5 Techniques to Increase AI Adoption Rates via Interpretability

Mike Purewal Fall 2021

Introduction | AI Adoption & Interpretability | MLOps & Interpretability | Conclusion

Outline

1. Introduction

2. AI Adoption

3. MLOps & Interpretability

Claim

- Convincing skeptical users to adopt an AI model requires constant & multi-faceted questions into interpretability
- Answering these questions timely & accurately is a necessary for user adoption
- MLOps is the discipline most closely associated with being able to answer the breadth of questions
- MLOps is essential to the adoption process

Takeaways: 5 Techniques

Support adoption campaigns by:

1. Engender trust demonstrating *robust model performance.*
2. *Baselining* via simple, non-ML models to gain transparency.
3. Tools for post-hoc explainability.
4. Demonstrating model transferability.
5. Tools to communicate model informative-ness.

"It's awesome, all right. Remind me again:
Why did we build this?

Adoption: Definition

- Adoption defined for this talk: The period between **specifying/building** a product and **using** a product.
- Other definitions:
  - Cultural absorption of ML mindset [Google]
  - *Implementation* (not necessarily *use*) of an ML model [McKinsey]
  - Organizational constraints [Gartner]

Adoption: You want me to use what?

How does an organization arrive in a situation where something is built, but there's resistance from end users?

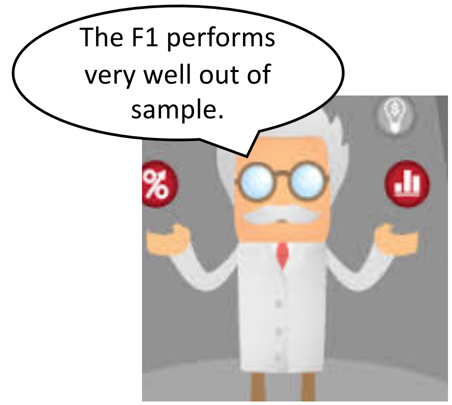
- **Enterprise initiatives & senior management demands**: Purposeful (sometimes) mismatch between strategic vision and end users. Desire for achieving change by 'forcing' use of more advanced tooling.
- **Lack of product/project management involvement**: There is development without clear product specification and communication between technical team and end users. This is usually (not always) facilitated by project or product management (different functions).

Introduction | AI Adoption & Interpretability | MLOps & Interpretability | Conclusion

Interpretability: Lipton[1]

"demand for interpretability arises when there is a mismatch between the formal objectives of [ML] ... and the real world costs in a deployment setting."

[1]Lipton: The Mythos of Model Interpretability



Why MLOps?

- MLOps (machine learning operations) is a practice that aims to make **developing** and **maintaining production** machine learning **seamless** and **efficient**.

- While MLOps is relatively nascent, the data science community generally agrees that it's an umbrella term for *best practices and guiding principles around machine learning – not a single technical solution*. Source: Valohai

Adoption & Interpretability

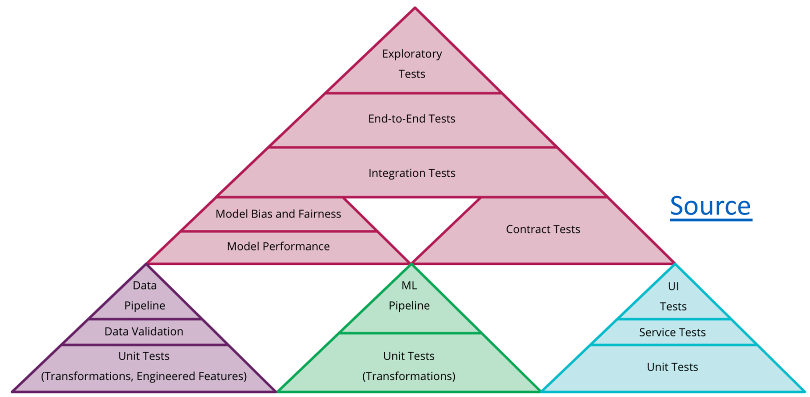
	Common Resistance Point	Interpretability Theme	Proposed Solution
1	The model will produce inaccurate output.	Trust	Demonstration of robust model performance, including data integrity.
2	The model is a black box, no *feel* of how it is operating.	Transparency	Moving towards simulatability, decomposition or algorithmic transparency.
3	The model is a black box, no *feel* of its *output*.	Post-hoc Interpretability	Dashboards showing explainability and drivers of results
4	The model only works in a lab. Non-stationary nature of business (SPACs, crisis, …).	Transferability	Demonstration of performance with increasing generality.
5	The model will replace my value-add.	Informative-ness	Make end users *better* at their jobs by *learning* from the model.

What is trust?

What does *trusting a model mean?*

- Perform well: (to a data scientist) Objective function score
- Deployed well: (to a software engineer) Unit tests
- Governed well: (to data team) Unit tests
- Qualitative: (to an end-user): "well-understood" model
- Achieve results: (to management) "real" objective
- Replace existing BAU: Relinquishing control to model,

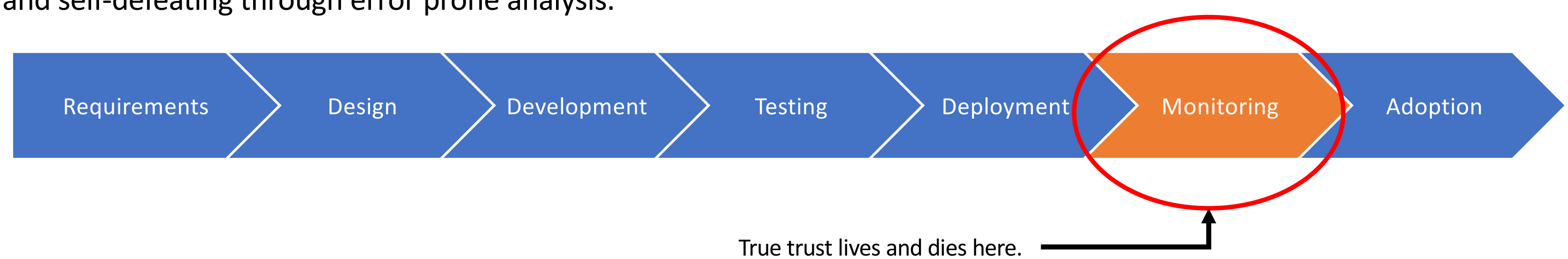
Data, Model, Code: Testing, Testing, Testing



Source

AI Model Lifecycle

- Waterfall development: engagement with stakeholders throughout, but usually 1x interactions pts -> monitoring and adoption are a **continuous touchpoint.**
- Without the proper monitoring infrastructure to support ongoing testing, building trust in an ad-hoc manner becomes onerous and self-defeating through error prone analysis.



Simulatability

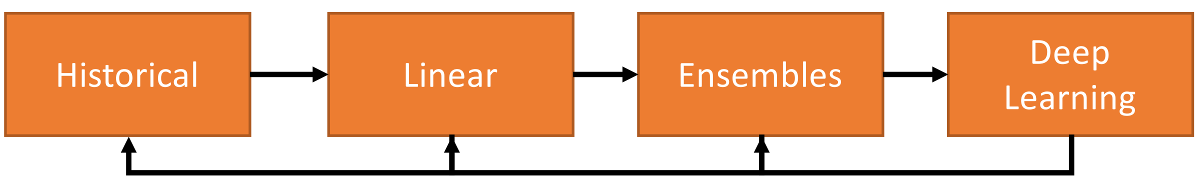
- Strict definition: A person can contemplate the entire model at once.
- Less strict definition I: Simple enough for a person to step through the calculations in a ‘reasonable’ amount of time.
- Less strict definition II: A low-level mechanistic understanding

Composability

- Tendency for “one model”:
  - Total Losses
  - Total Revenue
- Compose multiple models in sequence
- Benefits code maintenance, re-use, debugging, interpretation.

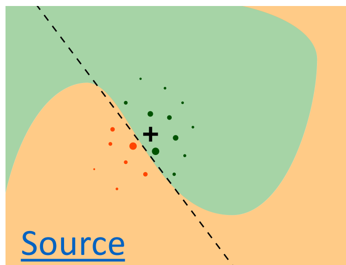
Algorithmic Transparency

- What level of complexity is needed to achieve the objective?
- Transparency ~ 1 / Complexity



LIME

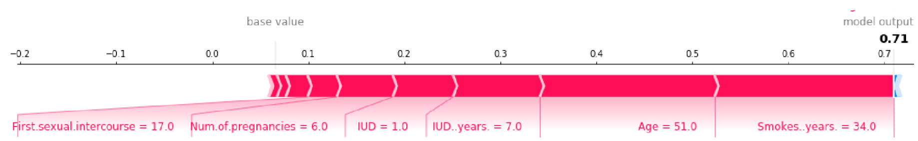
- Local interpretable model-agnostic explanations
- Explainable and sparse ('glass box') model is fit to black box output near specific prediction using simulated data.



SHAP

- SHapley Additive exPlanations
- SHAP is based on the game theoretically optimal Shapley Values.

A prediction can be explained by assuming that each feature value of the instance is a "player" in a game where the prediction is the payout. Shapley values – a method from coalitional game theory – tells us how to fairly distribute the "payout" among the features.



Technique: Permutation Importance

Beware Default Random Forest Importances (link)

Methodology:

- Record a baseline metric.
- Permute values of 1 feature; re-compute metric w/ test samples.
- Importance of feature: re-computed metric – baseline.
- More computationally expensive than the mean decrease in impurity.
- Does not require retraining the model after permuting each column.

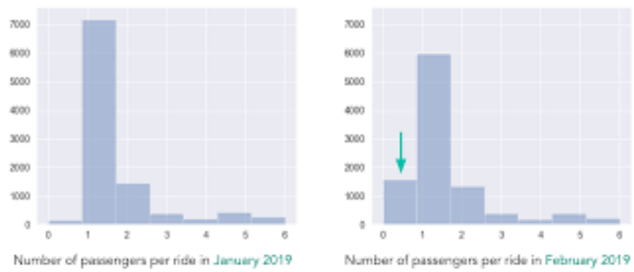
Properties of Explanations

- Symmetry
- Efficiency
- Dummy
- Additivity

Source

Monitoring

- Data Integrity, Data Drift (“X”)

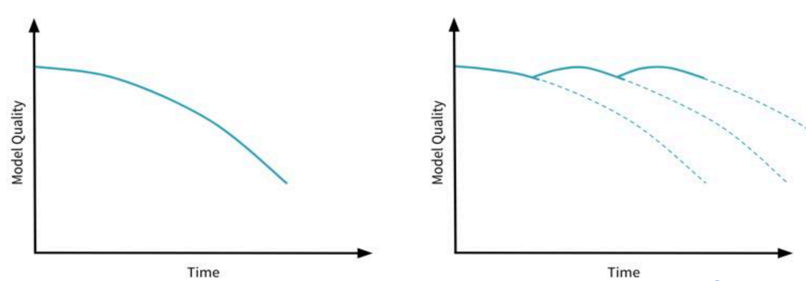


Source

- Performance Shifts / Model Drift / Re-training / Concept Drift (“X->Y”)

Static models

Refreshed models



Source

- Health / Operation Metrics
- Application KPI Performance, Performance by Segment
- Explain-ability
- Governance: MRM, Regulators, Bias/Fairness

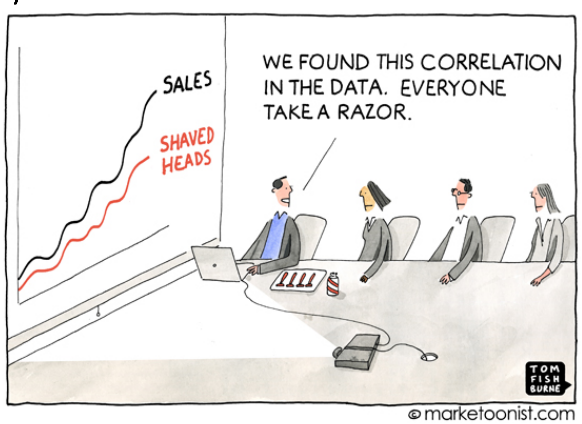
Monitoring Tools

- Splunk
- Great Expectations
- Elastic
- Broadcom AIOps
- Datadog
- AppDynamics
- BigPanda
- Dynatrace
- NewRelic
- databricks

Introduction → AI Adoption & Interpretability → MLOps & Interpretability → 5. Informative-ness → Conclusion

Tywman's Law

- The more unusual or interesting the data, the more likely they are to have been the result of an error of one kind or another. [Wiki, Text (p39)]
- Asymmetric explanation: Positive result->story; Negative result->limitation.
- At-risk cohorts (all me at some point in my career):
  - Inexperienced data scientists,
  - PhD students
  - Sleep deprived employees
  - Domain transfers
  - Anyone excitable



Value-Add: Making Users Smarter

Optimal		Human	
		Correct	Wrong
ML	Correct	x	x
ML	Wrong		

Potentially No Value-Add		Human	
		Correct	Wrong
ML	Correct	x	
ML	Wrong		x

Potentially Insightful		Human	
		Correct	Wrong
ML	Correct		x
ML	Wrong	x	

Drop ML		Human	
		Correct	Wrong
ML	Correct		
ML	Wrong	x	x

- First show model is correct when the human is.
- Once established, an informative model generates examples where *the model is correct*, but the human is **wrong.**
- It should also give an understanding why it performed better.

Introduction | AI Adoption & Interpretability | MLOps & Interpretability | Conclusion

Takeaways

- Convincing skeptical users requires constant & multi-faceted questions into model interpretability

- Answering these questions timely & accurately is a necessary for user adoption

- MLOps is the discipline most closely associated with being able to answer the breadth of questions

- MLOps is essential to the adoption process