Causal Modeling for Insurance Demand
- From Randomized Test to Double Machine Learning

Alina Guo, FSA, CERA

Director, Liberty Mutual APAC Data Science

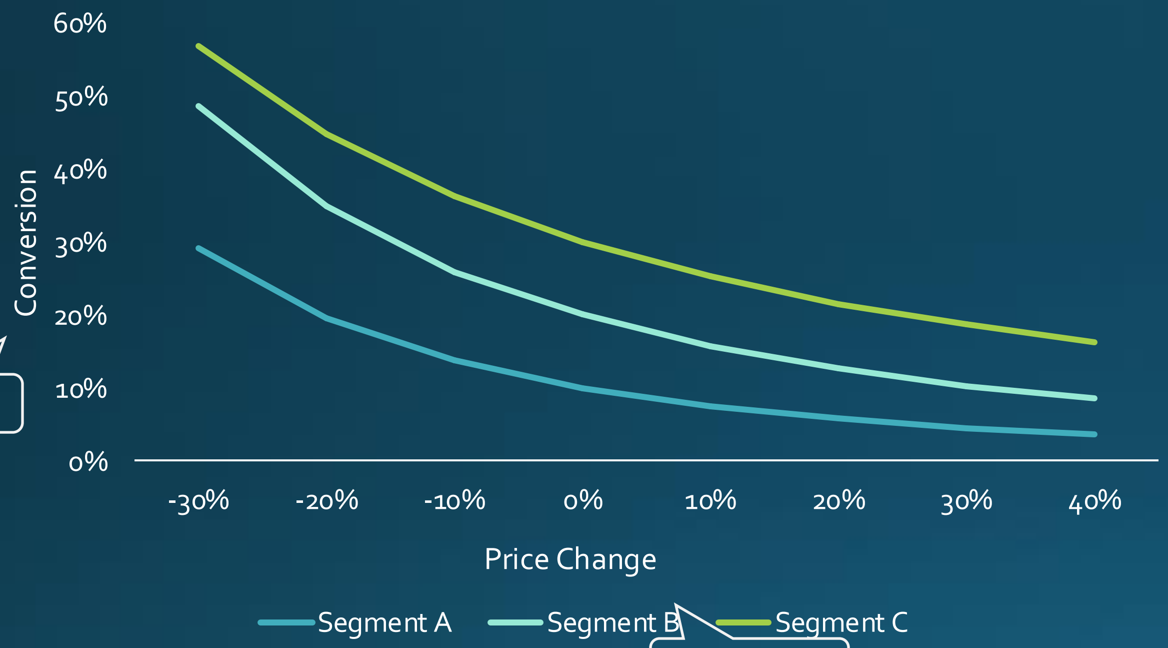
October 24th 2024

Outline

- Demand Curve/Elasticity
- Causal Concepts and Confounding Bias
- Causal Methods (RCT/DML)

Demand Analysis is Causal

Demand Models



Outcome

Treatment

Demand modeling wishes to discover the demand curve

Elasticity measured via demand curve

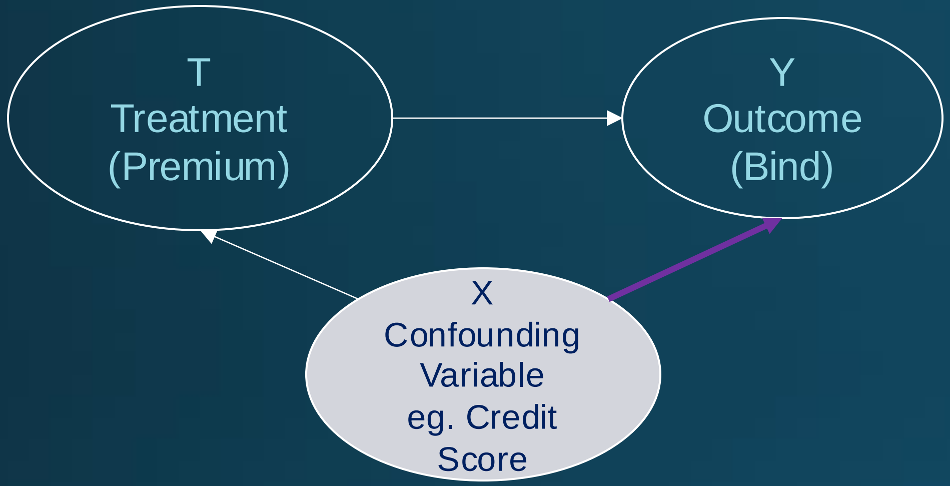
$$\text{Elasticity } (\mathcal{E}) = \frac{\%\text{ change in Quantity } (q)}{\%\text{ change in Price } (p)}$$

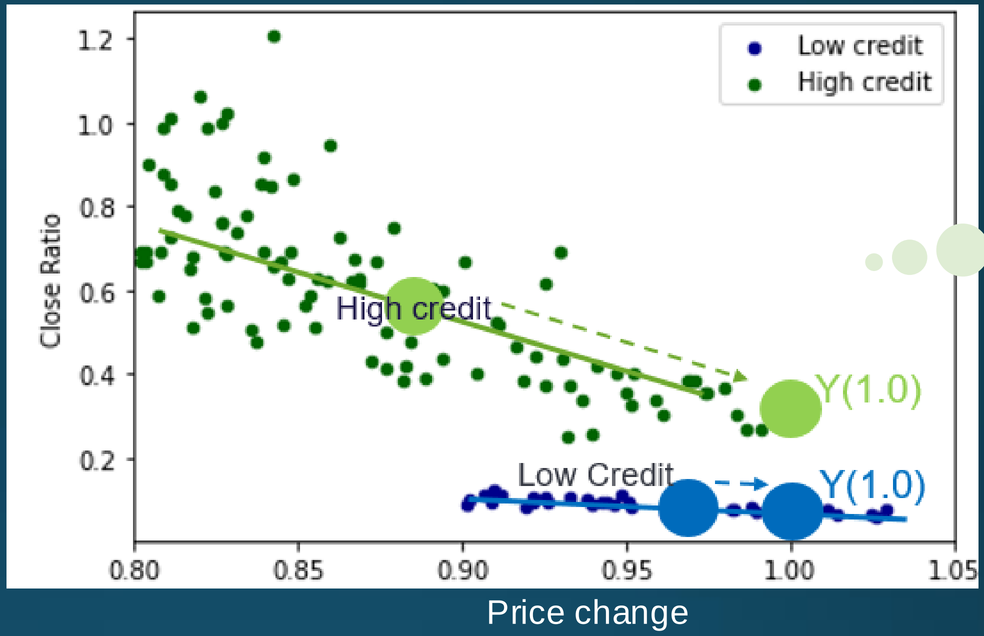
Knowing the demand curve or elasticity allows for projecting conversion or retention ratio across a wide range of hypothetical prices supporting cost/benefit analysis, as well as better pricing strategy

Confounding Bias

Causal Relationship: Price Change → Demand Change

Confounders – Key to the Causal Modeling

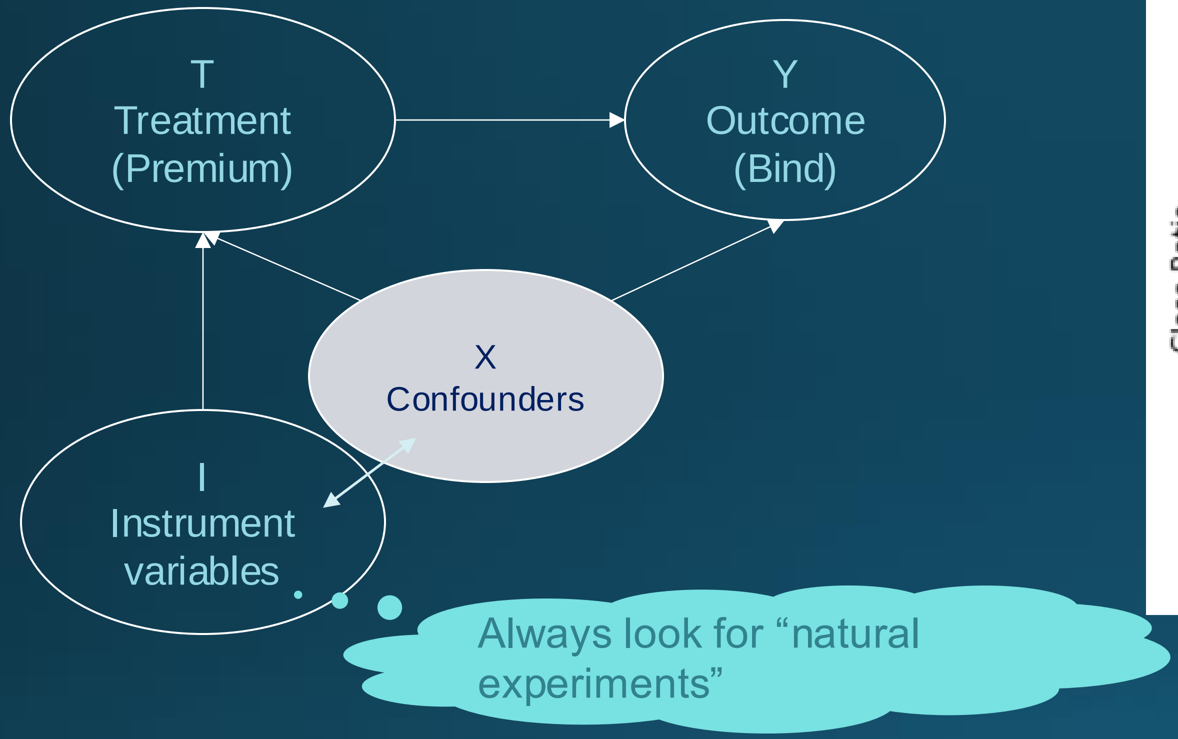


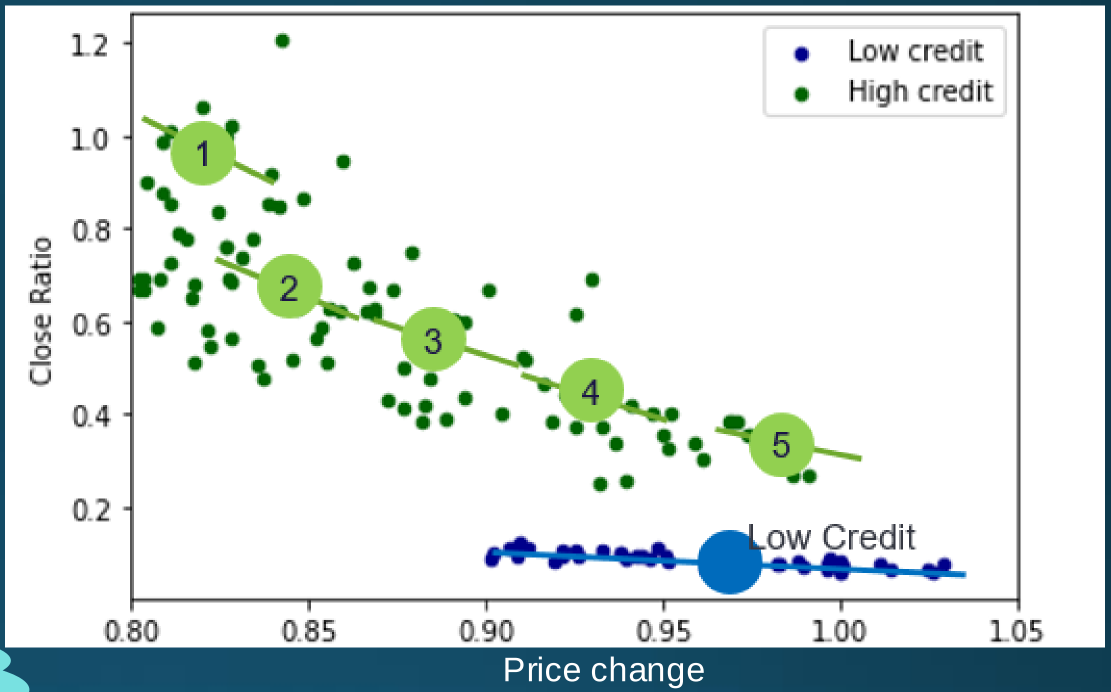


Conditioning on a confounder (credit) is equivalent to free-fitting as a control in a linear model

- Hard to know the causal effect of Premium on Bind because both share a common cause: Customer Credit Scoring.
- Control for Credit Score - Conditioning on the confounder (analyze demand separately for each credit group)
- However, Credit Score is always measured by other variables (Customer's income, occupation…), those variables are called "Instrumental Variables"

Instrumental variables

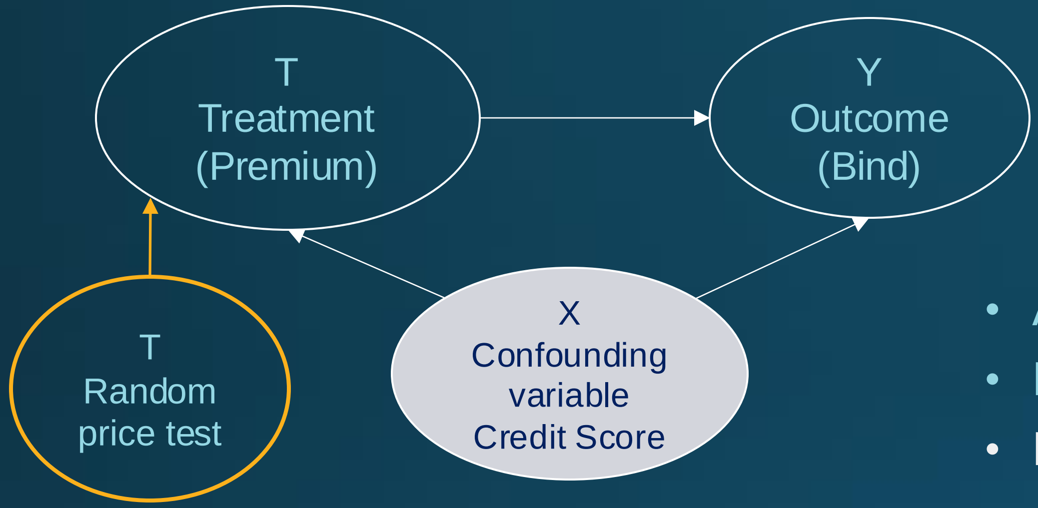




- **IVs** will cause the treatment and only correlated with the outcome through the treatment
- **IVs** can allow isolating causality even with unknown confounders
- Very hard to come by, especially in insurance data

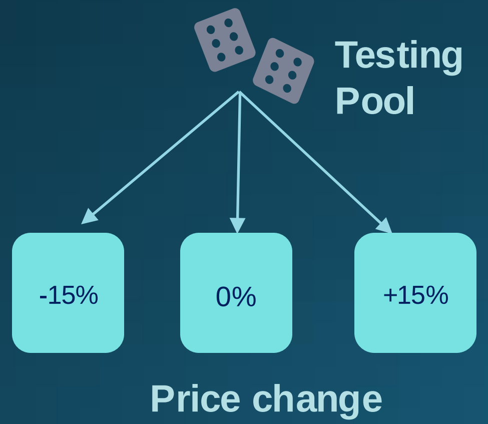
Any other way to avoid confounding bias?

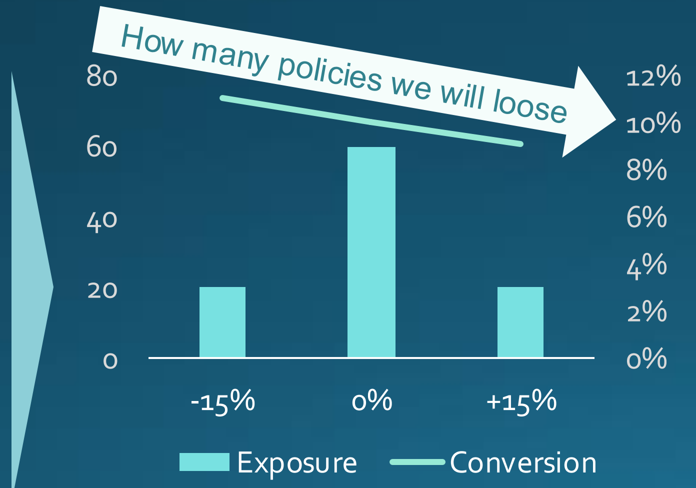
Causal Method: Randomized controlled trials (RCT)

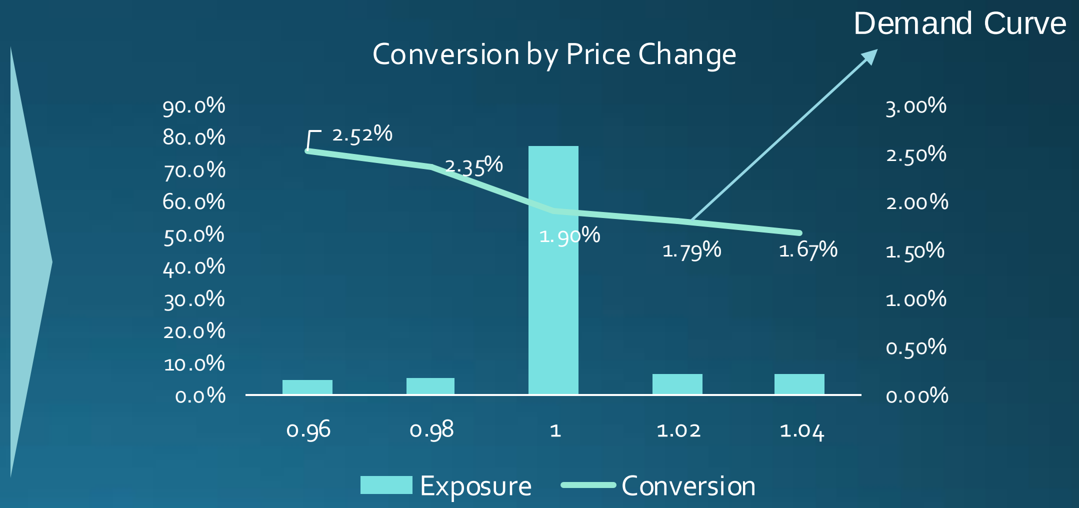


- Amazon: random price testing on 68 DVDs
- Orbitz: price test on Mac users and found this group is inelastic to expensive hotels
- Microsoft: tested pricing on Xbox One games
- …

- A randomized treatment is not caused by any variable
- Randomness is the gold standard in experimental design
- RCTs can be expensive and are not always allowed

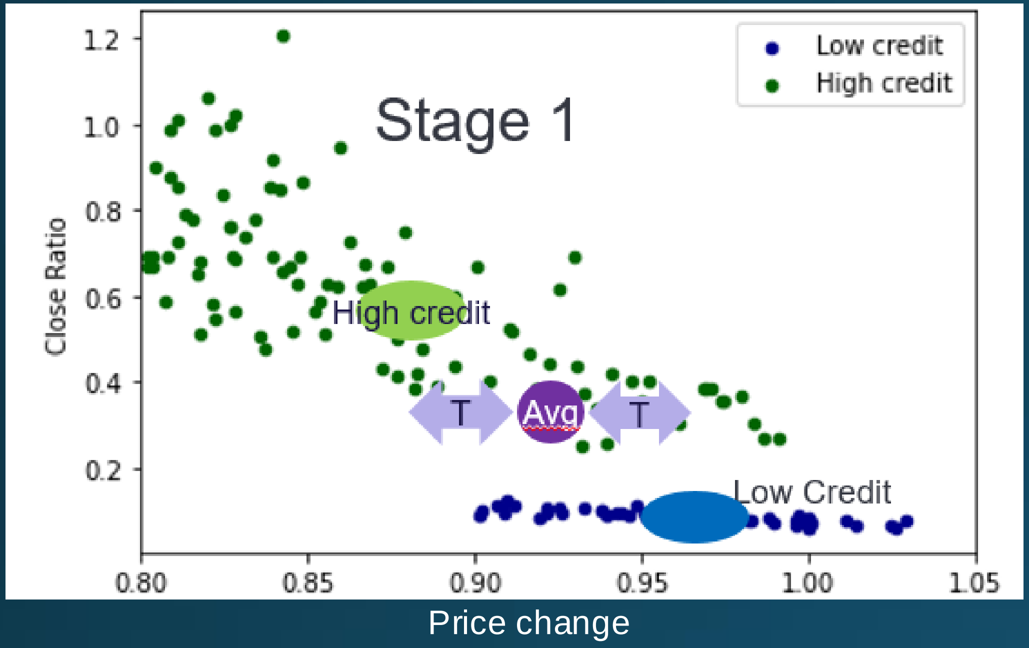


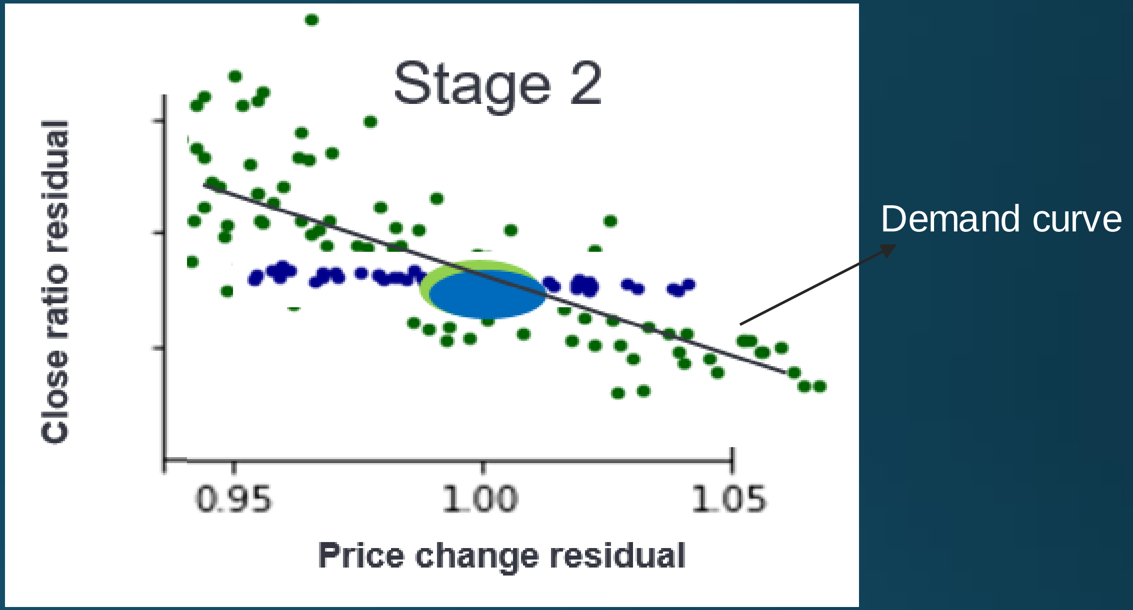




* Individual companies should do research and testing to determine the best price test buckets

Causal Method: Double Machine Learning (DML)





S1: Model the treatment and the outcome separately

X Confounders only

$T \sim X$ Treatment model

$Y \sim X$ Outcome model

S2: Estimate the causal effect of the treatment on the outcome

$$\left(Y - (Y \sim X)\right) \sim \left(T - (T \sim X)\right)$$

outcome residual treatment residual “price test”

The purpose of Stage 1 is to de-bias the treatment by removing the impact of confounding factors, so that stage 2 only considers treatment residual that drives outcome residual

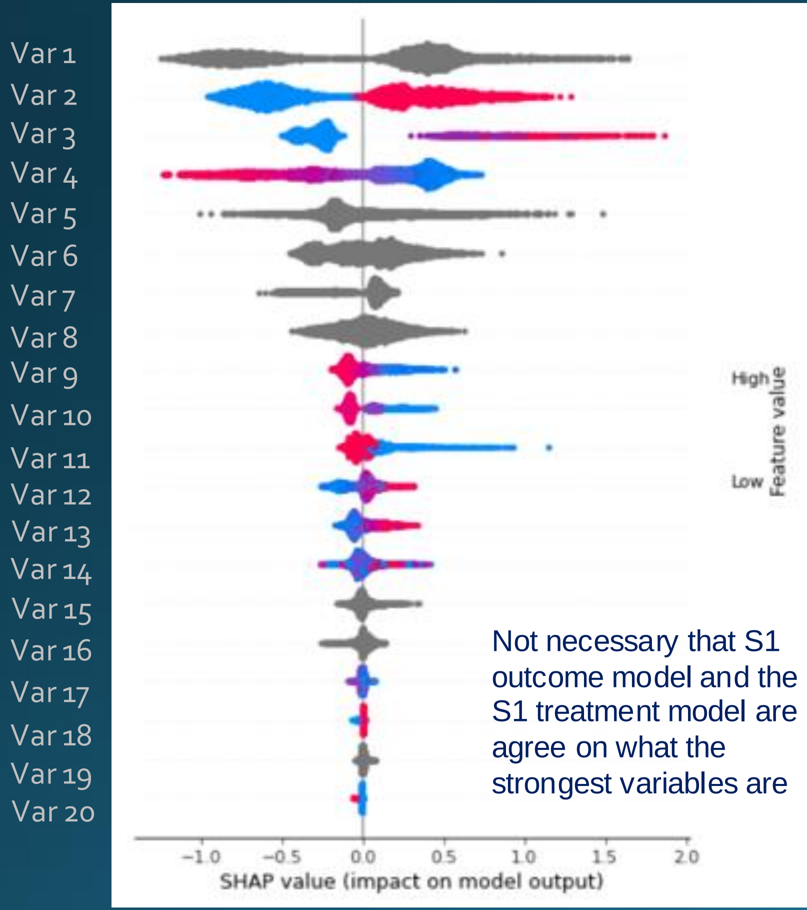
Case Study: 2-Stage Model

Stage 1 Treatment model

Y = Treatment = Price change

Stage 1 Outcome model

Y = Conversion



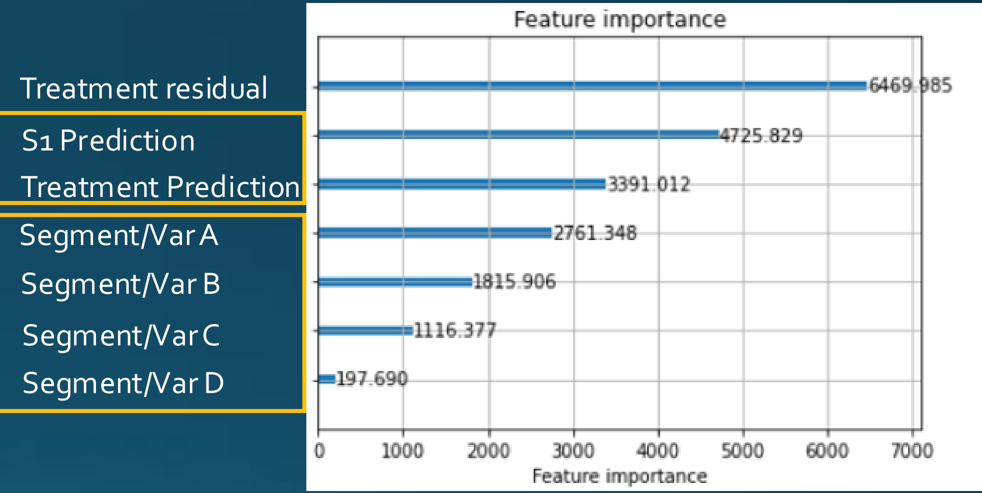
Not necessary that S1 outcome model and the S1 treatment model are agree on what the strongest variables are

Stage 2 Residual model

Y = Outcome residual

X = Treatment residual + Segmentations of interest to understand the demand curve

Segment the Elasticity



- In S2 the target is the conversion ratio residual. The residual model is accomplished by putting in the S1 conversion ratio prediction as an offset while keeping target = conversion. The main predictor is the treatment residual TM_res = price change / (S1 price change prediction)
- The best segmentor of elasticity is usually the conversion ratio - a higher elasticity in low conversion ratio segments (low S1 prediction) and vice versa low elasticity for high conversion ratio

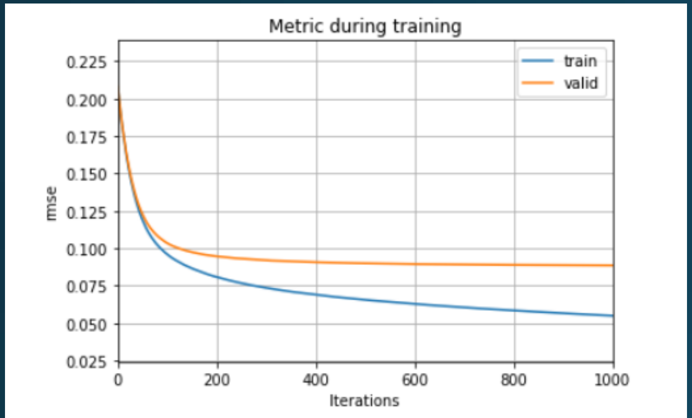
* Results are based on dummy data, only for illustrative purpose.

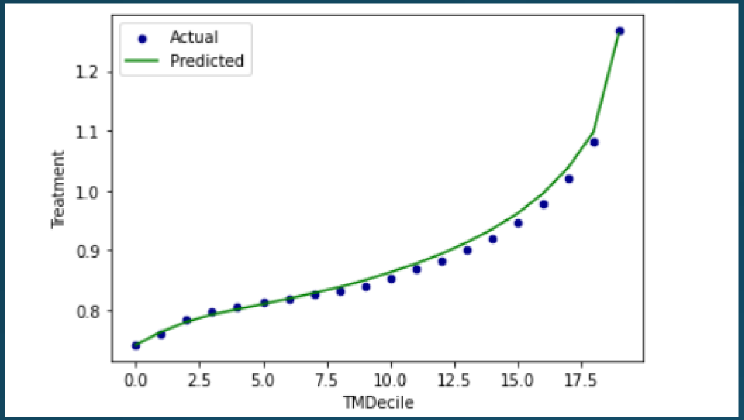
Case Study: Model Diagnosis

Stage 1 treatment model diagnostic plot and lift chart

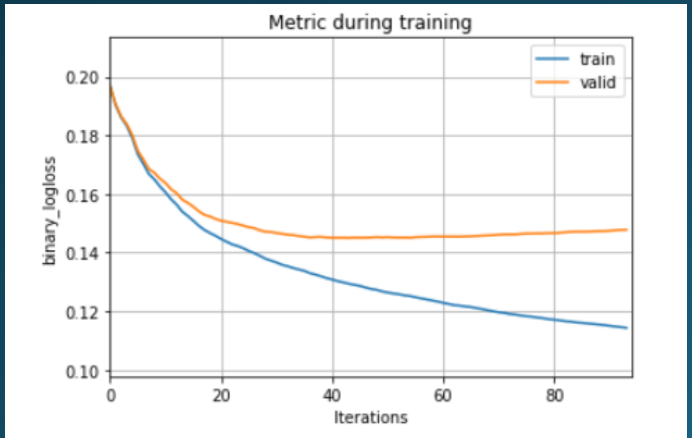
LightGBM

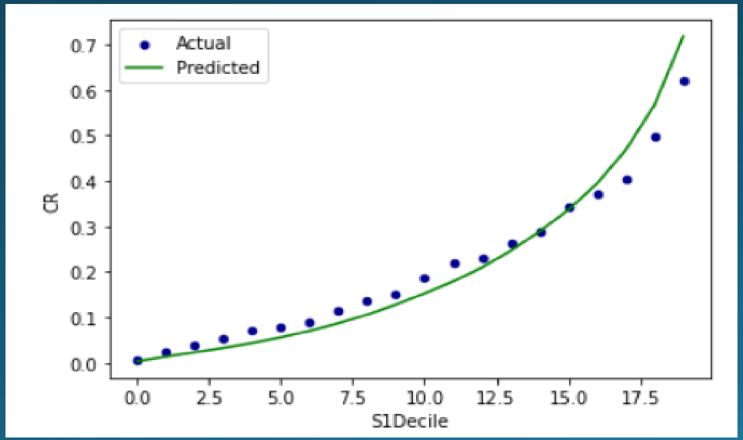
S1 prediction: 0.045
Actual conversion: 0.053





Stage 1 outcome model diagnostic plot and lift chart

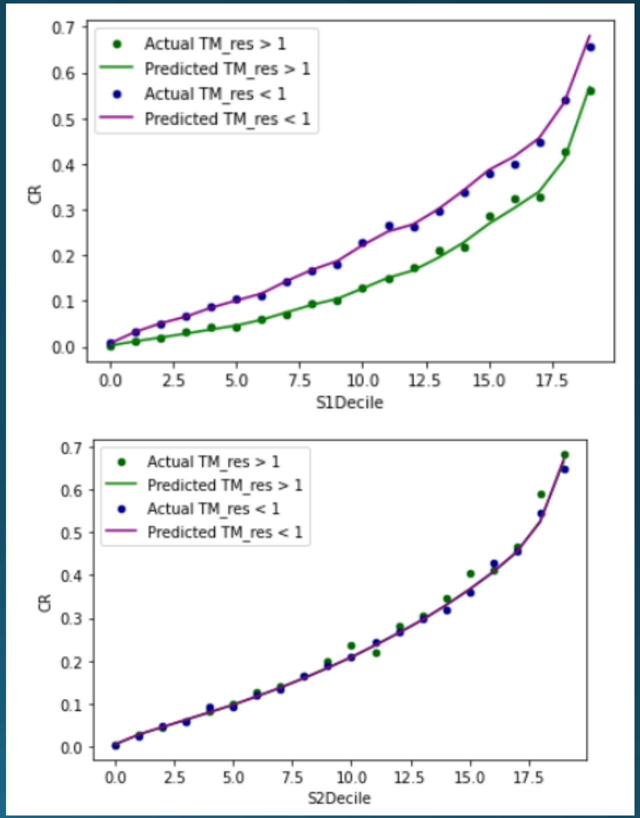




S1/S2 model lift chart

LightGBM

S2 prediction: 0.053
Actual conversion: 0.053



Conversion residuals are successfully predicted by S2 features!!

* Results are based on dummy data, only for illustrative purpose.

Reference Links – DML is Getting Popular!

- Double/Debiased Machine Learning for Treatment and Causal Parameters (2016)

  [The paper primarily investigates how to effectively estimate treatment effects and structural parameters in high-dimensional data environments by using DML. The key idea of the DML method is to use statistical or machine learning approaches, such as random forests, Lasso, ridge regression, deep neural networks, etc., to estimate the nuisance parameters and reduce bias through orthogonalization and cross-fitting techniques.]

- Estimating Causal Effects with Double Machine Learning -- A Method Evaluation (2024)

  [This paper provides a comprehensive review of the double machine learning (DML) framework for causal inference, particularly focusing on settings with multiple confounders.]

- Machine Learning in P&C Insurance: A Review for Pricing and Reserving (2020)

  [This review discusses the recent trends in machine learning applications in property and casualty (P&C) insurance, including pricing models that can benefit from advanced techniques like **double machine learning**. It provides insights into how these methods can improve the accuracy of predictions in insurance.]

- Double Machine Learning for Insurance Price Optimization (2023)

  [The paper aims to enhance insurance pricing models using DML, specifically addressing how to optimized pricing strategies while managing the complexities of high-dimensional data. The authors leverage machine learning models to control for the influence of numerous confounders (other risk variables like policyholders characteristics, external factors, etc…) and focus on price elasticity estimation.]

- Variables Selection in Double/De-biased Machine Learning for Causal Inference (2020)

  [The paper proposes an outcome-adaptive approach to variables selection that integrates DML. This approach allows the model to adjust the selection of variables based on the outcome of interest, thereby enhancing the robustness and efficiency of the estimates. The author emphasize the importance of identifying relevant variables for estimating causal effects accurately.]

……

Causal Method: GLM or ML?

Machine Learning (ML), or more specially **Double Machine Learning (DML) is superior to Generalized Linear Model (GLM) in dealing with <u>Confounding Bias</u>**

Comparative Analysis	GLM	DML
Accuracy of Causal Inference	Mainly used for <u>prediction</u>, not causal inference, and may not provide unbiased estimates of treatment effects.	Focuses on causal inference and provides an effective method for estimating causal effects through <u>orthogonalization</u> and <u>debiasing</u> techniques.
Orthogonalization and Debiasing	Very complex in practical operations.	Modeling the treatment and outcome variables separately, removing the impact of confounders, then use the residuals from each model to estimate the treatment effect.
Model Flexibility	Model specification needs to be set in advance	The use of Non-parametric techniques can capture complex, non-linear relationships
Reducing the Impact of Model Mis-specification	Easy to lead to estimation bias if having mis-specification, eg. incorrect link functions or distribution assumptions.	Reduce the risk by using two independent machine learning models to estimate treatment effects.
Handling High-Dimensional Data	Encounter inaccurate parameter estimation issues	Machine learning algorithms are good at handling
Overfitting	May overfit where there are many confounders.	Reduce the risk of overfitting by model selection techniques.

Thank you!

For any questions, feel free to reach out to me:
Alina.Guo@LibertyMutual.com