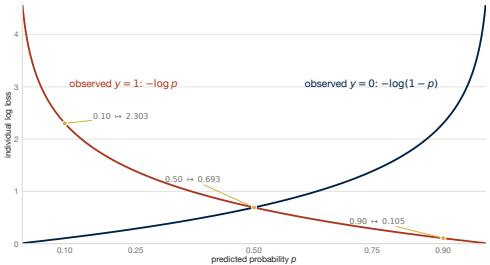
The rest of Part II answers those three questions in order.

# Log loss charges the probability assigned to the truth

Squared error fitted the fare line; a probability needs its own penalty. You report 0.90, the client says no: you gave the truth 0.10, so your penalty is $-\log(0.10) = 2.303$.

> **Log loss** is minus the log of the probability you assigned to what actually happened.

for an observed yes: $-\log(0.90) = 0.105$ · $-\log(0.50) = 0.693$ · $-\log(0.10) = 2.303$




For an observed yes, the penalty grows sharply as the predicted probability approaches zero.

**Confident and wrong is the most expensive way to be wrong. Lower average log loss is better; it is not a percentage of anything.**

# Minimising log loss is maximum likelihood

A second route to the same objective: ask which coefficients make the observed outcomes least surprising. Each record contributes $p_i$ if $y_i = 1$, and $1 - p_i$ if $y_i = 0$. Multiply, take logs, flip the sign, divide by $n$:

$$\prod_{i=1}^{n} p_i(\beta)^{y_i} \left[1 - p_i(\beta)\right]^{1-y_i} \quad \longrightarrow \quad -\frac{1}{n} \sum_{i=1}^{n} \Big[ y_i \log p_i(\beta) + (1 - y_i) \log \left(1 - p_i(\beta)\right) \Big]$$

left: the likelihood, the joint probability of the observed pattern · right: the average log loss

- the coefficients that minimise log loss are exactly the coefficients that give the observed yes/no pattern its largest joint probability
- no general closed-form solution exists; the computer climbs to the optimum numerically (Appendix F)

# Validation selects the full pre-call model

Both candidates were fitted on the training set, under yesterday's stratified 60/20/20 split: 24,712 training, 8,238 validation, 8,238 test records.

- **compact model**: seven pre-call fields: age, contacts this campaign, previous contacts, the three-month interbank rate (euribor3m), plus contact type, month, and previous campaign outcome
- **full model**: all 19 valid pre-call fields, 9 numeric and 10 categorical; everything except duration

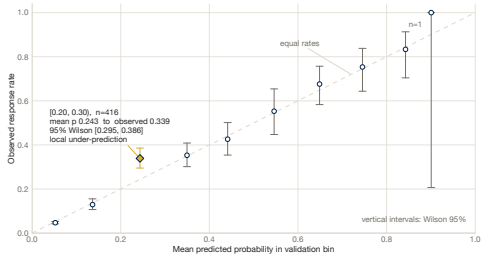
| model | validation log loss |
|---|---|
| training response-rate baseline | 0.352018 |
| compact valid-feature logistic | 0.276960 |
| full valid-feature logistic | 0.271758 |

We select the full model: lowest validation log loss. The test set stays sealed until day's end.

# The 0.30 region is under-predicted on validation

The 0.30 claim is a frequency, so audit it as one: group the validation records by fitted probability, then compare each group's average fitted value with its observed response rate. The bin $[0.20, 0.30)$:

| | |
|---|---|
| records | 416 |
| mean fitted probability | 0.243 |
| observed response rate | 0.339 |
| approximate 95% interval | [0.295, 0.386] |




Mean fitted 0.243; observed response 0.339.

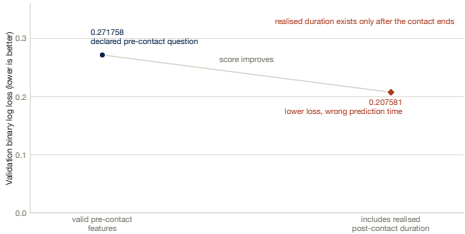
I show you this on purpose. Calibration is an empirical property, checked bin by bin, and no model earns it automatically. Small bins give noisy checks, so read the intervals too.

# Duration improves the score by answering the wrong question

The field **duration** records the length of the final call, and it exists only after that call ends. Add it anyway, and the validation log loss falls:

| model | validation log loss |
|---|---|
| valid pre-call features | 0.271758 |
| with post-call duration | 0.207581 |

This is Day 1's timeline leakage: the improvement answers an after-the-call question. The UCI documentation says to discard duration.




The lower loss uses post-call information.

> **The score fell, and the model became worse for its purpose.**

# Linear and logistic regression share a recipe, not a scale

Both models compute the same additive index $\eta = \beta_0 + \mathbf{x}^\top \beta$. They differ in what they do with it.

| element | linear regression | logistic regression |
|---|---|---|
| target | numerical $Y$ | binary $Y \in \{0, 1\}$ |
| output | $\widehat{y} = \eta$ | $\widehat{p} = 1/(1 + e^{-\eta})$ |
| fitting loss | squared error | binary log loss |
| coefficient scale | target units | log-odds; $e^{\beta_j}$ multiplies odds |
| baseline | training mean | training response-rate baseline |
| held-out evidence | MAE, RMSE, $R^2$ | log loss and calibration |

You have now walked the chain's middle links twice: loss, model, evaluation. The recipe did not change; the output scale did.

# Checkpoint: audit the fitted 0.30

Work from claim to evidence:

1. State the frequency claim made by a fitted probability of 0.30.
2. Explain why $e^{\widehat{\beta}} = 19.27$ is neither a probability multiplier nor a causal effect.
3. Name the validation result that selected the full model and the calibration weakness the audit uncovered.

A fitted probability can now be interpreted and audited. Nothing here yet says whether to pick up the phone; that requires a decision problem.

# Zillow shows why prediction error is not decision cost

> Zillow's house-buying arm used price predictions to make cash offers on homes. On 2 November 2021 Zillow announced it would wind the business down. The announcement included a workforce reduction of about 25%. Inventory write-downs for 2021 totalled $408 million.

My reading of the mechanism, offered as interpretation:

- under-price a house and you lose the deal to another buyer
- over-price it and you win the house, together with the loss

The prediction errors can be symmetric while the costs are not.

The same miss size can have opposite business consequences; the cost asymmetry becomes a formula next.

# Classification and intervention ask different questions

**Classify an existing state.** The fact is already fixed; you decide how to treat it.

- example: a transaction already made is fraudulent or it is not; you decide whether to send it for review
- a trustworthy probability plus the costs of the two errors can settle this

**Choose an intervention.** The action is meant to change the outcome.

- example: an offer designed to keep a customer who would otherwise leave
- you need to know what the action *changes*, and no rescaling of a response probability contains that

We take classification first. The intervention problem closes the day.

# Unequal error costs imply a non-0.5 threshold

A question for the room: should a fraud model and a spam filter share the same cut-off?

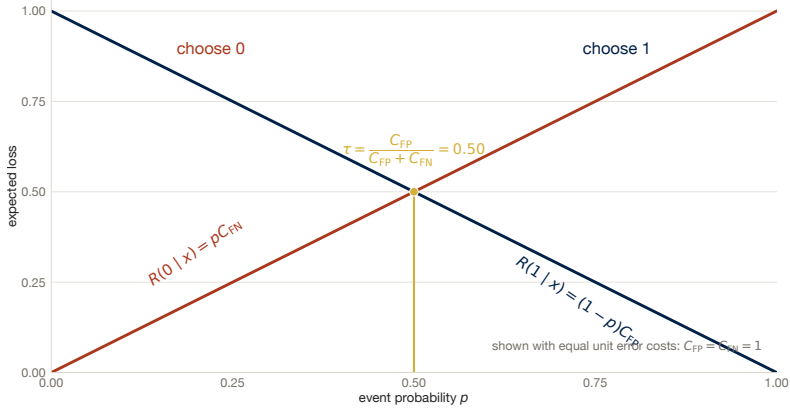
Correct decisions cost nothing; a false positive costs $C_{\mathrm{FP}}$, a false negative $C_{\mathrm{FN}}$. For a case with trustworthy probability $p$, compare the two expected losses and flag when the first is smaller:

$$\underbrace{(1-p)\,C_{\mathrm{FP}}}_{\text{flag it}} \;<\; \underbrace{p\,C_{\mathrm{FN}}}_{\text{wave it through}} \quad\Longleftrightarrow\quad p > \tau = \frac{C_{\mathrm{FP}}}{C_{\mathrm{FP}} + C_{\mathrm{FN}}}.$$

$C_{\mathrm{FP}}$: cost of flagging an innocent case · $C_{\mathrm{FN}}$: cost of missing a real one · $\tau$: the break-even probability

The threshold is the false-positive cost as a share of both error costs. It equals 0.5 only when the two mistakes cost the same.

# The threshold is where the two expected losses cross




Choose the lower-loss action on each side of the crossing.

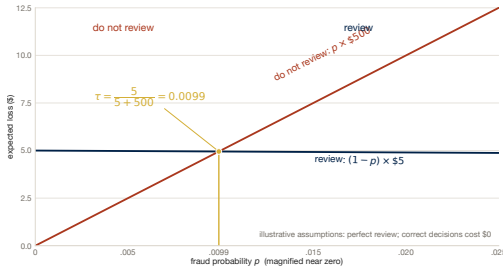
A fraud model and a spam filter need not share a cut-off: their mistakes carry different costs.

# A $500 miss can justify a 1% review threshold

An illustrative fraud setting, with invented costs:

- sending a legitimate transaction to review costs $5 of handling and customer friction
- missing a fraudulent one costs $500
- review catches the fraud it inspects

$$\tau = \frac{5}{5 + 500} = 0.0099$$




Costs of $5 and $500 set $\tau = 0.0099$.

Under these costs we review any transaction scored above about 1%.

Change the costs, change the threshold; the probability model can stay exactly the same.

# Thresholds belong to decisions, not model classes

| domain | false positive | false negative |
|---|---|---|
| marketing | a wasted call | a missed subscriber |
| fraud | a blocked legitimate payment | a stolen payment |
| medical screening | an unnecessary follow-up | a missed diagnosis |
| credit | a rejected safe borrower | an approved risky one |

Real institutions publish exactly such policies: US base credit scores run from 300 to 850, and lenders have commonly treated scores below about 620 as subprime. Those cut-offs are lender policy, not a property of the score.

One model class can serve all four rows. Its threshold does not travel between them.

# Under the stated call costs, 0.30 clears a 0.048 bar

Now answer the opening case with hypothetical bank costs, invented for illustration: a wasted call costs 4 units; a missed subscriber costs 80.

$$\tau = \frac{4}{4 + 80} = \frac{4}{84} \approx 0.048$$

Our client's 0.30 sits far above 0.048: under these costs, call.

Recall Day 1: at the 0.50 threshold, the model found only 23.3% of the subscribers. Now you can see what went wrong:

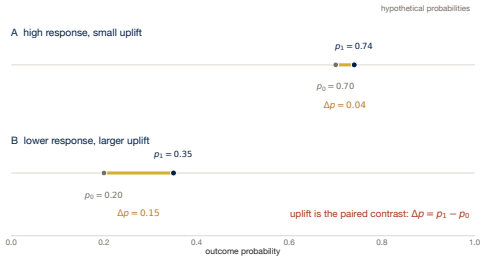
- nobody had costed the errors, and 0.50 was never a law
- with costs like these, the rational bar sits near 0.05, flags many more clients, and misses fewer subscribers

# Response and uplift compare different worlds

Even a calibrated 0.30 describes what clients like this one did. If calling is meant to *cause* a subscription, you need the difference the call makes. For the same client:

$$\Delta p(\mathbf{x}) = \mathbb{P}\big(Y(1) = 1 \mid \mathbf{X} = \mathbf{x}\big) - \mathbb{P}\big(Y(0) = 1 \mid \mathbf{X} = \mathbf{x}\big)$$

$Y(1)$: outcome if contacted · $Y(0)$: outcome if not · $\Delta p$: the **uplift**




Response uses one strategy; uplift compares two.

**A client can sit at 0.30 under both strategies: high response, zero uplift, wasted call.**

# Intervene only when uplift covers incremental cost

Let $b$ be the value of a success under either strategy, and $c$ the extra cost of intervening over the comparison. The value gain from intervening on a client is:

$$b\,\Delta p(\mathbf{x}) - c, \qquad \text{so intervene when} \qquad \Delta p(\mathbf{x}) > \frac{c}{b}.$$

$b$: what one extra success is worth · $c$: what the intervention itself costs · $c/b$: the break-even uplift

- the action pays when the extra successes it causes are worth more than its extra cost
- this simple bar assumes every success has the same value and the action has no side effects; when that fails, write out the full value function instead of the ratio

# At a 0.10 uplift bar, only B and C pay

A hypothetical randomised retention offer, with invented numbers: a success is worth €100 under either strategy ($b = 100$), and the offer adds €10 of cost ($c = 10$). The break-even uplift is $10/100 = 0.10$.

| customer | estimated uplift $\widehat{\Delta p}$ | value $100\widehat{\Delta p} - 10$, in euros | decision |
|---|---|---|---|
| A | 0.03 | −7 | no offer |
| B | 0.12 | +2 | offer |
| C | 0.25 | +15 | offer |

These uplift estimates would have to come from a randomised experiment. The historical UCI campaign file cannot supply them, and Part III ends with the reason why.

# With one offer, choose the largest positive value

Suppose the budget allows one offer. Choose C: the largest positive incremental value, +€15.

Two conditions sit under this choice:

- ranking clients by response probability can order them differently from ranking by incremental value, especially when values or costs differ across clients
- the whole calculation assumes that offering to one customer does not change any other customer's outcome

The general version, choosing the best $K$ offers under a budget, is a small optimisation problem; Appendix M states it in symbols.

# The campaign data support response, not uplift

Keep two claims separate as we close Part III.

**The predictive claim, audited today:**

- the selected model's test log loss is 0.273012, against 0.352018 for the training response-rate baseline
- we opened the test set once, for this number, and for nothing else

**The uplift claim, not available:**

- uplift compares outcomes under two contact strategies
- identifying it needs random assignment, or an observational design with measured confounders and both strategies present at the relevant feature values (Appendix J)
- the campaign file records one strategy as it was run: no comparison, no uplift

> **Response is what these data show. Uplift is what the contact decision needs.**

# Checkpoint: choose the right decision quantity

Diagnose each decision before calculating:

1. For an existing state, derive the error-cost threshold and apply it to the fraud example.
2. For an action meant to change an outcome, state the quantity that replaces response probability and its break-even rule.
3. Explain why the campaign file can audit the first quantity but cannot identify the second.

The distinction is now operational: classification needs a trustworthy probability and costs; intervention needs uplift and causal evidence. The opening cases can close.

# The taxi closes; the call remains conditional

**The $14.20 trip**

- distance only: $12.55, residual +$1.65
- distance plus duration: about $13.36, residual +$0.84
- held-out RMSE: $16.12 → $4.16 → $3.70

The trip was ordinary; the meter made the model's job easy.

**The 0.30 client**

- claim: about 30 in 100 similar records responded
- audit: log loss 0.271758 versus 0.352018, with one under-predicting calibration bin
- at costs 4 and 80: $0.30 > 0.048$, so call

If the call must cause the subscription, response is insufficient: that decision needs uplift and an experiment.

## Netflix paid for accuracy it did not fully deploy

> In September 2009, Netflix awarded a $1 million prize for a 10% RMSE improvement over its own recommendation system. Components of the winning work were used, but Netflix's engineers later wrote that the full winning ensemble was never fully put into production: the extra accuracy did not justify the engineering effort.

That episode belongs at the end of today.

> **A model earns its place by improving a decision at an acceptable cost, and today gave you the tools to say whether it does.**

# Two fitted models, two decision quantities

| Kind | Name | What it contributes |
|---|---|---|
| fitted model | linear regression | a numerical conditional-mean prediction and interpretable coefficients |
| fitted model | logistic regression | a binary response probability through the sigmoid |
| decision rule | cost threshold | an action when a calibrated probability clears the cost-implied bar |
| causal estimand | uplift | the outcome difference between two actions; not identified here |

Mean and prevalence are baselines. MAE, RMSE, $R^2$, log loss, and calibration judge predictions. None is another fitted model.

The model produces an estimate; the decision rule and the evidence determine what may be done with it.

## Five review questions will reopen Day 3

Write short answers tonight; we audit them in tomorrow's first minutes.

1. What does least squares minimise, and why does squared error point at the mean?
2. The slope 3.6529: state its units, its comparison, and its scope in one sentence.
3. A colleague reads $e^{\widehat{\beta}} \approx 19.27$ as "19 times more likely". Correct them, using the doubled-odds table's logic.
4. A false positive costs 4 and a false negative costs 80. Derive the threshold and apply it to a fitted 0.30.
5. Describe a situation where a calibrated response probability still cannot justify the action, and name the missing quantity.

These are tomorrow's opening audit list, exactly as yesterday's questions opened today.

# Day 3 replaces one global line with flexible partitions

The meter handed us a linear world, and the bank's log-odds line worked well enough. One question we cannot yet answer: what happens when the true rule has steps, interactions, and corners that no single line can hold?

Day 3 answers it:

- trees and ensembles: models that cut the feature space into pieces instead of fitting one global index
- flexible models bring a harder evaluation problem, so Day 3 also brings cross-validation
- and metrics matched to the task, including ROC and precision-recall curves

The discipline travels unchanged: baseline first, validation for choices, one test audit.

# Appendix A: differentiation yields the balance conditions

Differentiate the two-parameter objective with respect to each coefficient and set both derivatives to zero; then substitute the first condition into the second and solve:

$$\frac{\partial}{\partial \beta_0} \sum_i e_i^2 = -2 \sum_i e_i = 0, \qquad \frac{\partial}{\partial \beta_1} \sum_i e_i^2 = -2 \sum_i x_i e_i = 0,$$

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\widehat{\mathrm{Cov}}(X, Y)}{\widehat{\mathrm{Var}}(X)}, \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

$e_i = y_i - \beta_0 - \beta_1 x_i$: trip $i$'s residual at the candidate coefficients

These are the two balance conditions quoted in Part I, now with their algebra attached.

# Appendix A continued: several features give normal equations

Define the design matrix $D$: the augmented matrix whose first column is ones, followed by the feature columns. With distance and duration, row $i$ is $(1, \text{distance}_i, \text{duration}_i)$, so $\widehat{\mathbf{y}} = D\widehat{\beta}$.

$$S(\beta) = (\mathbf{y} - D\beta)^\top (\mathbf{y} - D\beta), \qquad \nabla_\beta S(\beta) = -2D^\top (\mathbf{y} - D\beta) = 0 \implies D^\top D\,\widehat{\beta} = D^\top \mathbf{y}.$$

$D$: one row per trip, one column per included term (ones first) · $S(\beta)$: total squared error

Software solves this linear system directly rather than forming an explicit inverse.

# Appendix B: OLS projects outcomes onto the feature space

The normal equations say $D^\top(\mathbf{y} - D\widehat{\beta}) = 0$. With $\widehat{\mathbf{y}} = D\widehat{\beta}$ and $\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}}$,

$$D^\top \mathbf{e} = 0.$$

The residual vector is orthogonal to every included feature column. Geometrically, $\widehat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the column space of $D$: no other linear combination of the included columns sits closer to $\mathbf{y}$ in squared distance.

You met the two balance conditions in Part I; this is their geometry, one condition per column of $D$.

# Appendix C: omitting a relevant variable tilts the coefficient

Suppose the outcome truly follows the two-variable rule, but you regress $Y$ on $X$ alone:

$$Y = \alpha + \beta X + \gamma Z + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X, Z] = 0 \qquad \Longrightarrow \qquad \operatorname{plim} \widehat{b} = \beta + \gamma \, \frac{\operatorname{Cov}(X, Z)}{\operatorname{Var}(X)}.$$

plim: the value the estimate settles at as the sample grows · $\gamma$: the omitted variable's own effect · $\operatorname{Cov}(X, Z)$: how the omitted variable moves with the included one

Bias appears when $Z$ affects $Y$ and moves with $X$. The short regression can still predict well; its coefficient just stops meaning what a careless reader wants it to mean.

# Appendix D: the variance formula must match the data

Under errors with constant spread and no correlation, the first formula applies; heteroskedasticity replaces it with a sandwich estimator:

$$\mathrm{Var}(\widehat{\beta} \mid D) = \sigma^2 (D^\top D)^{-1}, \qquad \widehat{\mathrm{Var}}(\widehat{\beta}) = (D^\top D)^{-1} D^\top \widehat{\Omega}\, D\, (D^\top D)^{-1}.$$

$\widehat{\Omega}$: squared-residual information, with finite-sample variants in practice

- repeated clients, clustered groups (observations that move together), and time dependence each need their own correction
- robustness to one problem does not buy robustness to the others

Before you trust an interval, ask how the data were produced.

# Appendix E: simple probability effects trade away the bounds

For binary $Y$ with probability $p(\mathbf{x})$,

$$\operatorname{Var}(Y \mid \mathbf{X} = \mathbf{x}) = p(\mathbf{x})\big[1 - p(\mathbf{x})\big],$$

so the error spread changes whenever the probability changes: the model is heteroskedastic by construction, and robust standard errors are standard practice with it.

- OLS still estimates a linear approximation to the conditional mean, and the coefficients read as constant probability changes: the model's appeal
- its costs: predictions can leave $[0, 1]$, and the marginal effect cannot vary with the starting probability

# Appendix F: Newton's method fits logistic regression iteratively

For the total negative log likelihood $\ell(\beta)$, the gradient and Hessian have closed forms; Newton's method evaluates them at the current coefficients and updates:

$$\nabla \ell(\beta) = D^\top(\mathbf{p} - \mathbf{y}), \qquad \nabla^2 \ell(\beta) = D^\top W D, \quad W = \operatorname{diag}\{p_i(1-p_i)\},$$

$$\beta^{(t+1)} = \beta^{(t)} - \left[D^\top W^{(t)} D\right]^{-1} D^\top \left[\mathbf{p}^{(t)} - \mathbf{y}\right].$$

$\mathbf{p}$: the current fitted probabilities · $W$: weights largest where $p_i$ is near 0.5

The update repeats until the objective or the coefficients stabilise. You will never run this by hand; knowing the shape of the update is enough.

# Appendix G: logistic probability effects peak near 0.5

For a numerical feature,

$$\frac{\partial p_\beta(\mathbf{x})}{\partial x_j} = \beta_j \, p_\beta(\mathbf{x})\big[1 - p_\beta(\mathbf{x})\big], \qquad \Delta p = \sigma\big(\operatorname{logit}(p) + \beta_j\big) - p.$$

$\sigma$: the sigmoid · $\operatorname{logit}(p)$: the log-odds of $p$ · $p(1-p)$: the local steepness factor

The factor $p(1-p)$ peaks at $p = 0.5$ and vanishes at the ends. That single factor generates the doubled-odds table in the main deck: the same coefficient moves the probability most in the middle of the scale and hardly at all near 0 or 1.

# Appendix H: a penalty restores finite separated coefficients

If some combination of features separates the yes and no records perfectly, the unpenalised likelihood keeps improving as coefficients grow without bound, and no finite maximum exists. An $L_2$-penalised fit restores one:

$$\min_{\beta_0,\beta} \Big\{ \ell(\beta_0, \beta) + \lambda \|\beta\|_2^2 \Big\},$$

$\lambda$: the penalty strength, chosen on validation data · the intercept $\beta_0$ is left unpenalised

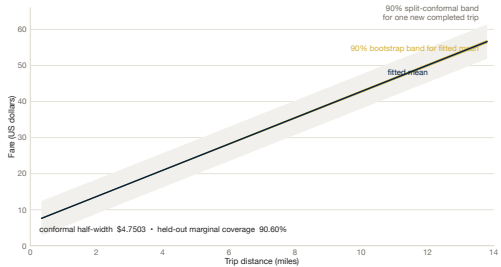
- standardise numerical features first, so the penalty treats every scale alike
- perfect separation is rare in a large file, but small samples and rare categories produce it easily

# Appendix I: calibration residuals form a prediction interval

Fit the model on a proper training sample. On a separate calibration sample, compute the absolute residual scores $s_i = |y_i - \widehat{f}(\mathbf{x}_i)|$, and take:

$$q = s_{(\lceil (n_{\text{cal}}+1)(1-\alpha) \rceil)}$$

$$C_\alpha(\mathbf{x}) = \left[\widehat{f}(\mathbf{x}) - q,\ \widehat{f}(\mathbf{x}) + q\right]$$




A new-trip interval includes residual variation.

$q$: the calibration residual at the required rank

- $\alpha = 0.10$ for a 90% interval

Under exchangeability, coverage is at least 90% for a new observation; no subgroup promise. Taxi sample: half-width \$4.7503, coverage 90.60%.

# Appendix J: uplift requires two comparable strategies

The target is $\Delta p(\mathbf{x}) = \mathbb{E}[Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}]$. Random assignment to specified intervention and comparison strategies identifies it, under three conditions:

- **consistency**: the observed outcome is the assigned strategy's potential outcome
- **no interference**: one client's assignment does not touch another's outcome
- **overlap**: both strategies occur at the relevant feature values

For observational data, a common route additionally assumes:

$$(Y(1), Y(0)) \perp A \mid \mathbf{X}, \qquad 0 < \mathbb{P}(A = 1 \mid \mathbf{X} = \mathbf{x}) < 1.$$

$A$: the strategy received · left: no unmeasured confounding · right: both strategies present everywhere relevant

Train/test separation, however careful, establishes none of this.

# Appendix K: squared loss targets the conditional mean

Write $Y - a = [Y - m(\mathbf{x})] + [m(\mathbf{x}) - a]$. Given $\mathbf{X} = \mathbf{x}$, the first bracket has conditional mean zero, so the cross term vanishes; the variance term does not involve $a$, and the squared term is smallest at $a = m(\mathbf{x})$:

$$\mathbb{E}\big[(Y - a)^2 \mid \mathbf{X} = \mathbf{x}\big] = \mathrm{Var}(Y \mid \mathbf{X} = \mathbf{x}) + \big[m(\mathbf{x}) - a\big]^2.$$

$a$: any candidate prediction · first term: unavoidable spread · second term: your avoidable distance from the mean

Linear regression restricts the predictor to a line; its population target is the best linear approximation:

$$\beta^* = \arg\min_{\beta_0, \beta} \mathbb{E}\big[(Y - \beta_0 - \mathbf{X}^\top \beta)^2\big].$$

# Appendix L: assumptions depend on the inferential purpose

| purpose | requirement |
|---|---|
| unique coefficients | no column is an exact combination of the others |
| useful prediction | the line approximates well; the population stays stable |
| valid uncertainty | a variance method matching error spread and dependence |
| causal interpretation | a real intervention comparison, or an identification strategy |

For the logistic model, the matching list:

- the log-odds form fits reasonably well, and repeated clients are handled
- future clients resemble the validation clients, and calibration is checked rather than assumed

Normality is not required for OLS; squared loss stays outlier-sensitive; extrapolation is unsupported.

# Appendix M: capacity turns uplift into a ranking problem

With at most $K$ interventions available,

$$\max_{a_1,\dots,a_n \in \{0,1\}} \sum_{i=1}^{n} a_i \left(b_i \widehat{\Delta p_i} - c_i\right) \qquad \text{subject to} \qquad \sum_{i=1}^{n} a_i \le K.$$

$a_i = 1$: client $i$ receives the offer · $b_i \widehat{\Delta p_i} - c_i$: client $i$'s expected incremental value · $K$: the budget

- when every client shares the same $b$ and $c$: rank by estimated uplift and take the top $K$ with positive value
- when values or costs differ: rank by $b_i \widehat{\Delta p_i} - c_i$ instead

You applied that rule when you chose customer C.

# References: data, cases, and empirical scope


- New York City Taxi and Limousine Commission. *TLC Trip Record Data*, January 2024 yellow-taxi records; and *Taxi Fare* (rates effective 19 December 2022), the metered fare rules referenced in Part I.
- Moro, S., Rita, P., and Cortez, P. *Bank Marketing.* UCI Machine Learning Repository, CC BY 4.0.
- Zillow Group (2021). "Zillow Group Reports Third-Quarter 2021 Financial Results & Shares Plan to Wind Down Zillow Offers Operations." Press release, 2 November; and Zillow Group, Inc. (2022), Form 10-K for fiscal year 2021, SEC filing.
- Netflix Prize (2006 to 2009), Grand Prize awarded 21 September 2009; Amatriain, X., and Basilico, J. (2012). "Netflix Recommendations: Beyond the 5 Stars (Part 1)." Netflix Technology Blog, 6 April.
- Consumer Financial Protection Bureau, "Borrower Risk Profiles"; myFICO (Fair Isaac Corporation), "What Is a FICO Score?"


The taxi results use 50,000 completed January 2024 trips; the bank results use `bank-additional-full.csv`: 41,188 records, 20 input variables.