FAST-05 Lab Guide: Applied Data Science and Machine Learning for Business Users

Provider: 7formations | Course: FAST-05 | Instructor: Wadood Chaudhary | Five labs, one per session (Lab 5 includes the capstone memo review) Audience: trainees (steps, worksheets) and instructors (setup, answer notes, timing). Instructor-only notes are marked **Instructor**.

All companies, people, figures, and records are fictional. The clinic dataset contains no names, dates of birth, diagnoses, or other PHI. It uses appointment IDs only. Labs run in Microsoft Excel with Copilot and Forecast Sheet (primary), Power BI, Google Sheets with Gemini, or ChatGPT data analysis, per the employer's approved tool, or on paper with a calculator. This is not a coding class; nobody builds a model.

0. Lab environment and safety rules (read before Lab 1)

Tools

- Use the employer-approved analysis tool with a work account, or the 7formations sandbox. Excel with Copilot generally needs the file saved to OneDrive or SharePoint and data formatted as a table; feature availability varies by license and has changed during 2026. The instructor confirms before each cohort.
- Paper fallback: every lab can be done with the printed tables and a calculator. What is scored is interpretation and verification.

Data

- Analysis tools invite uploading whole files. At work: only approved tools, only the columns you need, no personal information (G.L. c. 93H, 201 CMR 17.00), no PHI unless the tool is covered by a HIPAA business associate agreement and policy allows it. De-identification has specific rules (see the HHS reading).
- Class datasets are fictional. If you describe your own work, use general terms.

Fictional organizations

- **Pioneer Valley Mutual Insurance Agency** (Springfield): fraud-flag scores on homeowners claims (Lab 2, Lab 5).
- **Charles River Family Dental** (offices in Newton and Waltham): appointment no-show data (Lab 3). IDs only.
- **Bayside Precision Parts** (60-person CNC shop, Worcester County): demand forecast for part BP-2207 (Lab 4).

Completion

- Submit W1 to W5 and the capstone memo. Meets on every rubric criterion (Appendix B) and 80% on the applied quiz.

Lab 1 (Session 1): Map a prediction

Scenario. Before judging any model, you need to know what it predicts, from what, and who acts. You'll map a fictional example, then one from your workplace.

Goal. Complete W1 for the clinic no-show example and for one prediction or score used in your job (described generally).

Data. Example: Charles River Family Dental's scheduling system shows a "no-show risk" score for each appointment.

Steps

1. **Map the example (10 min).** Fill W1 for the clinic: inputs (lead time, weekday, reminder sent, prior misses), label (attended or not), output (score 0-1), who acts (front desk), action (reminder call above 0.6), cost if wrong (empty chair vs. annoyed patient).
2. **Rules vs patterns (5 min).** Which parts of the clinic's process are rules (48-hour reminder for everyone) and which are learned?
3. **Your prediction (15 min).** Map a CRM lead score, ERP forecast, quality or fraud flag, or staffing forecast you see at work.
4. **Pair review (10 min).** Partner asks: "Who acts on a 0.55? What happens if it's wrong both ways?"

Deliverable: W1 Prediction map

Field	Clinic example	My workplace
What is predicted		
Inputs (features)		
Past outcome learned from (label)		
Output and form (score, class, number)		
Who acts, and what they do		
Cost if it wrongly says yes		
Cost if it wrongly says no		
Data sensitivity notes		

Completion criteria. Both maps complete; both error directions costed in plain words; the person who acts is named by role; no real personal data.

Lab 2 (Session 2): Confusion-matrix card game and threshold memo

Scenario. Pioneer Valley Mutual is piloting a fraud-flag score on 1,000 homeowners claims; 20 were later confirmed fraudulent. Review costs about $60 of adjuster time per flagged claim; missed fraud averages $8,000. Adjusters can review 80 claims a month.

Goal. Complete W2: confusion matrices at two or three thresholds, costs, and a threshold memo tied to business cost.

Data. Card deck (printed or in activities.json) summarizing outcomes:

Threshold	Flagged	Fraud caught (TP)	False alarms (FP)	Missed fraud (FN)	Correctly passed (TN)
0.7	50	14	36	6	944

Threshold	Flagged	Fraud caught (TP)	False alarms (FP)	Missed fraud (FN)	Correctly passed (TN)
0.6	80	17	63	3	917
0.5	120	18	102	2	878
Flag nothing	0	0	0	20	980

Steps

1. **Round 1 (10 min).** Teams deal cards (each card = one claim with a score and true outcome) at threshold 0.7 and tally the four boxes. Check against the table.
2. **Metrics (10 min).** Compute accuracy, precision, recall for each row.
3. **Costs (10 min).** Cost = flagged × $60 + missed × $8,000.
4. **Round 2 (5 min).** Repeat at a threshold your team argues for.
5. **Memo (10 min).** 150 words to the claims manager: decision, metrics in plain words, costs, capacity, review date.

Deliverable: W2 Threshold and cost memo

Threshold	Accuracy	Precision	Recall	Review cost	Missed-fraud cost	Total	Within capacity?

Memo (150 words):

Completion criteria. Metrics correct for at least two thresholds; costs correct; memo recommends a threshold with reasons tied to cost and capacity, in plain language.

Lab 3 (Session 3): Data-quality hunt in the clinic dataset

Scenario. Charles River Family Dental is considering a vendor no-show model trained on its own 2023-2024 appointment history, and the office manager wants to use it to double-book high-risk slots. Before anyone trains anything, you check the data.

Goal. Complete W3: at least three real data problems, each sized, located, and tied to a consequence; plus a fairness note.

Data. Sandbox file `CRFD_appointments_2023_2024.csv` (1,200 rows; fictional; IDs only). Excerpt:

appt_id	office	booked	appt_date	weekday	lead_days	reminder	prior_misses	insurance	no_show
A-10231	Newton	2024-02-01	2024-02-15	Thu	14	N	0	Private	N
A-10232	Newton	2024-02-01	2024-02-20	Tue	19	N	1	Private	Y
A-10233	Newton	2024-02-02	2024-02-06	Tue	4	N	0	Private	N
A-10233	Newton	2024-02-02	2024-02-06	Tue	4	N	0	Private	N

appt_id	office	booked	appt_date	weekday	lead_days	reminder	prior_misses	insurance	no_show
A-10234	Waltham	02/03/2024	03/04/2024	Mon		N	2	Public	1
A-10235	Newton	2024-02-05	2024-02-07	Wed	2	N	0	Private	N
A-10236	Waltham	02/05/2024	02/26/2024	Mon		N	0	Public	0
A-10237	Newton	2024-02-06	2024-03-12	Tue	35	N	3	Self-pay	Y

Summary of the full file (from the sandbox): Newton 1,080 rows (90%), Waltham 120 rows (10%; the two offices see similar volumes in reality); lead_days blank in 112 rows (104 of them Waltham); Waltham rows use MM/DD/YYYY dates and 1/0 for no_show; 18 exact duplicate rows; no Saturday appointments although Waltham opens Saturdays; text reminders started January 2025 (after the data ends).

Steps

1. **Explore (10 min).** Use the tool to profile the file: row counts by office, blanks per column, distinct values in no_show, date formats. (Excel: Copilot "summarize this table," or PivotTable; Power BI: column profiling; ChatGPT/Gemini: ask for a profile of the columns, and show the calculation.)
2. **Verify (5 min).** Confirm one profile number by hand (e.g., COUNTIFS for office = Waltham).
3. **Hunt (15 min).** Record each problem on W3: what, how many, where, consequence, fix.
4. **Fairness (10 min).** Who would be treated unfairly if this model were used to overbook? Which features could act as proxies?
5. **Report (5 min).** Teams report their most damaging finding.

Deliverable: W3 Data-quality findings

#	Problem	Size and location	Consequence for a model	Fix or question for the owner
1				
2				
3				

Fairness note (3-4 sentences):

Completion criteria. 3+ real problems, each quantified and with a consequence; fairness note names an affected group and a lower-harm alternative; one tool number verified manually.

Lab 4 (Session 4): Forecast vs. actuals and override decision

Scenario. Bayside Precision Parts' ERP produced a monthly forecast for hydraulic fitting BP-2207 with an 80% range. The production planner must decide whether to trust July's forecast.

Goal. Complete W4: errors, one AI finding confirmed by hand, a Forecast Sheet (or equivalent) run, and a logged override decision.

Data. Sandbox table `Demand_BP2207`:

Month (2026)	Forecast	Low (80%)	High (80%)	Actual
Jan	1,200	1,050	1,350	1,180
Feb	1,220	1,060	1,380	1,250
Mar	1,250	1,080	1,420	1,310
Apr	1,270	1,090	1,450	1,240
May	1,290	1,100	1,480	1,620
Jun	1,310	1,110	1,510	1,700
Jul	1,330	1,120	1,540	(pending)

Sales note (fictional): "Granite State Hydraulics' new program adds about 350 units/month starting May; confirmed through December."

Steps

1. **Ask the tool (10 min).** Use the improved prompt (activity F05-S4-A3): errors, MAE for Jan-Apr and May-Jun, months outside the range. Ask it to show formulas.
2. **Confirm by hand (8 min).** Recalculate MAE Jan-Apr and the Jan-Jun average actual with your own formulas (=ABS, =AVERAGE).
3. **Forecast Sheet (8 min).** Excel: Data → Forecast Sheet on Jan-Jun actuals (or the employer's tool). Compare its July range with the ERP's. Note: with six points, any statistical forecast is weak; say so.
4. **Decide (10 min).** Override July or not? Log it.
5. **Discuss (9 min).** When is overriding the wrong call?

Deliverable: W4 Forecast override log

Item	Value
AI finding checked (what, AI value, my value, match?)	
MAE Jan-Apr / May-Jun	
Months outside range	
Original July forecast and range	
Override value (or "no override")	
Reason and source	
Approver (role)	
Review date and how you'll judge it	

Completion criteria. One AI finding confirmed manually; errors and outside-range months correct; override decision with reason, source, approver, review date.

Lab 5 (Session 5): Vendor report-card review and capstone memo

Scenario. A carrier offers Pioneer Valley Mutual access to "ClaimGuard Score v3" from Sentinel Analytics (both fictional). Before staff act on the scores, the agency wants questions answered.

Goal. Complete W5 (vendor questions) and present your capstone decision memo (3 minutes).

Data. Fictional model card excerpt:

```
ClaimGuard Score v3 (Sentinel Analytics) - Model card summary
Intended use: prioritize property and auto claims for special investigation.
Training data: 2019-2022 closed auto claims, 6 Midwest carriers (2.1M).
Performance: 95% accuracy on held-out 2022 claims (fraud prevalence 3%).
Features: claim amount, days since policy start, time of loss, prior
claims, ZIP code, repair vendor, report channel.
Fairness: tested for score differences by age band and gender; none
significant.
Limitations (p. 7): not validated for policies in force under 2 years
or for commercial lines. Performance on homeowners claims "expected to be
comparable."
Maintenance: retrained annually.
```

Steps

1. **Read (10 min).** Mark strengths, limits, and gaps (activity F05-S5-A1).
2. **Find the buried limit (5 min).** Each team names one.
3. **Questions (15 min).** Write 6+ vendor questions on W5 covering data fit, metrics at your threshold, fairness and proxies, drift and notification, explanations, and regulatory alignment (MA DOI Bulletin 2024-10).
4. **Monitoring (5 min).** Name an owner, a monthly measure, and an escalation trigger.
5. **Capstone presentations (25 min).** Three minutes each; peers score the explain-it-back for clarity.

Deliverable: W5 Vendor questions list

#	Question	Why it matters to us	Answer needed before go-live?

Monitoring plan: owner / measure / trigger / action.

Completion criteria. 6+ specific questions across the categories; buried limit identified; monitoring plan with owner and trigger; capstone presented.

Appendix B: Rubric (Not yet / Meets / Strong)

Criterion	Not yet	Meets	Strong
Metric interpretation	Metrics wrong or not tied to cost	Correctly ties metric to business cost	Includes capacity and customer impact
Data judgment	Fewer than 3 real problems	Finds 3+ real data problems, sized	Links problems to fairness and proposes fixes
Verification	AI finding not checked	AI finding confirmed by manual calculation	Explains any difference between tool and manual result
Communication	Jargon or unclear	Plain-language explanation a non-expert understands	Peers score it clear; includes next steps and owner