Session 5 Handout: Evaluating Models You Did Not Build

FAST-05 Applied Data Science and ML for Business Users · 7formations · Instructor: Wadood Chaudhary · All examples fictional

Why this session matters

Most AI in small and mid-sized businesses arrives inside software you buy: CRM lead scores, ERP forecasts, carrier fraud scores, scheduling tools. You can't inspect the code, but you can read what the vendor publishes, ask sharp questions, and set up simple monitoring. This session turns the course into a decision memo your manager can act on.

Key concepts

1. Reading a model card

A **model card** (Mitchell et al., 2019) documents a model's intended use, training data, performance (ideally by group), and limitations. Many vendors publish something similar. Read it for:

Section	Ask
Intended use	Is our use inside it?
Training data	When, where, which customers? Like ours?
Performance	Which metrics, at what threshold, on what base rate?
Fairness	Which groups tested? Proxy features?
Limitations	What's buried on page 7?
Maintenance	How is drift detected between retrains?

2. Questions for vendors

1. How does it perform on data like ours (industry, region, product, customer type)?
2. What are precision and recall at the threshold we'd use?
3. Which groups were tested for fairness? Have proxy features been tested?
4. How do you detect drift, and how will we be notified?
5. What explanation can we give an affected customer or employee?
6. How does your governance align with applicable rules (for insurance in Massachusetts, Division of Insurance Bulletin 2024-10 on AI systems)?
7. Can we turn it off or override it, and is that logged?

3. Monitoring and escalation

Keep it simple and owned:

- **Owner:** a named role.
- **Measure:** e.g., monthly flagged vs. confirmed counts; forecast error.

- **Trigger:** e.g., confirmed rate below 15% two months running.
- **Action:** pause, escalate to vendor, revert to manual process.

4. Explaining a prediction

For customers: explain the process, not the model. No scores, no "the AI thinks." Say what happens next, when, and who to contact. For managers: decision first, counts not formulas, costs, and what you'll monitor.

Worked example: ClaimGuard Score v3 (fictional)

Buried limits found: homeowners performance only "expected to be comparable" (not validated); policies under 2 years excluded; 95% accuracy at 3% prevalence (no precision or recall); ZIP code used, but fairness tested only by age and gender; retrained annually with no drift monitoring.

Recommendation (excerpt): pilot 90 days on policies older than 2 years; threshold set to adjuster capacity; vendor to supply homeowners precision/recall and proxy testing before full use; claims supervisor monitors monthly; pause trigger defined.

Explain-it-back (customer): "Like many claims, yours has been selected for a standard review before payment. This doesn't mean anything is wrong. An adjuster will call within 5 business days."

Checklist: before relying on a vendor model

- ☐ Intended use matches ours.
- ☐ Training data comparable, or validation on our data provided.
- ☐ Precision and recall at our threshold.
- ☐ Fairness tested including proxies.
- ☐ Drift monitoring and notification agreed.
- ☐ Owner, measure, trigger, action set.
- ☐ Customer explanation drafted and approved.

Do and don't

Do	Don't
Read the limitations first	Read only the headline metric
Ask for performance on data like yours	Accept "expected to be comparable"
Name a monitoring owner	Rely on the vendor to tell you
Explain process to customers	Tell customers "the AI flagged you"
Write a decision memo	Adopt because competitors did

Try this at work

1. Find the documentation for one AI feature in a system you use.

2. Fill the model-card table above.
3. Send three of the vendor questions to your system owner.

Glossary

Term	Meaning
Model card	Standard documentation of a model's use, data, performance, limits
Validation	Testing performance on data the model hasn't seen, ideally like yours
Monitoring	Ongoing measurement of model performance in use
Escalation trigger	A measurable condition that prompts action
Explain-it-back	Plain-language explanation of a prediction for an affected person

Key takeaways

- Read the limits, not just the headline.
- Ask for precision and recall on data like yours.
- Monitoring needs an owner, a measure, and a trigger.