causaLens

Why machine learning "succeeds" in development but fails in deployment

info@causaLens.com

www.causaLens.com

causaLens

Causal AI

Why machine learning "succeeds" in development but fails in deployment

The pandemic has exposed weaknesses in conventional machine learning (ML) algorithms, which have been unable to adapt to the new normal. However, even in periods of relative stability, these algorithms are liable to fail. 87% of ML projects in industry never make it beyond an experimental phase, according to one estimate.

Google research recently demonstrated an often-overlooked but key reason for real-world failures of ML systems. The crux of the problem is that large numbers of models perform equally well in training. Some are well-behaved in the real world, but others fail disastrously — and conventional ML pipelines can't tell them apart.

The forty-strong Google team concluded that causality-based methods provide a "promising solution" to the problem of specifying models that perform in deployment. Causal AI pipelines create opportunities for domain experts to constrain models, while causal discovery algorithms zero in on causal predictors, that are resilient against real-world stresses and environmental shifts.

We explore the problem and illustrate why Causal AI makes far better predictions in the real world, not just in the environment under which the model is trained.

The "*underspecification*" problem

The problem (see the Figure below) goes by many names, including "underspecification", "the multiplicity of good models" and the "Rashomon effect" — after the 1950 Kurosawa film which tells the story of four witnesses giving incompatible descriptions of the same incident. Like the witnesses in *Rashomon*, many equivalently good models give incompatible descriptions of the data and make widely varying predictions.

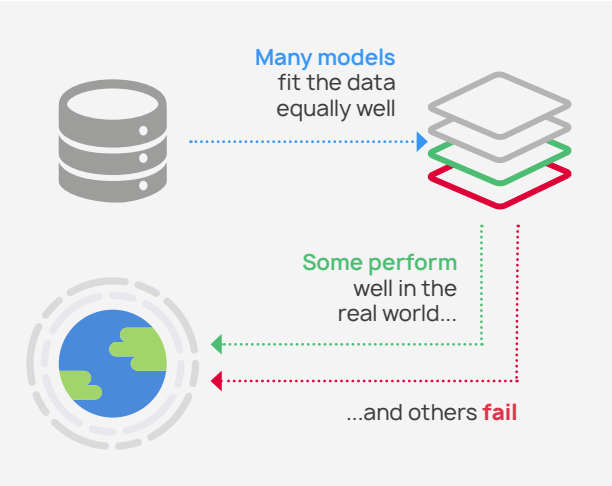

Figure: an illustration of underspecification in conventional ML pipelines

Underspecification is exacerbated by several factors in conventional ML pipelines. Conventional ML models are known to be biased towards learning spurious correlations, which are likely to be fragile under real-world conditions.

> **Worryingly, chance plays a pivotal role in determining whether conventional ML pipelines work in the real world"**

There are also strong theoretical grounds for believing that, when there are many models that are roughly as accurate as each other in training, some will have desirable properties beyond just making good predictions — such as fairness, interpretability and simplicity. However these key properties are often sidelined in conventional ML systems, which tend to produce needlessly complex "black box" models.

Worryingly, chance plays a pivotal role in determining which model is ultimately selected. Researchers have demonstrated that tweaking seemingly irrelevant parameters, such as the random seed value (a randomly generated number that determines how a model is initialized in training), while holding all other aspects of the ML pipeline fixed, leads to totally different models being trained.

Underspecification is not just a theoretical problem. AutoML users and other business users of conventional ML systems often find that the same data science pipelines output completely different models each day. The Google team details how it undermines applied ML systems in the contexts of computer vision, natural language processing and medical diagnostics.

The problem illustrated

Consider an insurer pricing motor insurance premiums.

Actuarial datasets are big and growing, as traditional risk proxies are increasingly augmented with new kinds of data: telematic data and data about the vehicle itself, especially any on-board Advanced Driver Assistance Systems (ADAS).

Typically, insurers rely on "generalized linear models" (GLMs), simple ML models that relax some of the assumptions of ordinary linear regression, to calculate insurance premiums. GLMs are popular because, although simplistic, they are explainable, and so meet regulatory requirements, and are easy for actuaries to interact with.

info@causaLens.com
www.causaLens.com

But GLMs are also underspecified, making them fragile in production. If there are fifty features in the dataset from which five variables are selected, then there are approximately two-million combinations of features that can be used in the GLM. Many of these models will be roughly as good as one another in development, but will give very different pictures of the underlying risks in real-world settings:

Picture 1:

$$risk = \beta_0 + \beta_1 age + \beta_2 driving\_years + \beta_3 driver\_distractedness + \beta_4 claims\_history\_freq + \beta_5 anti\text{-}lock\_braking$$

Picture 2:

$$risk = \beta_0 + \beta_1 age + \beta_2 territory + \beta_3 speed\_vs\_exepcted\text{-}speed + \beta_4 claims\_history\_severity + \beta_5 blindspot\_monitor$$

> **Most insurers rely on ML models that are both too simplistic and underspecified, making them fragile in production"**

A minority of insurers use more powerful ML algorithms, like neural networks, for risk pricing. Insurers have been slow to adopt more advanced algorithms in part because they are "black boxes" that fail to offer the transparency or fairness characteristics that regulators require.

These more powerful ML models suffer from a more extreme form of underspecification. There are vast numbers of ways of parametrizing, say, deep learning models that achieve roughly equal loss in training. Many of these possible models are likely to pick up on misleading correlations in big actuarial datasets.

Causal specifications

There are two broad solutions to underspecification: more testing and more specification. One option for more testing is to retest the equivalently-good models to try to weed out the ones that make errant predictions in deployment. This can be done on data drawn from the real world, assuming it's possible to get access to fresh data. Another option is to conduct "stress tests" — tests that deliberately discard the simplifying assumption that the real world shares an identical distribution with the training data, (which has been called "the big lie of machine learning").

In addition to more rigorous testing, the other possible solution to underspecification is to supply more constraints. As advocated by the Google research group, Causal AI provides promising techniques for creating "causal specifications" to create models that perform in the real world.

One upshot of underspecification is that, as the Google researchers state, "there is a need to find better interfaces for domain knowledge in ML pipelines". Traditional ML pipelines have limited scope for integrating domain knowledge, which is largely confined to feature selection. Causal AI facilitates far deeper integration of domain knowledge: highly intuitive causal diagrams enable experts to convey information that can assist the AI in parameter selection.

> **Causal AI leverages causal discovery algorithms and empowers domain experts to specify robust models"**

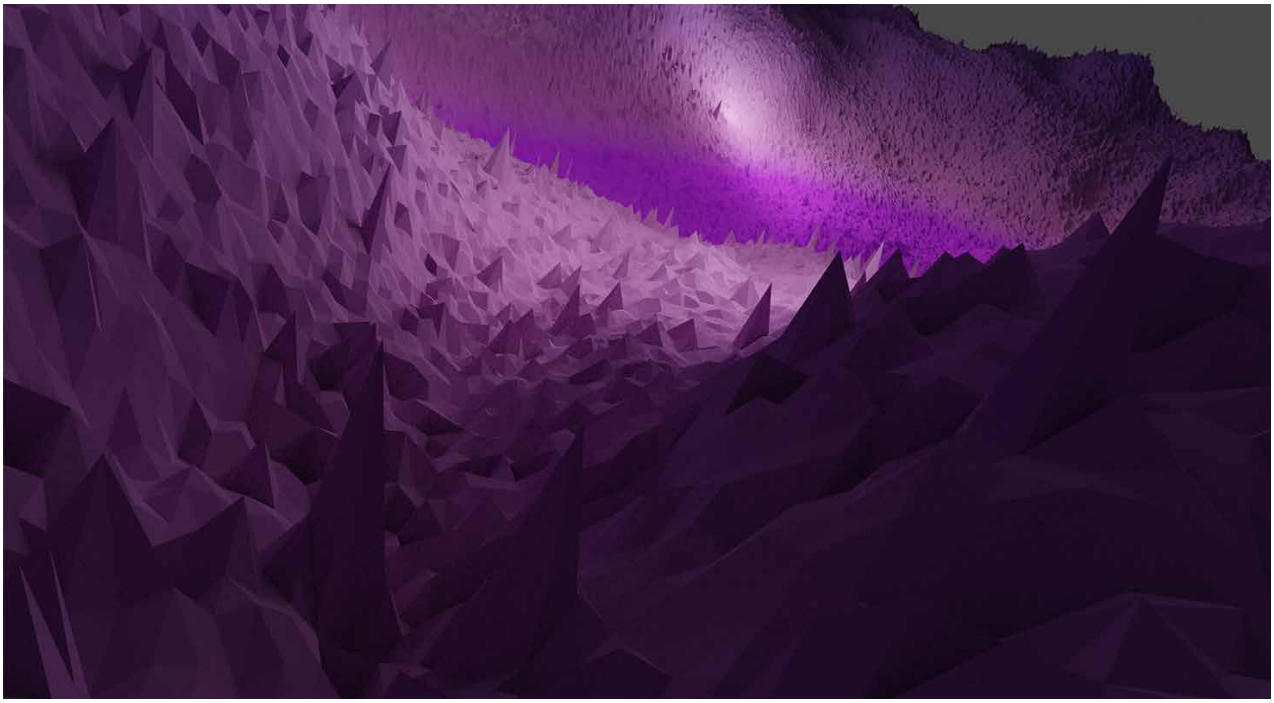

Figure: artwork depicting the "loss landscape" of a deep learning model during training. There are a vast number of ways of weighting the variables in the model that have roughly the same loss (a measure of the amount of error made in the development environment). Photograph: www.losslandscape.com

causaLens

Levels of AI maturity in insurance

Level 0 — No ML

- Regional insurers & tech laggards.
- Use **static pricing rules** that may not have changed in decades.
- Results in adverse selection, leading to unsustainable loss ratios, and lost revenue due to overcharging.

Level 1 — Manual ML

- Late majority (1 in 2 insurers).
- **GLMs.**
- **Moderate underspecification problem.**
- Pricing models are labor-intensive, slow to bring to production and inaccurate.
- Improved loss ratio of new business by 0.8-1.5% over L0, but losing ground to AI leaders and InsurTech.

Level 2 — Auto ML

- Use of representation learning systems (**deep learning and neural networks**) that automate feature engineering.
- **Radical underspecification problem.**
- Models are not explainable, do not incorporate actuarial expertise, and encode spurious correlations and biases.
- Improved loss ratio of new business by 2.1-4.2% over L0, but high failure rates in deployment lead to inconsistent results.

Level 3 — Causal AI

- Being actively explored by AI vanguard.
- **Causal AI** systems make predictions that are explainable, fair, adaptable to new conditions, and incorporate actuarial expertise.
- **Properly specified: models are robust in deployment.**
- Counterfactual analysis improves tail risk estimation.
- Interventional analysis allows Causal AI to model price elasticities and design targeted marketing interventions to prevent churn.

Figure: levels of AI maturity in insurance risk pricing. Impact estimates based on McKinsey analysis.

Returning to the motor insurance example, actuaries are gathering insights into the ways in which ADAS is impacting claims. There seems to be a nascent trend that more cutting-edge ADAS features are reducing claims frequency, by averting accidents, but are increasing claims severity, as the cost of repairing sophisticated on-board computer systems hikes up the total claim amount. Experts can convey this information succinctly to a Causal AI, via causal diagrams, in order to narrow down the search space of model parameters.

More fundamentally, while conventional ML analyses observable data patterns, Causal AI aims to learn the data-generating process: the underlying system of causes and effects that gives rise to observable data. By focussing on underlying causes, the AI disregards incidental and spurious properties of the training data that mislead conventional algorithms.

The causal model that the AI learns is also robust to stresses, perturbations and distribution shifts that inevitably occur in the real world. A large body of research explores the connections between causal models and robustness across different environments. To cite two prominent examples: "invariant causal prediction" is a causal discovery algorithm based on the idea that direct causes are invariant across different environments; and "invariant risk minimization" learns high-level causal representations that generalise in new distributions. "The problem of robustness in its broadest form", as AI luminary Judea Pearl writes, "requires a causal model of the environment".

Conventional machine learning systems fail when the world changes, while Causal AI is more adaptable and robust to change. We've found that Causal AI adapts 3x faster than conventional algorithms during the crisis. However, even when we enter the next normal, AutoML and other conventional ML algorithms will continue to break down, and underspecification will be partly to blame. Causal AI is by far the best technology for properly specifying ML models — using both domain expertise and causal discovery algorithms — in order to work well in the real world, and in a transparent and explainable manner.

About Us

causaLens is pioneering Causal AI, a new category of intelligent machines that understand cause and effect - a major step towards true AI. Its enterprise platform is used to transform leading businesses in Finance, IoT, Energy, Telecommunications and others.

Current machine learning approaches, including AutoML solutions, have severe limitations when applied to real-world business problems and fail to unlock the true potential of AI for the enterprise. For instance, in the case of predictions, they severely overfit and do not adapt when the environment changes. causaLens' Causal AI Platform goes beyond predictions, providing transparent causal insights and suggesting actions that directly improve business KPIs.

causaLens is run by scientists and engineers, the majority holding a PhD in a quantitative field.

Contact us on **info@causaLens.com** or follow us on **LinkedIn** and **Twitter**.