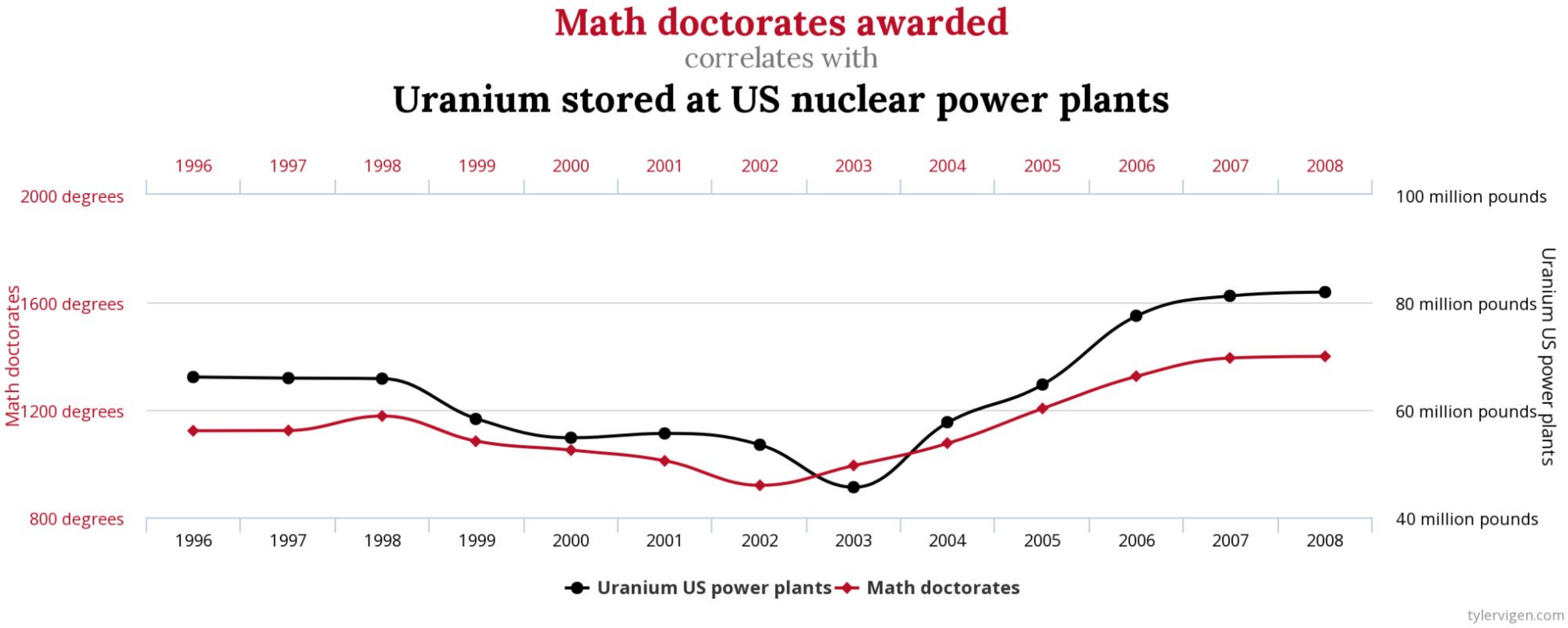
Rising above Machine Learning: The Case of Causality in SME Lending

Siddharth Dixit
Data Scientist (Intern)
causaLens, London



Correlation = 95.23%

Drawbacks of conventional ML

- **Fail under distribution shifts:** Rely on the assumption that the data distribution of the training and test sets are very similar
- **Limited to providing only predictions**: In many decision-making problems, predictions alone are not the best solution [Athey 2017]
- **Black Boxes**:
    - Biased
    - Unfair
    - Unaccountable
    - Fail to pass compliance & regulatory checks

REAL-WORLD EXAMPLES

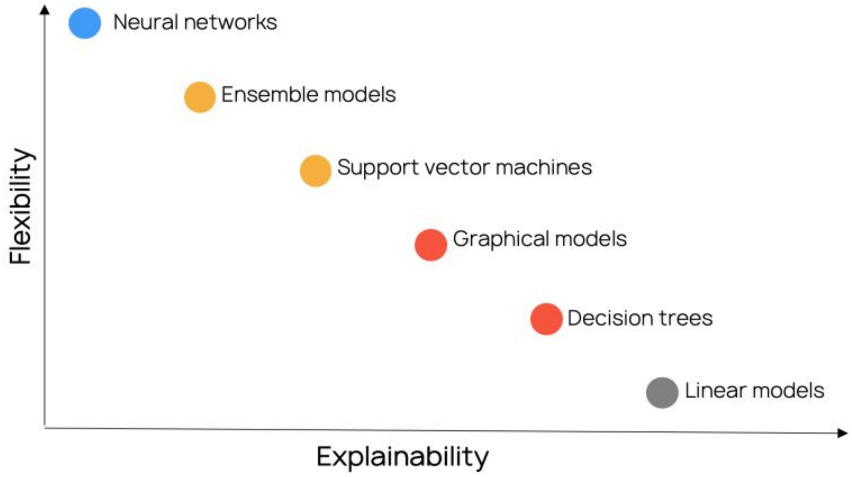
"Over 35% of UK banks have reported that their ML models have been unable to adapt to COVID-19."

BANK OF ENGLAND

WIRED

The Apple Card Didn't 'See' Gender—and That's the Problem

The way its algorithm determines credit lines makes the risk of bias more acute



There is a trade-off between flexibility and explainability in conventional machine learning

Truly Explainable AI: Putting the "cause" in "because"

Causal Inference

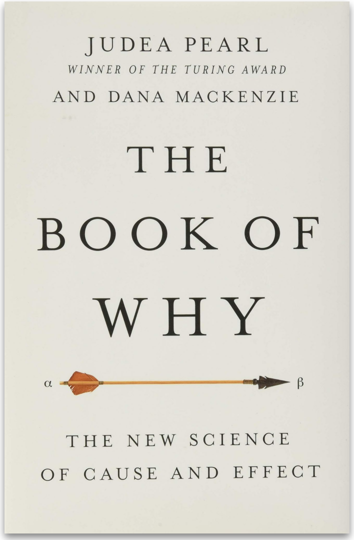
"More has been learned about causal inference in the last few decades than the sum total of everything that had been learned about it in all prior recorded history"- Gary King, Harvard University

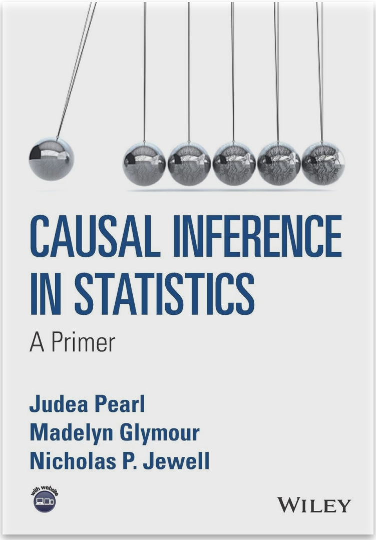
Branch of statistics concerned with the study of cause-and-effect relationships

What does it mean to *cause* something?

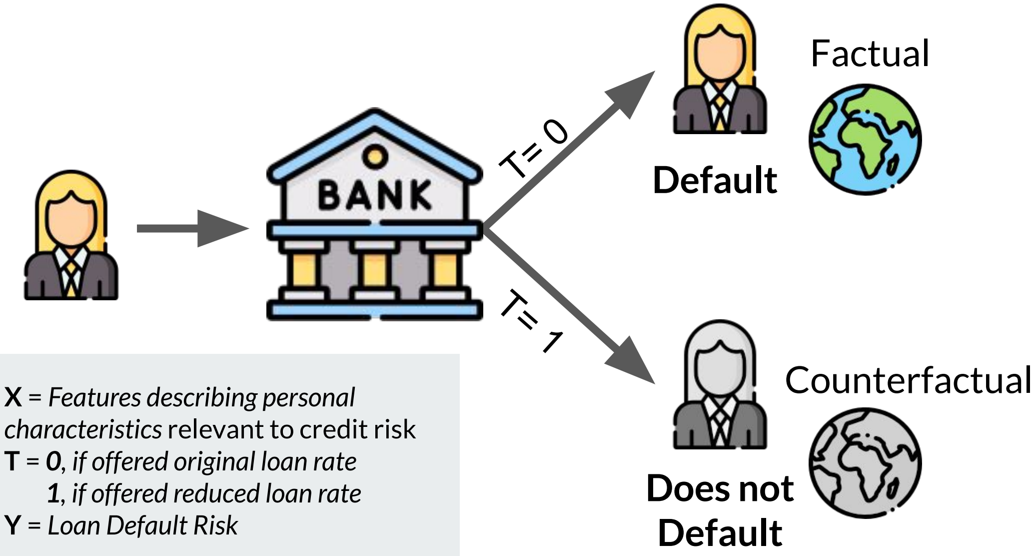
We say Treatment (T) ***causes*** Outcome (Y), if changing T leads to a change in Y while keeping *everything else constant*

The ***causal effect*** is the magnitude by which Y is changed by a unit change in T.





Fundamental Problem of Causal Inference



X = *Features describing personal characteristics relevant to credit risk*
T = **0**, *if offered original loan rate*
1, *if offered reduced loan rate*
Y = *Loan Default Risk*

Person	T	$Y_{T=1}$	$Y_{T=0}$
$P_1(X)$	1	0.3	?
$P_2(X)$	0	?	0.4
$P_3(X)$	0	?	0.9
$P_4(X)$	1	0.4	?
$P_5(X)$	0	?	0.4

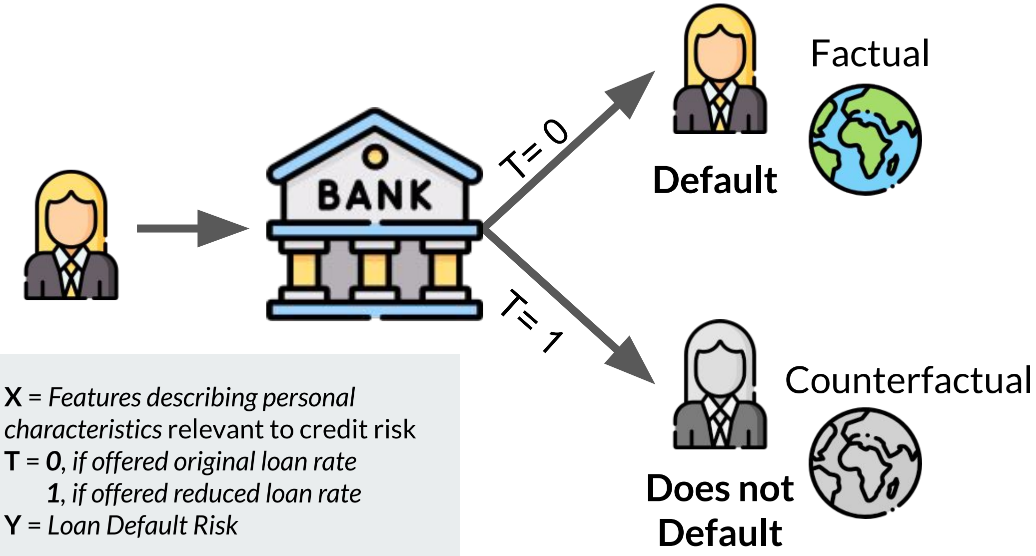
Causal effect of offering reduced loan rate = $\mathbf{E}\,[\,Y_{T=1} - Y_{T=0}\,]$

For any individual, we only observe the outcome in one of the two cases:

- When treatment is *not administered* [Y(T=0)] ~ Factual (Counterfactual)
- When treatment is *administered* [Y(T=1)] ~ Counterfactual (Factual)

Causal Inference boils down to estimating the counterfactual

Fundamental Problem of Causal Inference



X = *Features describing personal characteristics relevant to credit risk*
T = **0**, *if offered original loan rate*
1, *if offered reduced loan rate*
Y = *Loan Default Risk*

Person	T	$Y_{T=1}$	$Y_{T=0}$
$P_1(X)$	1	0.3	0.6
$P_2(X)$	0	0.1	0.4
$P_3(X)$	0	0.6	0.9
$P_4(X)$	1	0.4	0.5
$P_5(X)$	0	0.3	0.4

Causal effect of offering reduced loan rate = $\mathbf{E\,[\,Y_{T=1} - Y_{T=0}\,]}$

For any individual, we only observe the outcome in one of the two cases:

- When treatment is *not administered* [Y(T=0)] ~ Factual (Counterfactual)
- When treatment is *administered* [Y(T=1)] ~ Counterfactual (Factual)

Causal Inference boils down to estimating the counterfactual

The Three Layer Causal Hierarchy

"The Seven Tools of Causal Inference with Reflections on Machine Learning" [Pearl 2018]

Current ML Systems

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

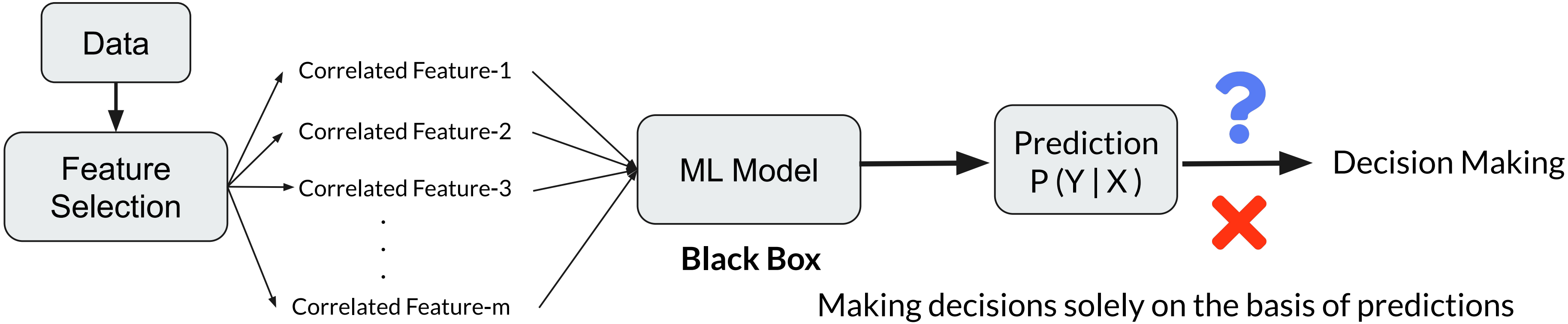
Fig. 1. The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Causal ML

Uses tools from Causal Inference to make ML models:

- More robust to distribution shifts [Causal Discovery]
- More explainable [Causal Discovery]
- More functional for decision making [Treatment Effect Estimation]

Conventional Machine Learning Workflow



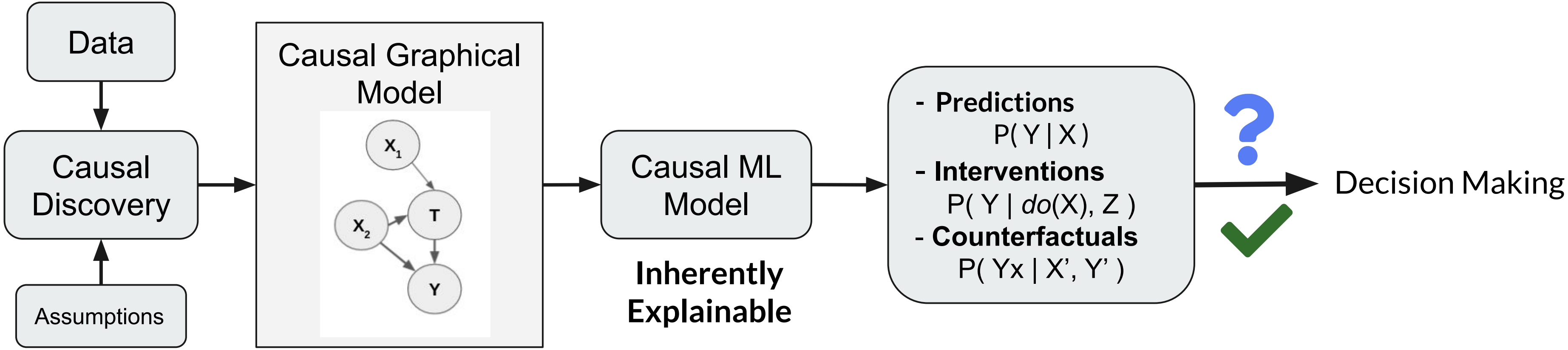
Making decisions solely on the basis of predictions often does not yield the best results!

Causal ML

Uses tools from Causal Inference to make ML models:

- More robust to distribution shifts [Causal Discovery]
- More explainable [Causal Discovery]
- More functional for decision making [Treatment Effect Estimation]

Causal Machine Learning Workflow

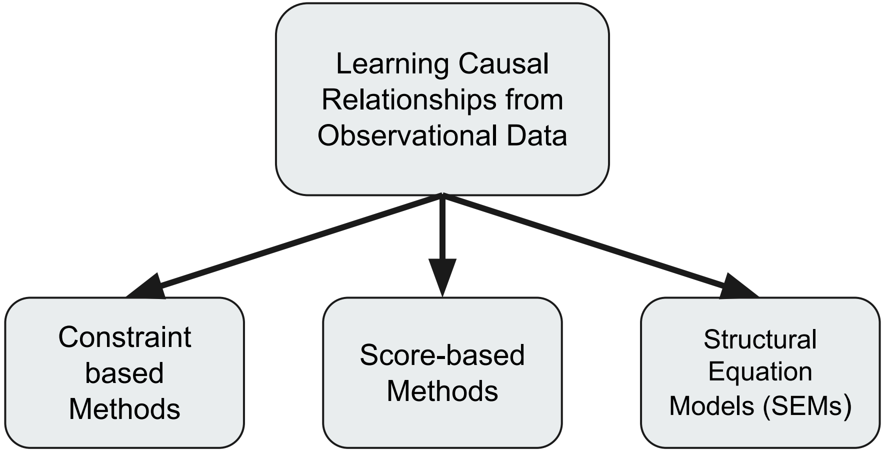


Causal Discovery

"As X-rays are to the surgeon, graphs are for causation." ~ Judea Pearl

The gold-standard of discovering causal relations is to use randomized experiments, which in many cases are too expensive, too time-consuming, or even impossible!

AIM: Identify causal relationships from purely observational or quasi-experimental data.



Causal Discovery

"As X-rays are to the surgeon, graphs are for causation." ~ Judea Pearl

Constraint-based (CB) algorithms

- Learn a set of causal graphs that satisfy the conditional independence embedded in the data.
- Statistical tests are utilized to verify if a candidate graph satisfies the independence based on the *faithfulness assumption*

Faithfulness: Conditional independence between a pair of variables:

$$X_i \perp X_j \mid Z \quad \text{for } X_i, X_j, Z \subseteq X \setminus (X_i, X_j)$$

can be estimated from a dataset X iff Z *d-separates* X_i and X_j in the causal graph G = (V, E) which defines the data-generation process of X .

d-separation: Is a criterion for deciding, from a given a causal graph, whether a set X of variables is independent of another set Y, given a third set Z. The idea is to associate "*dependence*" with "*connectedness*" (i.e., the existence of a connecting path) and "*independence*" with "*unconnected-ness*" or "*separation*".

Peter Clark (PC) Algorithm: Works in a two-step fashion. First, it learns an undirected (skeleton graph) from data. Then, it detects the directions of the edges to return an equivalent class of causal graphs [Peter et al. 2000]

Other popular methods for i.i.d. data include IC algorithm, FCI algorithm (that takes unobserved confounders into account).
Some constraint-based time-series causal discovery methods such as a modified FCI, TiMINO also exist.

Causal Discovery

“As X-rays are to the surgeon, graphs are for causation.” ~ Judea Pearl

Score-based (SB) algorithms

- The faithfulness assumption made by Constraint Based methods can be violated
- To relax the faithfulness assumption, SB algorithms replace conditional independence tests with the goodness-of-fit tests.
- SB algorithms learn causal graphs by maximizing the scoring criterion $S(X, G')$ which returns the score of the causal graph G' given data X.
- Intuitively, low scores should be assigned to the graphs which embed incorrect conditional independence.
- For goodness-of-fit tests, two components need to be specified:
    - **Structural equations:** Often assumed to be linear with additive Gaussian noise, which introduces parameters θ. Each structural equation describes how a variable is causally influenced by its parent variables and a noise term.
    - **Score function:** Maps a candidate causal graph to a scalar based given a certain parameterization of structural equations. Popular scoring functions include BIC, AIC, among others.

Greedy Equivalent Search (GES) Algorithm [Chickering 2003]

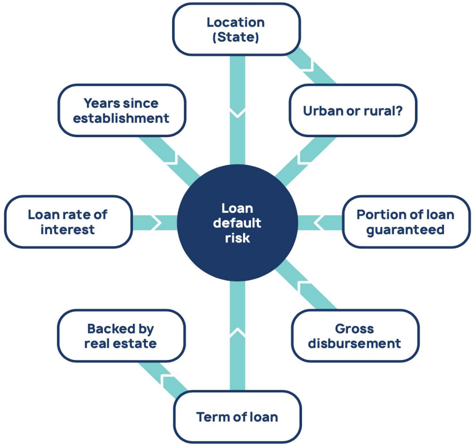
Searching for the causal graph with maximal score is known as structural learning. Since it is not computationally feasible to score all possible causal graphs exhaustively, GES, a two phase procedure is used to reach a locally optimal solution.

Causal Discovery

"As X-rays are to the surgeon, graphs are for causation." ~ Judea Pearl

Structural Equation Models (SEMs)

- Also known as Functional Causal Models (FCMs)
- Offer an easy framework for expressing complex causal relationships and our assumptions about the data generating process.
- Under this framework, a variable X_j can be written as a function of its direct cause P_{X_j} (Parent of X_j) and some noise term $\boldsymbol{\varepsilon}_j$ as $f\,(\,P_{X_j}\,,\boldsymbol{\varepsilon}_j\,)$
- Facilitates evaluation of interventions & counterfactuals
- Popular methods include Linear Non-Gaussian Acyclic Model (LiNGAM), ICA-LiNGAM, among others



A toy causal model for business loan default risk. While edges represent *direct causes*, directed paths represent *indirect causes*.

Backed by real estate = f (Term of loan, $\boldsymbol{\varepsilon}$)

Loan default risk = g (Term of loan, … , Rate of Interest , $\boldsymbol{\varepsilon}$)

Causal Markov Assumption: A node is independent of all its non-descendants given its parents. This gives a nice factorization of the joint distribution.

Estimating Treatment Effects

- The objective for causal inference is to estimate the treatment effect from observational data.
- Treatment effect or the impact of interventions can be measured at the *population*, *treated group*, *subgroup*, and *individual* levels.

Average treatment effect (ATE): Is the treatment effect at the population level:

$$ATE = E[Y(T = 1) - Y(T = 0)]$$

Average treatment effect on the treated (ATT): Is the treatment effect for the treated group:

$$ATT = E[Y(T = 1)|T = 1] - E[Y(T = 0)|T = 1]$$

Also known as *Local Average Treatment Effect (LATE)*

Conditional average treatment effect (CATE): Is the treatment effect at the subgroup level.

$$CATE = E[Y(T = 1)|X = x] - E[Y(T = 0)|X = x]$$

Also known as *Heterogeneous Treatment Effect*

Individual treatment effect (ITE): Is the treatment effect at the individual level.

$$ITE = Y_i (T = 1) - Y_i (T = 0)$$

CASE STUDY: SME CREDIT RISK MODELING

BACKGROUND

- Small and medium-sized enterprises (SMEs) account for **99%** of all businesses.
- They power economies and, in doing so, generate approximately **$850 billion** of revenue for banks each year.
- Yet, relative to big corporates and consumers, SMEs are underserved—the global SME finance gap is **$5.2 trillion**. There are vast untapped business opportunities for any lenders that are up to the challenge of serving SMEs at scale.

Lending organizations—against a backdrop of market volatility and economic crisis—look to maximize *application approval rate* while keeping the *default rate* as low as possible.

Thin-file, high-risk and underserved SMEs present especially big challenges for lenders.

DATASET

RAW

Variable name	Data type	Description of variable
LoanNr_ChkDgt	Text	Identifier – Primary key
Name	Text	Borrower name
City	Text	Borrower city
State	Text	Borrower state
Zip	Text	Borrower zip code
Bank	Text	Bank name
BankState	Text	Bank state
NAICS	Text	North American industry classification system code
ApprovalDate	Date/Time	Date SBA commitment issued
ApprovalFY	Text	Fiscal year of commitment
Term	Number	Loan term in months
NoEmp	Number	Number of business employees
NewExist	Text	1 = Existing business, 2 = New business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise code, (00000 or 00001) = No franchise
UrbanRural	Text	1 = Urban, 2 = rural, 0 = undefined
RevLineCr	Text	Revolving line of credit: Y = Yes, N = No
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Currency	Amount disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan status charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approved loan

[Li et al. 2018]

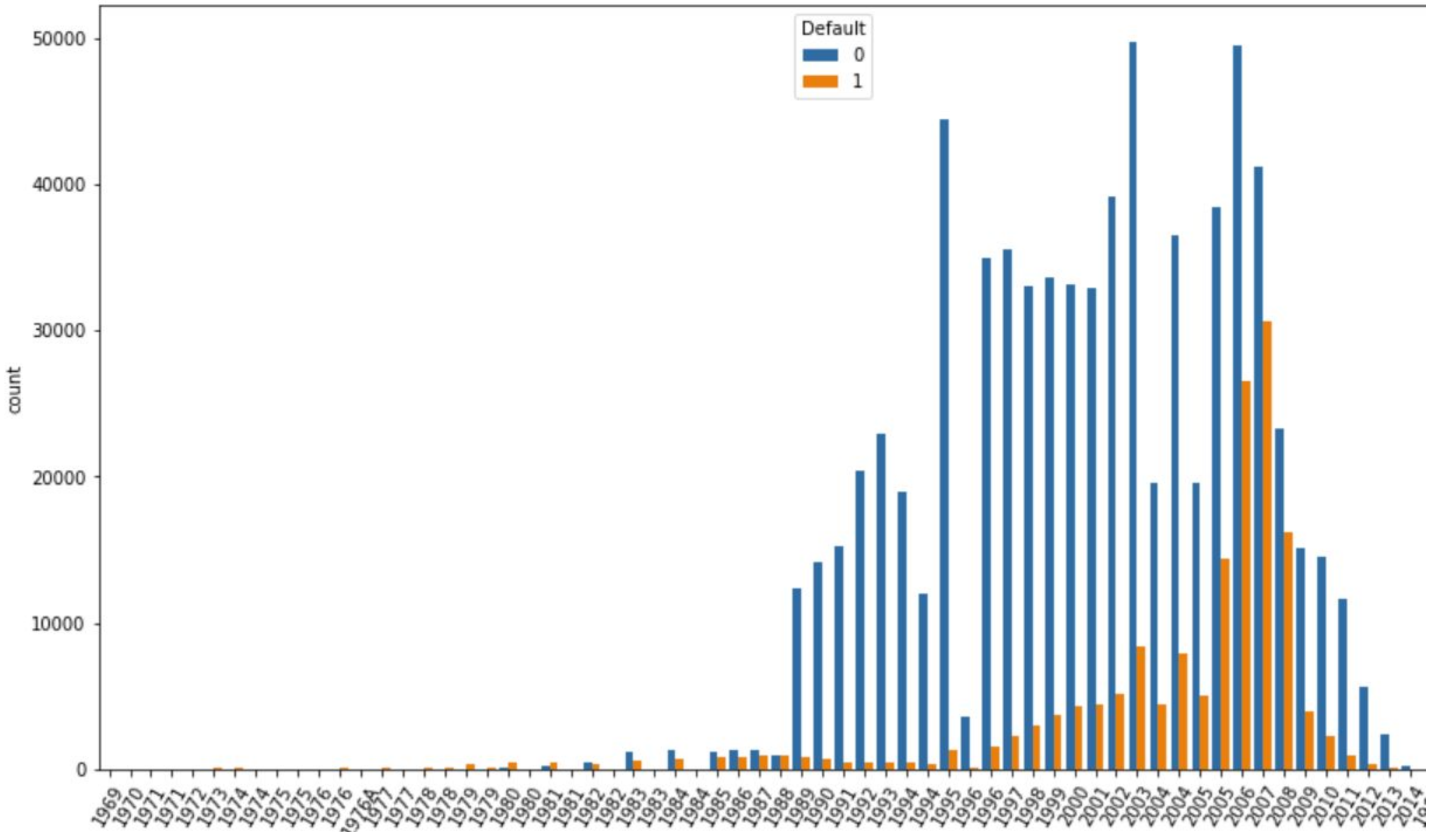
PROCESSED

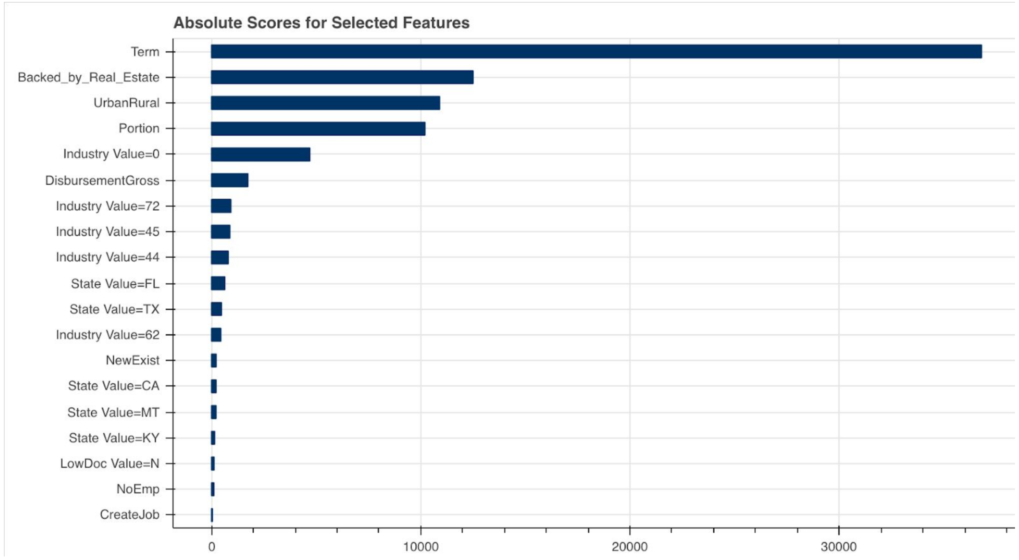
Features: [Index = *Approval Date*]

- *Term of Loan*
- *Backed by real_estate*
- *Urban or Rural*
- *Portion of loan guaranteed*
- *Gross Disbursement*
- *Industry Category*
- *State (Location)*
- *Age of Business*
- *# Employees*
- *# new jobs created in last 12 months*

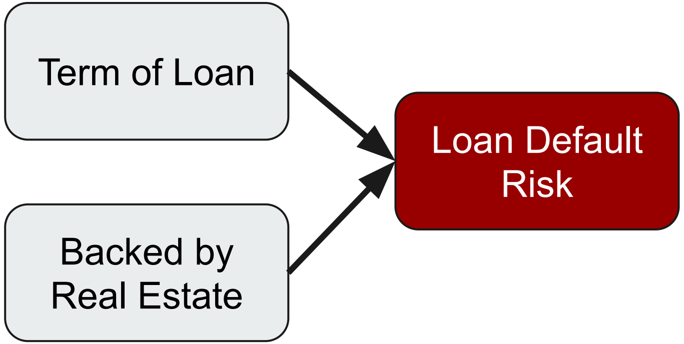
Target: *Default*

US Small Business Administration (SBA) Default Rate over the years





Signal Testing



Observation: *Term of Loan* & *Real Estate Backing* both affect *Loan Default Risk*

Causal Discovery Report

Target	Causal Discovery Method	Discovered Drivers
Default	VAR-NOTEARS	5 out of 19
Time Delta = **0**	p_value=None	(see the table for more details)

About the method:

For more information, please see the documentation.

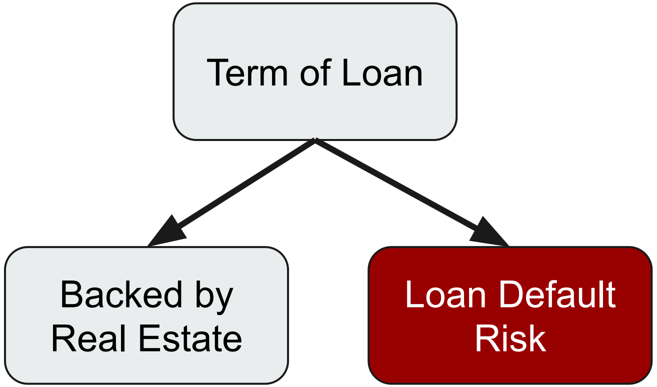
Results

Show 10 entries

Search:

Feature Name	Abs Coef	Threshold
Term	0.229	0.0010
UrbanRural	0.102	0.0010
State Value=FL	0.033	0.0010
State Value=TX	0.031	0.0010
Portion	0.026	0.0010
Backed_by_Real_Estate	0.00	0.0010
Industry Value=0	0.00	0.0010
DisbursementGross	0.00	0.0010
Industry Value=72	0.00	0.0010
Industry Value=45	0.00	0.0010

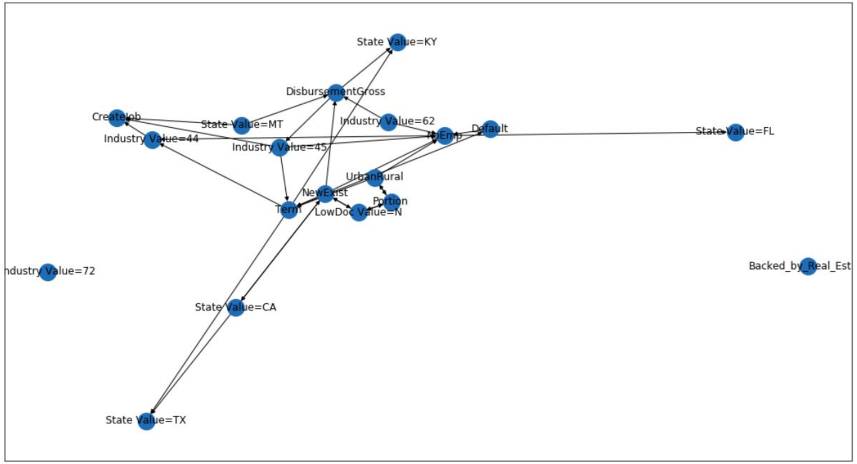
Causal Discovery

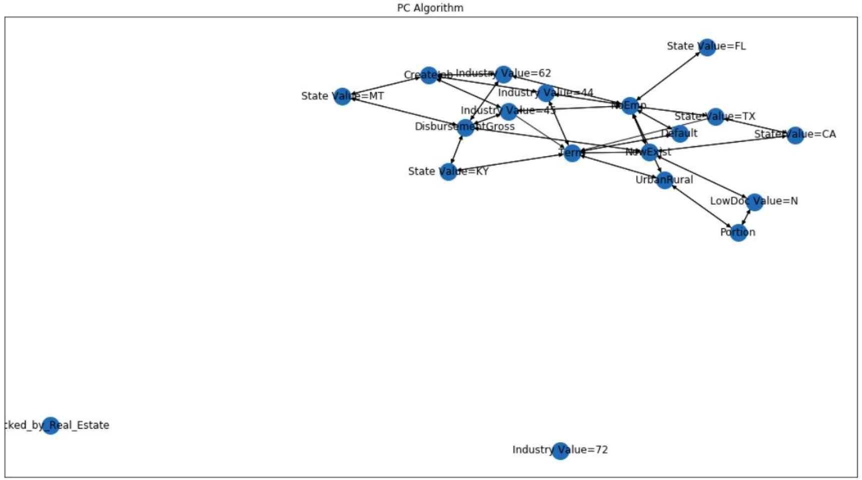


Ground Truth: *Real Estate Backing* was a **spurious correlation**—*Term of Loan* being the confounder.

This also makes sense because *Term of Loan* is a function of the expected lifetime of the assets. Hence loans backed by real estate will have terms 20 years or greater and are the only loans granted for such a long term, whereas loans not backed by real-estate will have terms less than 20 years.

Causal Discovery Toolbox





Causal Discovery using *Greedy Equivalent Search*

Causal Discovery using *PC Algorithm*

Methodology

- **Baseline model (Banks' balance sheet)**: Since the dataset only comprises of accepted loan applications, baseline refers to the banks original decision on the loans.

- **State-of-the-art AutoML:** ML model built on top of features that correlate with the target. AutoML platform handles the process of building several ML models and tuning them to select the best performing model.

- **Causal AI model:** ML model built on causal drivers—identified by causal discovery. Once again, model building and tuning is automated.

Assuming *interest rate (i) = 4%*

PnL = $(P * ((1+i)^{(n-1)}))$*(Not default) + -1*$(P *(1-f))$*(Default)

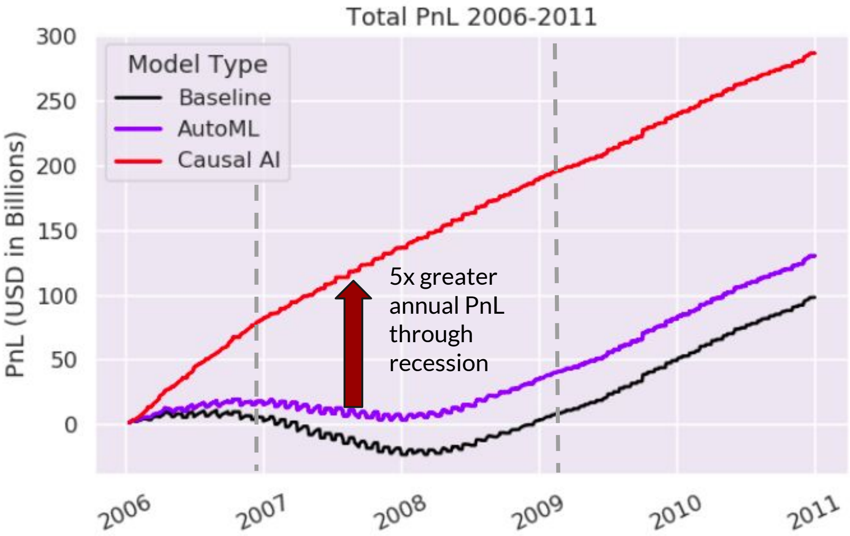
Profit   Loss

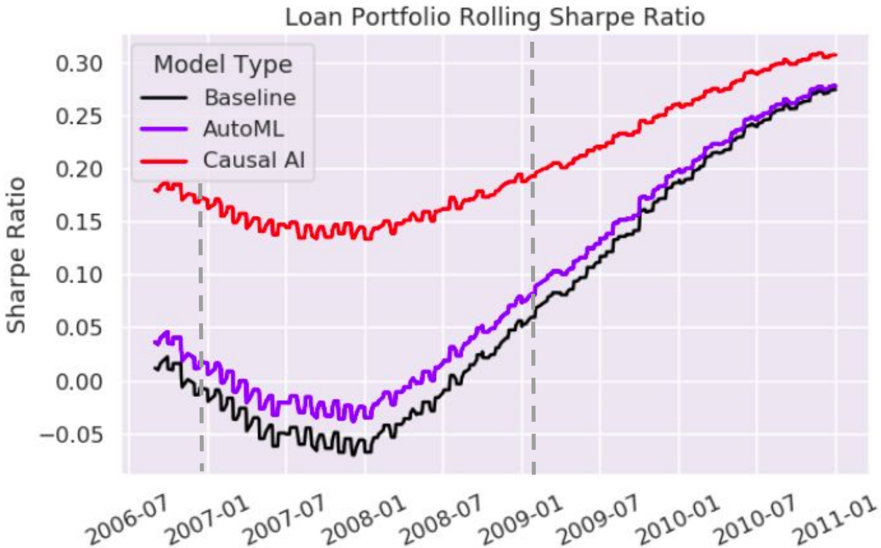
Where P := Disbursement amount
f := Portion of loan guaranteed by SBA
n := Term of loan

Sharpe Ratio = (Mean of PnL) / (Std dev of PnL)

Date	experiment	Baseline_PNL	CausalAI_PNL	AutoML_PNL
1990-01-02	training	561110.551100002	561110.551100002	561110.551100002
1990-01-03	training	2253565.37517998	2253565.37517998	2253565.37517998
1990-01-03	training	288411.813121068	288411.813121068	0
1990-01-03	training	-200000	-200000	0
1990-01-03	training	121531.44436325	121531.44436325	0
1990-01-03	training	-280000	-280000	0
1990-01-04	training	1765150.41411785	1765150.41411785	1765150.41411785
1990-01-04	training	163277.392461024	163277.392461024	0
1990-01-05	training	1360087.63034566	1360087.63034566	1360087.63034566
1990-01-05	training	2228982.38010661	2228982.38010661	2228982.38010661
1990-01-08	training	-400000	-400000	0
1990-01-08	training	9658132.27143635	9658132.27143635	9658132.27143635
1990-01-08	training	-1200000	-1200000	-1200000
1990-01-08	training	27743.3142484948	27743.3142484948	27743.3142484948

Results





Causal AI model trained on [Small Business Administration (SBA) loans data] (1990-2005), consistently outperforms conventional ML and the baseline (banks' balance sheets) on out-of-sample data.

While conventional ML models would have lost money as default rates skyrocketed during the recession (global financial crisis of 2007-08), Causal AI was able to identify key causal drivers of loan risk to consistently turn a profit.

Read more about Causal AI…

causalens.com/white-papers