

AI · 14-SLIDE SUMMARY

Traditional ML in practice: predicting follow-on funding from Crunchbase data

A step-by-step ML project, from the business question to a model whose top tenth of rounds raises again at 69%, against 42% on average.

From business question to honest evaluation

PART 1 · SLIDE 3 Goal 03 Business question and objectives	PART 2 · SLIDES 4-9 Implementation 04 Twelve-step pipeline 05 Data collection and EDA 06 Label definition 07 Point-in-time features 08 Train/test split 09 Training and tuning	PART 3 · SLIDES 10-12 Results 10 Shortlist performance 11 Drivers of follow-on funding 12 Limitations and next steps	PART 4 · SLIDE 13 Learnings 13 Six lessons for any model
--	--	---	---

After a funding round, will the company raise again within 24 months?

An investor backs dozens of companies and can't support all of them equally. A follow-on forecast tells them where to spend partner time, when to set aside reserves, and who may need help before the money runs out.

OBJECTIVE 1

Beat simple rules

majority class and a one-line rule, on ROC-AUC

OBJECTIVE 2

A useful shortlist

the top 10% raise again clearly more often than average

OBJECTIVE 3

Explain the drivers

signals an investor recognises

52,928

funding rounds (Crunchbase on Kaggle)

31,719

companies

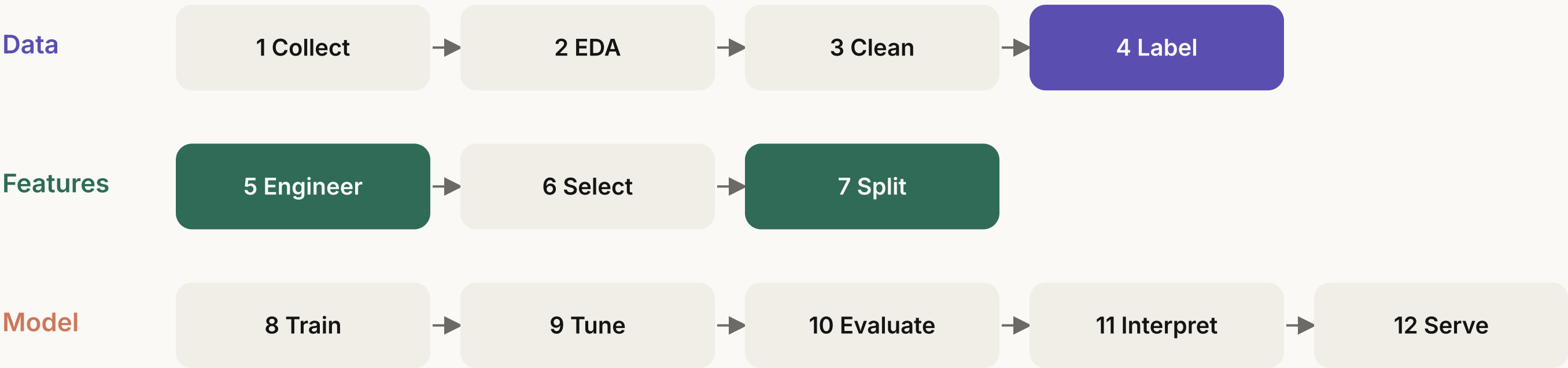
2013

the snapshot ends in December

1 CPU

Python, pandas, scikit-learn

Twelve steps in three phases.



The filled steps decide whether the score can be trusted: the label, the features and the split. Tuning, the step most tutorials dwell on, changed almost nothing.

Collect and explore: two problems hide in plain sight.

DUPLICATION RISK

2.4 investor rows per round

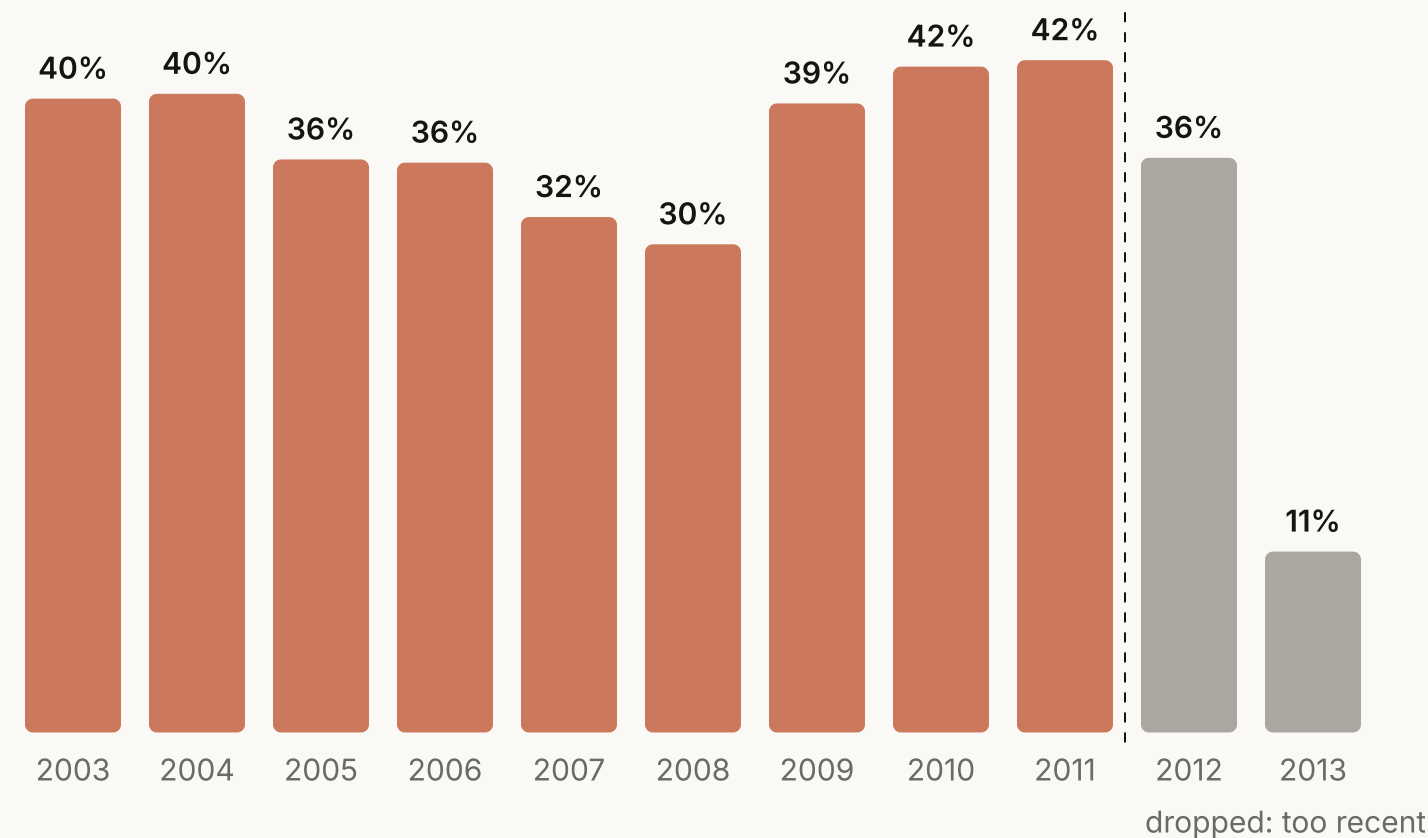
Joining from the investments table would copy each round once per investor, up to 36 times. So the table is built from funding_rounds: one row per round, 52,928 rounds.

TIME

Recent rounds look like dead ends

The data stops in December 2013, so a round from 2013 had no time to be followed. The fall on the right is the end of the data, not a change in startups.

FOLLOWED BY ANOTHER ROUND WITHIN 24 MONTHS, BY YEAR



EDA also showed that Crunchbase's own **is_last_round** flag is wrong for 6,481 rounds that have a later round in the data, so it can't serve as the label.

The label is a product decision.

NAIVE APPROACH

"raises another round, ever"

Counts rounds years later the same as rounds months later, and makes every recent round look like a dead end.

USED HERE

target = 1 if the next round came within 24 months

A fixed horizon makes every round comparable. Rounds after December 2011 are dropped: they don't have 24 months of future.

```
rounds['next_round_at'] = rounds.groupby('object_id')['funded_at'].shift(-1)
rounds['target'] = (rounds['next_round_at'] <= rounds['funded_at'] + pd.DateOffset(months=24)).astype(int)
rounds = rounds[rounds['funded_at'] <= snapshot_end - pd.DateOffset(months=24)] # censoring
```

31,252

labelled rounds after censoring

38%

followed by another round within 24 months

267

rounds dropped in cleaning (no or impossible date)

Features: only what was known on the round date.

POINT-IN-TIME FEATURES (USED)

Stage
rounds_so_far, raised_so_far

Momentum
months since the previous round

Age
company age at the round

The round
amount (log), participants, \$0 flag

Investor
lead investor's prior deals

Context
round type, sector, country

Money on a log scale; dates become ages and gaps; categories one-hot encoded (top 10 values).

THE TEMPTING SHORTCUT: SNAPSHOT TOTALS (EXCLUDED)

funding_rounds, total_rounds, funding_total_usd are company totals measured in 2013, so they include the rounds after the one being predicted.

If a company's total is one round, that round is by definition the last. **A model given them learns Crunchbase's bookkeeping, not investor behaviour.**

SELECTION BECAME A LEAKAGE SCREEN

Out first: anything measured after the round date: snapshot totals, the company's eventual status, milestone and relationship counts. Then 10 point-in-time features remain, so SelectKBest adds nothing.

Test: adding the two totals back lifts ROC-AUC from 0.69 to **0.965**, too good to be true. They stay out.

The split is part of the model.

Naive: random 80/20 by row



copies of one round, and rounds of one company, land on both sides

Used here: split by time



trained on the past, applied to the next rounds: how it would really be used

Fit preprocessing on training years only

Imputation, scaling and the category vocabulary. Fitting them on all rows first would leak test statistics into training.

Base rates differ by period

35% of training rounds were followed up, 42% of test rounds: another reason to judge models on ranking, not accuracy.

Baselines first; tuning comes last and changes little.

ROC-AUC ON 2010-2011 ROUNDS (AXIS FROM 0.40)



18

tuned settings for gradient boosting: learning rate, leaf size, number of trees

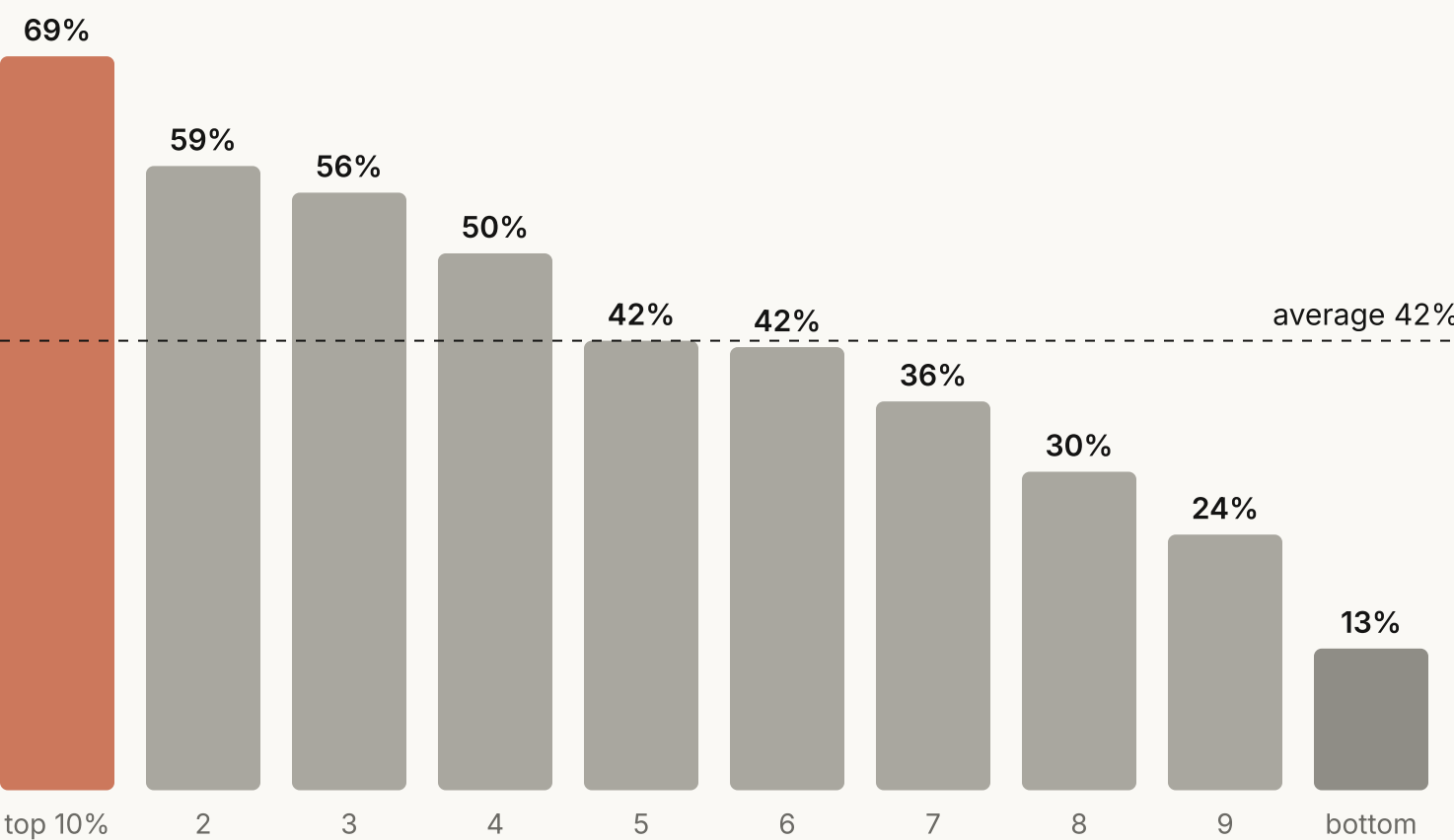
0.689 vs 0.690

best tuned setting against the defaults, on the test years

Tuning moved the score by a thousandth. **The data, the label and the split decide the result.**

The shortlist works: the top 10% raise again at 69%, the bottom at 13%.

FOLLOWED BY ANOTHER ROUND WITHIN 24 MONTHS, BY SCORE DECILE



OBJECTIVE 1 · MET

0.69 vs 0.61

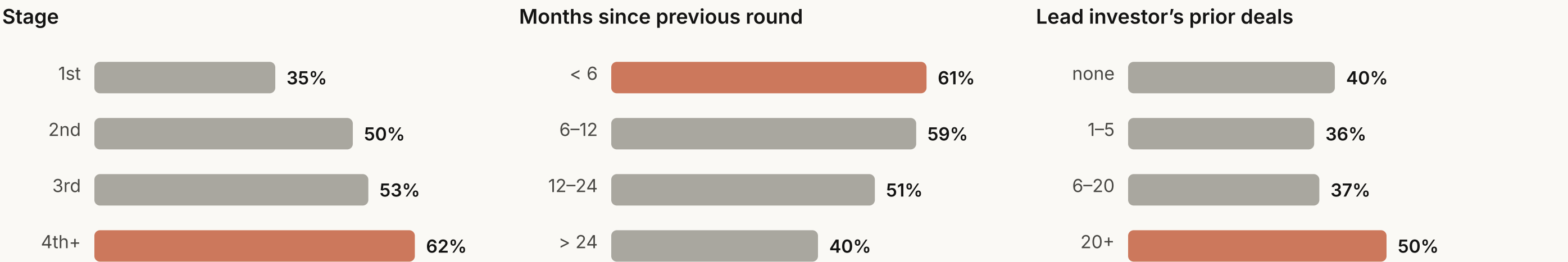
Beats the one-line rule; real but modest. Most of the signal is simple and close to linear.

OBJECTIVE 2 · MET

1.6× lift

Separates “likely to raise, set aside reserves” from “unlikely, look closer” well enough to act on.

Objective 3: stage, momentum and investor track record drive follow-ons.



Later stage keeps raising

35% of first rounds are followed up, against 62% from the fourth.

Momentum compounds

Raised within six months: 61%. After a gap of two years: 40%.

Experienced leads follow up

Lead investor with 20+ earlier deals: 50%, against 36-40% for the rest.

Permutation importance also ranks **company age** and **country** high: US companies raise again at 49% against 26% elsewhere, partly because Crunchbase tracked them better.

A prioritisation aid, not a decision-maker.

COVERAGE

68% of rounds are US

Crunchbase in 2013 was crowd-sourced; small and non-US rounds are under-reported, so “no next round” sometimes means “not recorded”.

MISSING SIGNALS

Money and timing only

No team, product, traction or market: what investors actually judge. That caps the score near 0.69.

AGE OF THE DATA

2010–2011 behaviour

The method transfers to today's data; the learned weights don't.

Next steps: serve the saved pipeline behind an API (raw inputs, no hand-made encodings), calibrate the probabilities, add investor-network and sector-momentum features.

What carries over to every model, rented or built.

01

The label is a product decision

“Raises again eventually” and “raises again within 24 months” are different questions.

02

Point-in-time or nothing

Every feature must be knowable on the prediction date. Add a suspect column back: if the score leaps, it leaks.

03

The split is part of the model

Split the way the model will be used: by time, by entity.

04

Baselines first

A one-line rule scored 0.61; without it, 0.69 means nothing.

05

Tuning comes last

18 settings moved the score by a thousandth. Data, label and split decide.

06

Serving is part of training

Save the whole pipeline, or the model can't leave the notebook.

THE SHORT VERSION

The next round follows **later stages, fresh momentum** and **experienced lead investors**: 62% of fourth-plus rounds raise again (35% of first rounds), 61% when the last raise was under six months ago (40% after two years), 50% when the lead investor has 20+ prior deals (36–40% otherwise).

Together the model's top tenth raised again at **69%**, its bottom tenth at **13%** (average 42%). A moderate signal (ROC-AUC 0.69): use it to triage reserves, not to decide. It can't see team, traction or market, and small and non-US rounds are under-recorded in Crunchbase.

READ huyentran.tech/writings/traditional-ml-in-practice
