Data Science Tasks

Prediction, causality and the Causal Forest

Luis Moneda, Data Scientist at Nubank

About me

Work

- Data Scientist at Nubank

Education

- Bachelor in Economics (FEA-USP)
- Bachelor in Computer Engineering (Poli-USP)
- MsC Computer Science student (IME-USP)

Twitter: @lgmoneda
LinkedIn: Luís Moneda
E-mail: lgmoneda@gmail.com
Blog: http://lgmoneda.github.io/

Outline

1. Data Science Tasks
2. Prediction Vs Causal
3. The causal inference challenge
4. Causal Inference approaches
5. Causal Forest

Data Science Tasks

Data Science Tasks

Description	Prediction	Causal Inference
- Computing proportions - Mean - Clustering - Visualizations	Mapping inputs (X) to output y.	Using data to calculate certain feature of the world if the world had been different: **counterfactual** prediction.

Reference: Miguel A. Hernán, John Hsu, Brian Healy. "Data science is science's second chance to get causal inference right: A classification of data science tasks", arXiv:1804.10846v2

Data Science Tasks - Examples

Description	Prediction	Causal Inference
What proportion of women aged 60-80 years had a stroke last year?	What is the probability of having a stroke next year for women with certain characteristics?	Will taking the drug A reduce, on average, the risk of stroke in women with certain characteristics?

Data Science Tasks - Confusion Matrix

What you're trying to do

Approach you're using	Description	Prediction	Causal Inference
Description	You're able to provide a snapshot of your data.	Nice benchmark, poor performance.	Misleading results, bad decisions.
Prediction	Why predict if you have the actual?	Low error predictions.	Biased estimations.
Causal Inference	Cost ineffective, but cool!	Cost ineffective, not the best performance.	Unbiased estimation of actions' effects.

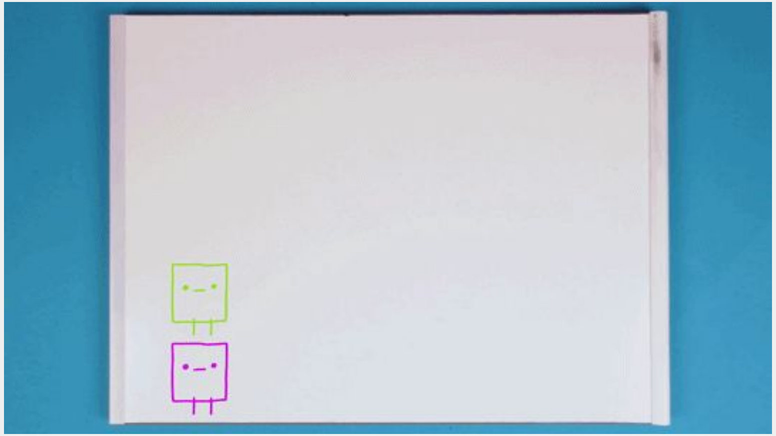
Prediction Vs Causal

Prediction

Most of successful applications today in DS are merely predictive!

Why?

1) A large dataset with inputs and outputs;
2) An algorithm that establishes a mapping between inputs and outputs;
3) A metric to assess the performance of the mapping.



All the information required is in the data!

Am I facing a prediction or causal problem?

$$\frac{d\pi(X_0, Y)}{dX_0} = \frac{\partial \pi}{\partial X_0}(Y) + \frac{\partial \pi}{\partial Y}\frac{\partial Y}{\partial X_0}$$

$\boldsymbol{\pi}$: Pay-off function

$\boldsymbol{X_0}$: Decision

$\boldsymbol{Y}$: Outcome

Illustrative example: Umbrella x Rain dance

Reference: Prediction Policy Problems. By Jon Kleinberg. Jens Ludwig. Sendhil Mullainathan. Ziad Obermeyer.

Causal

Confusion

- Spurious correlation
- Anecdote
- Science reporting

It's hard!

- Definition it is tricky
- Causal inference requires untestable assumptions
- I can only observe one potential outcome for each case

Causal - Notation

W: Treatment assignment

X_i: Features / Characteristics

Y: Observed outcome

Y^1: Outcome that would be observed if treated

Y^0: Outcome that would be observed if not treated

Causal - Core concepts

Potential outcome

The outcome we would see under each possible treatment option (Y^n).

Counterfactual

Slightly different than potential outcomes, but often used interchangeably.

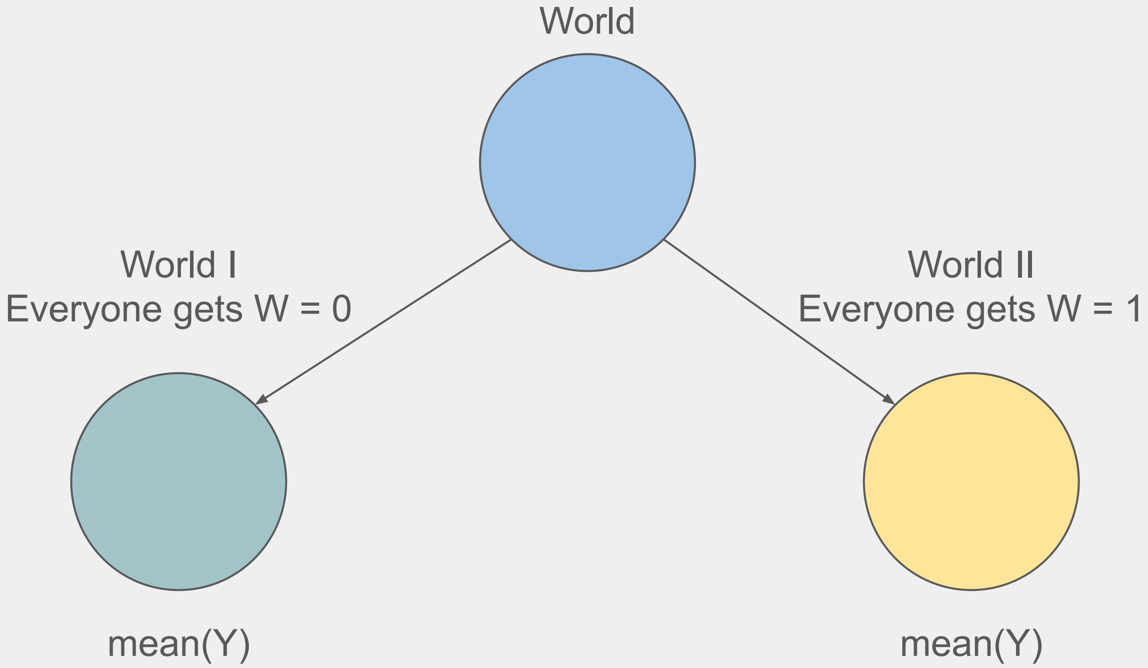
What would have happened had the action been different?

Before treatment decision is made, any outcome is a potential outcome: Y^1 or Y^0 .
After treatment there's an observed outcome Y^A and a counterfactual one Y^{1-A}.

Confounding

Anything that can impact both W and Y .

Causal - Causal Effect



Average Causal Effect = $E[Y^1 - Y^0]$

Causal - Randomized Controlled Trial (RCT)

- Golden standard;
- Solves all our problems!
- It has its own challenges, but once solved the results are robust;
- People in academia are used to do it.

The challenge

If random testing is a way to avoid all the difficulties of estimating causal effect, why do we even bother?

- It may not be **ethical**
- It can be **costly**

The challenge is estimating causal effect using either just **observational data** or using it with some random test data.

Causal - Assumptions

SUTVA

The outcome Y depends only on the individual features. No interaction/interference between individuals.

Consistency

The observed outcome under the treatment W must match the potential outcome $Y = Y^W$

Ignorability

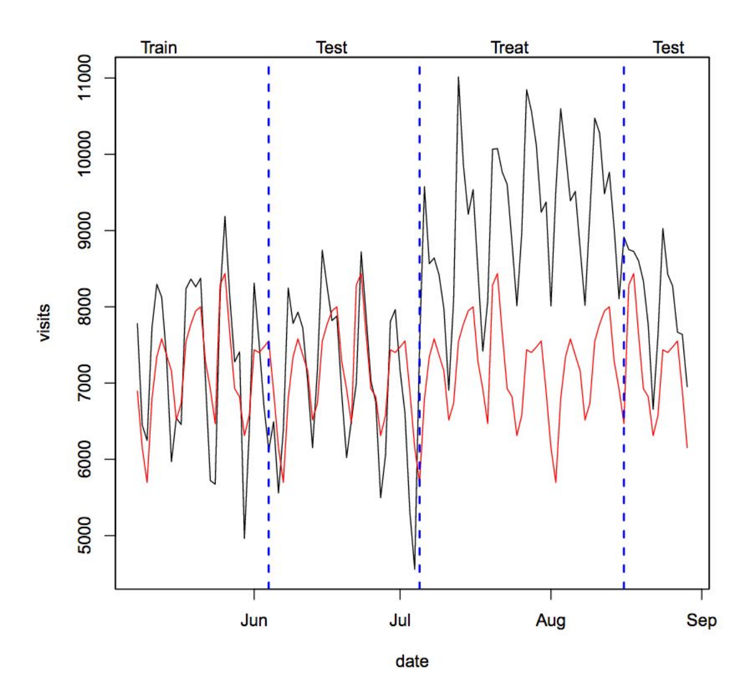
No unknown confounders: Y^W orthogonal W given X.

Positivity

The chance of being treated is positive: $P[W = 1 \mid X = x] > 0$ for all x. The treatment can't be deterministic.

Causal Inference Approaches

Causal Inference - Train Test Treat Compare



Reference: Causal inference in economics and marketing by Hal R. Varian

Causal Inference - Regression Discontinuity

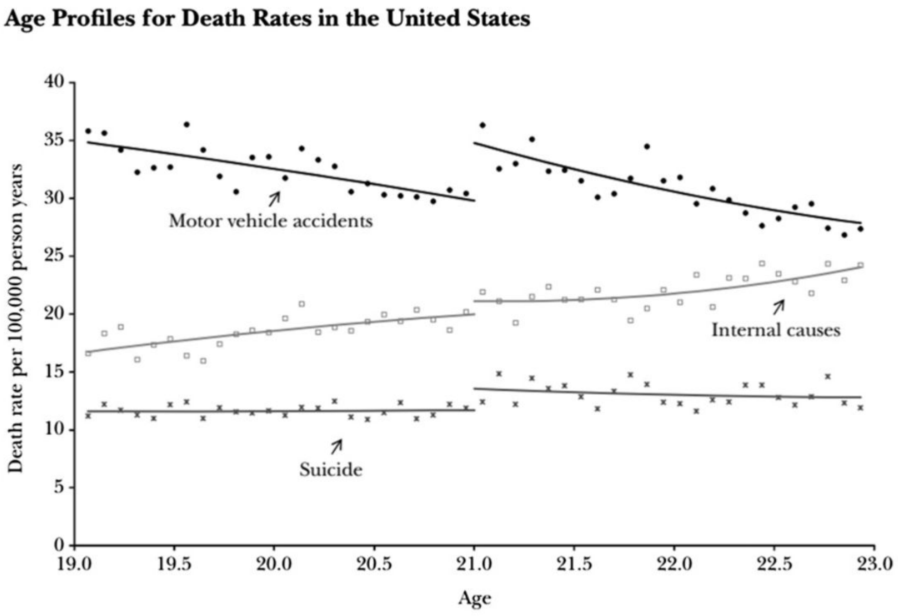


Fig. 2. Death rates by age by type of death. Reprinted with permission from ref. 13.

Reference: Causal inference in economics and marketing by Hal R. Varian

Causal Inference - Diff-in-diff

s_{TA} = sales after ad campaign for treated groups

s_{TB} = sales before ad campaign for treated groups

s_{CA} = sales after ad campaign for control groups

s_{CB} = sales before ad campaign for control groups

We assemble these numbers into a 2×2 table and add a third column to show the estimate of the counterfactual.

The counterfactual is based on the assumption that that the (unobserved) change in purchases by the treated would be the

Period	Treatment	Control	Counterfactual
Before	s_{TB}	s_{CB}	s_{TB}
After	s_{TA}	s_{CA}	$s_{TB} + (s_{CA} - s_{CB})$

same as the (observed) change in purchases by the control group. To get the impact of the ad campaign, we then compare the predicted counterfactual sales to the actual sales:

Reference: Causal inference in economics and marketing by Hal R. Varian

Causal Forest

Causal Forest - What is it about?

Goal: Heterogeneous treatment effect using **observational data**, estimating the effect on individuals rather than the average for the whole population or subgroups.

How: trying to learn the causal effect by **grouping similar observations** in the same leaf and comparing the treated and untreated.

Why it's interesting: for decision making in causal inference problems you need confidence intervals since you can't validate in the data.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - Definitions

Observed data: (X_i, Y_i, W_i)

Unconfoundedness: $\{Y_i^1, Y_i^0\} \perp W_i \mid X_i$

Treatment effect: $\tau(x) = \mathbb{E}[Y_i^1 - Y_i^0 \mid X_i = x]$

Treatment propensity: $e(x) = \mathbb{P}[W_i = 1 \mid X_i = x]$

Honesty

A tree is honest if, for each training sample i, it only uses the response Y_i to estimate the within-leaf treatment effect τ or to decide where to place the splits, but not both.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - From CART to Causal

- **CART:** $\hat{\mu}(x) = \frac{1}{|\{i:X_i \in L(x)\}|} \sum_{\{i:X_i \in L(x)\}} Y_i$
- **Causal:** $\hat{\tau}(x) = \frac{1}{|\{i:W_i=1, X_i \in L(x)\}|} \sum_{\{i:W_i=1, X_i \in L(x)\}} Y_i - \frac{1}{|\{i:W_i=0, X_i \in L(x)\}|} \sum_{\{i:W_i=0, X_i \in L(x)\}} Y_i$
- **Ensemble of B trees:** $\hat{\tau}(x) = B_{-1} \sum_{b=1}^{B} \hat{\tau}_b(x)$

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - Learning

1) Draw a random subsample of size s from {1, ... , n} without replacement, and then divide into two disjoint sets of size I and J, both of size s/2;
2) Grow a tree via recursive partitioning. The splits are chosen using any data from the J sample, but without using Y-observations from the I-sample;
3) Estimate leaf-wise responses using only the I-sample observations.

The splits are done maximizing the variance of the estimated effect using the J sample. Each leaf should contain k or more I-sample observations of each treatment class.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Causal Forest - What is happening inside it?

- The estimation in the leafs addresses the effect of treatment;
- The idea is that in each leaf it behaves like a random experiment in a sub group
- The restriction of having k or more examples of each treatment helps to make it closer to a random experiment;
- The more the treatment is far from randomness the harder it’s to work with a small k, the restriction of having at least k examples of each treatment;

At the end of the day: I’m just comparing treated and not treated examples using a tree to split it smartly and build a fair group to do this comparison for individual/sub groups examples.

Reference: Estimation and Inference of Heterogeneous Treatment Effects using Random Forests by Stefan Wager and Susan Athey

Conclusion

- Know how to classify the problem you're trying to solve as one of the DS tasks or as a mixed component of them
- Causal problems are harder, but it's better to face it than trying to solve with the wrong tools
- Be a causal warrior and object against causal conclusions made without observing the assumptions

Questions?

Apply to Nubank!

http://nubank.workable.com

Contact

Twitter: @lgmoneda

LinkedIn: Luis Moneda

E-mail: lgmoneda@gmail.com