Estimating Feature Causality Using Machine Learning

Felix Zhu

Abstract

Companies often need to know how customers respond to new features that are released. They often do so through key metrics that represent company performance. Typically, Randomised Control Experiments are the traditional method of causal attribution, in the presence of confounders. It does so through the mechanism of back-door adjustment. However, given the volume of features and the financial cost of implementing experiments, this is not always possible. The following article examines the literature on how Machine Learning can also perform back-door adjustment through explanatory methods such as ALE Plots and Explainable Boosting Machines. Furthermore, in the presence of multicollinearity and uncertain causal structures, a novel technique is proposed which creates a range of causal attribution with a best estimate and upper bound.

Contents

Chapter 1

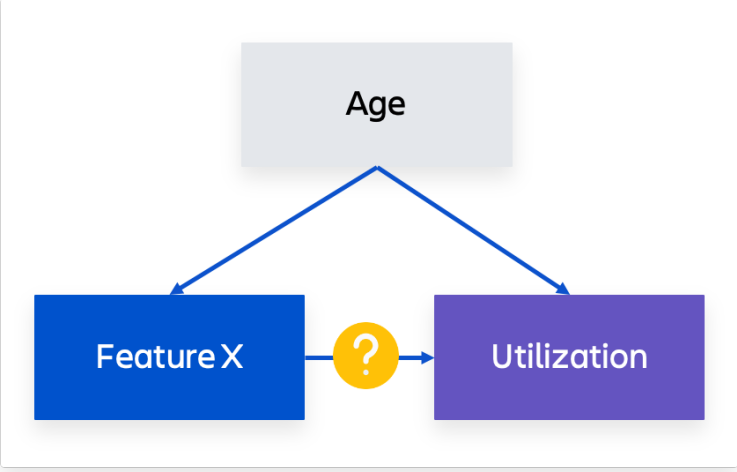
Motivation

As an example, imagine we are working at a social media company with subscribing customers. An important metric that we track is Utilization, describing how engaged they are with the product. Imagine we have just released Feature X with the aim of improving Utilization.

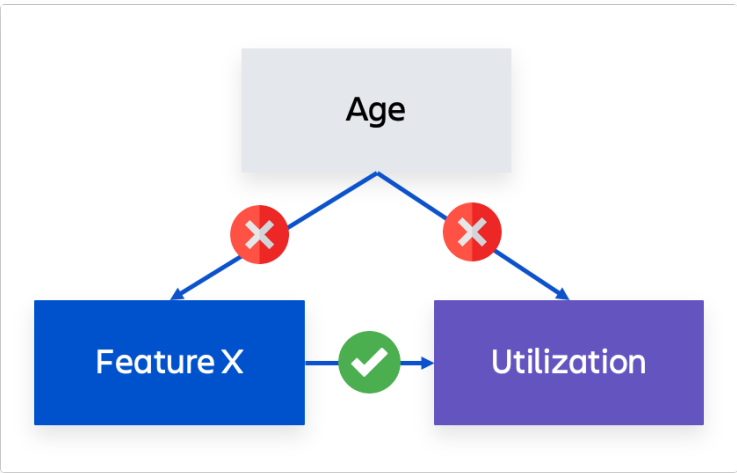
How do we determine whether these new features have improved Utilization?

Unfortunately, it's not as simple as plotting Feature X against Utilization as there are many factors that go into this. On top of how our customers use Feature X, we also need to consider other characteristics such as their age.

Perhaps younger customers will naturally use Feature X, but at the same time have higher Utilization as our product is more popular with the younger generation. How do we know that Feature X had any incremental impact on Utilization beyond how old they are?



Typically we would use the tried and tested Randomised Control Experiment (RCT) to isolate our feature of interest (Cartwright 2010). By randomly selecting customers that use Feature X, age no longer plays a part in the decision to use it, and we are able to isolate its impact on Utilization. In other words, we have controlled for the age confounder and isolated the causal link between Feature X and Utilization. The full process is called a back-door adjustment (Pearl 1993).



Chapter 2

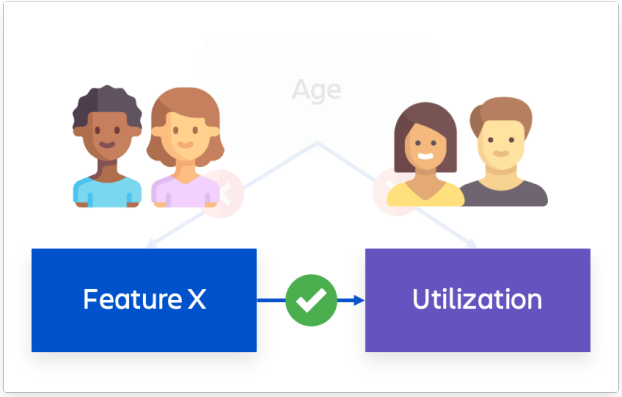
Attribution using Machine Learning

What happens when we don't have the time or resources to perform the experiment?

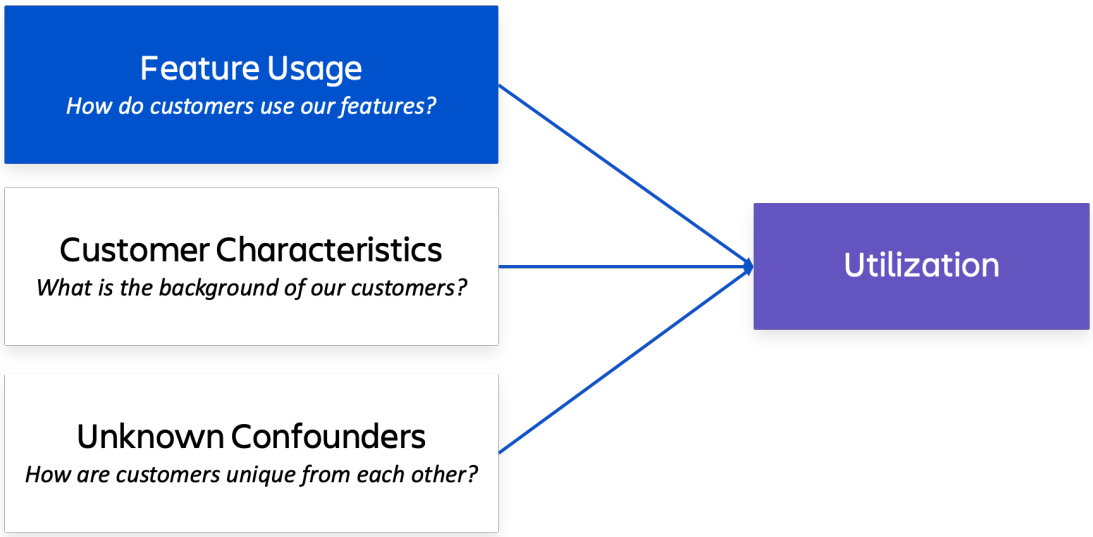
Luckily, RCTs are not the only way to perform back-door adjustments. Indeed, if we knew beforehand that age was going to play a part in Feature X and Utilization, then we can directly apply Pearl's formula to attribute causality.

$$P(Utilization|do(FeatureX = X)) = \int P(Utilization|FeatureX = X, Age = Y)dP(Y) \tag{2.1}$$

The formula states that if we isolate different ages, and compare Feature X to Utilization within, we will be able to attribute causality. Indeed, this is quite intuitive. If Utilization goes up with Feature X for customers between 20-30, as well as those between 40-50, then we can be confident that age did not play a part, and Feature X was the cause of Utilization.



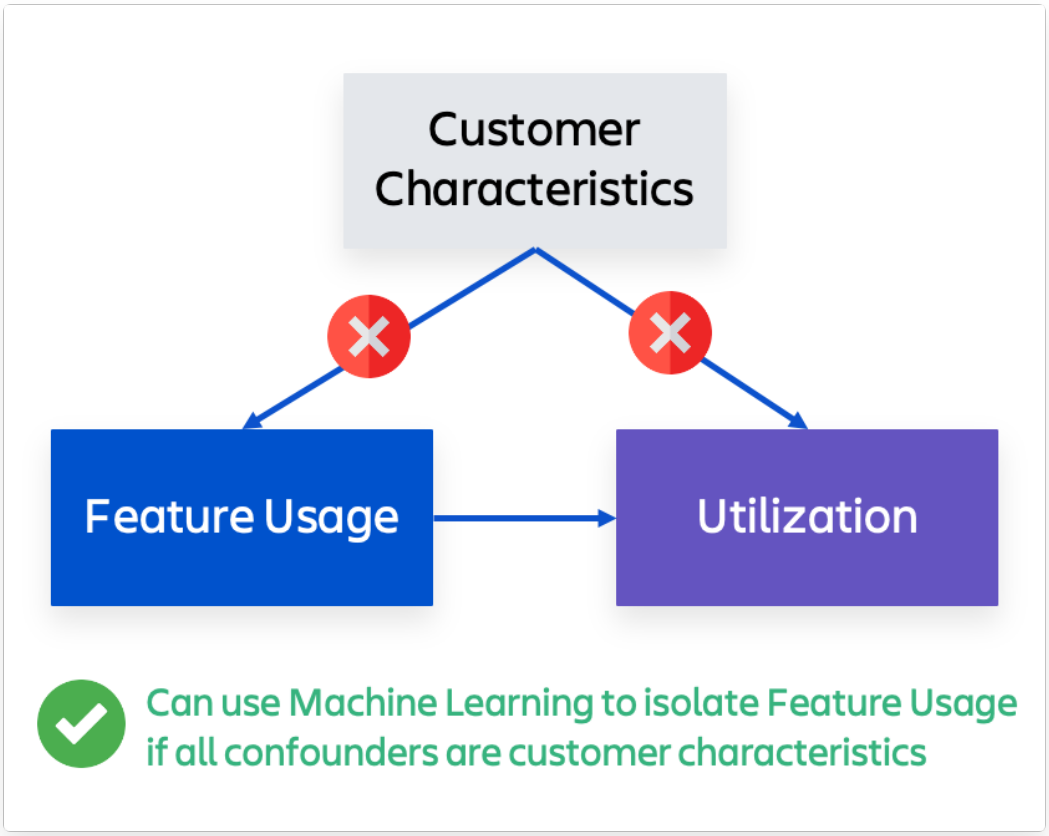
It is easy to hold a single variable constant, but what happens if we have hundreds of confounders at play?



Luckily, we have a commonly-used tool at our disposal that accounts for confounding variables, called Machine Learning. There are many models out there, the most popular of which include Random Forest and XGBoost, and more recently, Explainable Boosting Machines. We often use them to just predict Utilization, but we are now getting to the stage where we can begin asking them, why Utilization?

Zhao & Hastie (2021) show that these Machine Learning models perform the same back-door adjustment as RCTs, standardizing for all confounders that are specified within the model. If we have sufficient data and a strong understanding of the causal structure within the data, then we can attribute causality using it.

However, the problem is that we often do not know the extent of the confounders at play, let alone have access to the data. This limits the widespread use of Machine Learning in causal analysis. For the sake of this article, let us assume that the Customer Data we have does account for all the confounding variables. It is not a stretch to assume so given the extent of data that modern companies have access to.



Upon fitting the Machine Learning model, we can use explanatory methods such as ALE Plots (Apley & Zhu 2020) and Glassbox Methods (Nori et al. 2019) to visualise how models have isolated these features. The explanatory methods will show the average incremental impact of the target feature on the output, holding all included confounders constant.

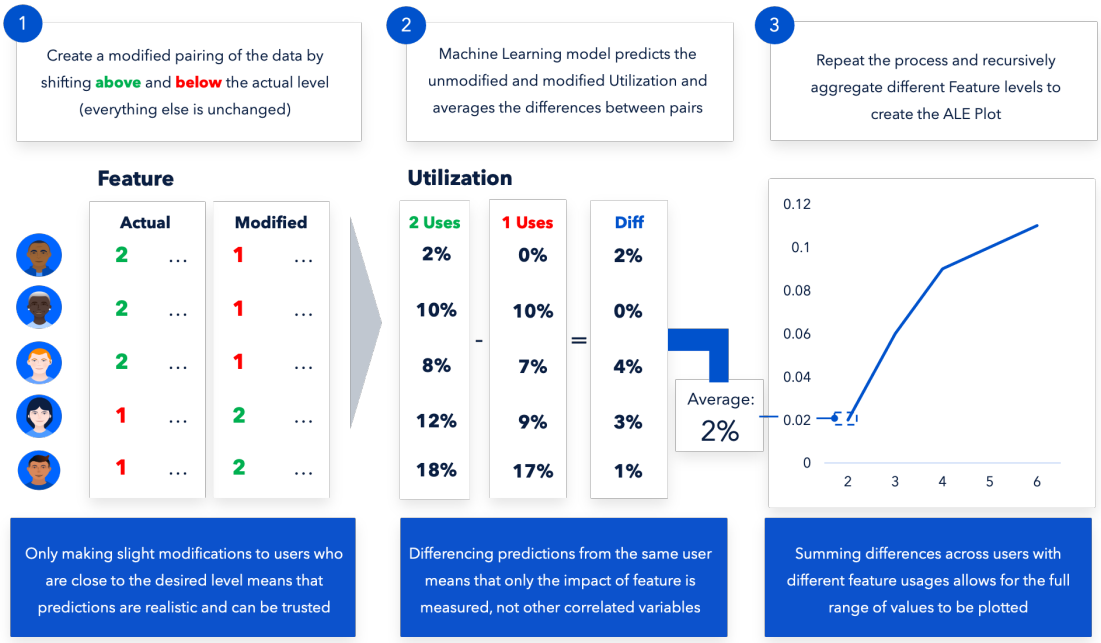
As summarized by Zhao & Hastie (2021), a successful attempt at causal interpretation requires at least three elements:

- **Good Predictive Model** so the estimated black-box function is (hopefully) close to the law of nature.
- **Domain Knowledge** about the causal structure to assure the back-door condition is satisfied.
- **Visualisation tools** such as ALE Plots and Glassbox Methods.

Accumulated Local Effects (ALE)

Accumulated Local Effects (ALE) describe how features influence the prediction of any machine learning model (Apley & Zhu 2020). They are a black-box methodology meaning they work without knowing the explicit transformation between input and output.

ALE is very good at summarizing incremental effects of features and is preferred over the more well-known Partial Dependence Plots (PDP) as they are true to the data and not biased by correlated features. This is because ALE only makes slight modifications to create plausible hypothetical scenarios and then takes differences to eliminate correlation effects.



Explainable Boosting Machine

Explainable Boosting Machines (EBMs) are Machine Learning models created by Microsoft that are often as accurate as state-of-the-art black-box models whilst remaining completely transparent about how they transform inputs into outputs (Nori et al. 2019). The interpretation of feature importance is hence Glass-box as opposed to Black-box like ALE Plots.

EBMs fit the following formula to the data:

$$g(E[y]) = \beta + \sum f_i(x_i) + \sum f_{i,j}(x_i, x_j) \tag{2.2}$$

They take advantage of modern machine learning techniques such as Bagging or Boosting to do so with a high degree of accuracy. Indeed, the advantage of this model is that

it has a clear additive functional form where each feature contributes to predictions in a modular way. This comes in handy during interpretation time when the impact of features can clearly be isolated by examining its individualised functions.

Chapter 3

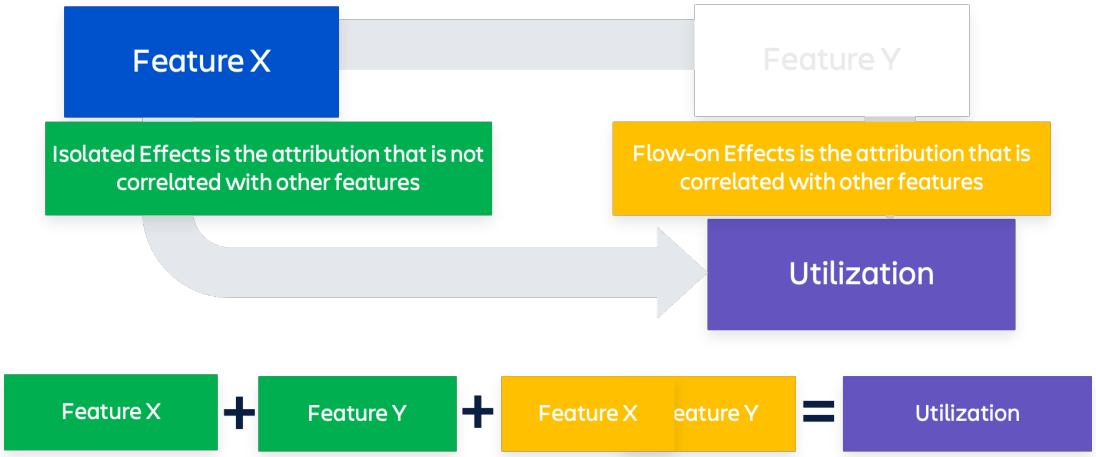
Attribution Under Causal Structure Uncertainty

What happens if we are not sure whether Feature X causes Feature Y?

It is very obvious that the direction of causality stems from Age to Feature X and not the other way around. After all, using Feature X will not suddenly make someone younger or older.

However, it gets complicated when we introduce a Feature Y that customers use in conjunction with Feature X. While we see that Utilization increases with Feature X and Feature Y, we are not sure which one it is due to. While historically they have moved together, our company wishes to pick one feature to invest in and we need to recommend the one with a higher impact.

We can split attribution into two parts based on correlation. Firstly, we can identify an isolated effect, describing the utilization attributed to the non-correlated components of the features. Secondly, we can identify a flow-on effect, describing the utilization attributed to the correlated components of the features. These two effects can be split based on how correlated the features are. Summing them, we get the total impact on utilization.



The most extreme case is when features are perfectly correlated and all attribution is through a flow-on effect. This is commonly known as multicollinearity and makes it so that we cannot observe features in isolation (Zidek et al. 1996). The traditional fix is to drop correlated features (Zainodin & Yap 2013).

Does dropping correlated features make sense in the context of causal attribution?

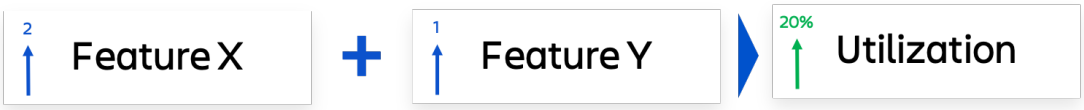
Let's take a look at an example where Feature X and Feature Y are perfectly correlated and all impacts on Utilization are flow-on effects.

Tenant	Feature X	Feature Y	Utilization
A	1	2	20%
B	2	4	40%
C	3	6	60%
D	4	8	80%
E	5	10	100%

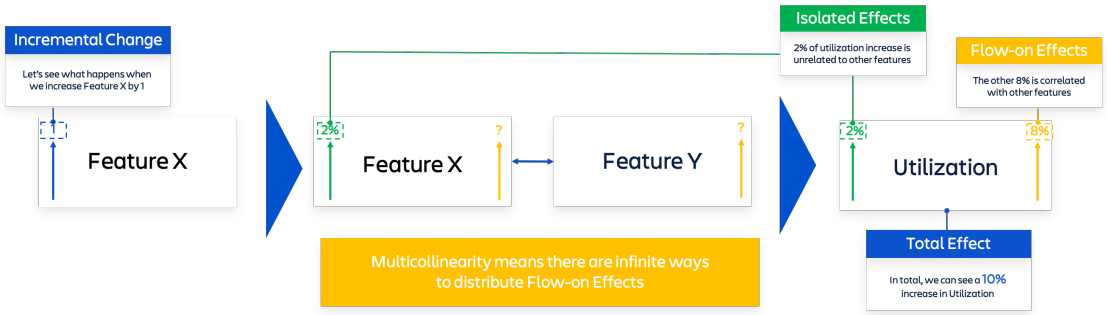
Treating Feature X as an isolated effect, we make the deduction that Feature Y leads to 20% Utilization and Feature X has no impact.

Conversely, treating Feature Y as an isolated effect, we make the deduction that Feature X leads to 10% Utilization and Feature Y has no impact.

None of that made sense, and that is because when it comes to multicollinearity, there are infinite solutions and it truly only makes sense to analyze features together, rather than in isolation (Basu & Maji 2022).



So clearly, we cannot drop correlated features as that would lead to over attribution. This means we must find a way to distribute flow-on effects when they occur. In this diagram, increasing XPC leads to a total 10% increase in Utilization, however, when adjusting for correlation between features, 2% are isolated effects, and the other 8% are flow-on effects. This is a case of imperfect multicollinearity.



The true way to split the flow-on effect depends on causality and is where an RCT would be helpful. Running an experiment on Feature X would allow us to directly see its impact on Utilization. If we see that it drives 10% Utilization, then we can give all the attribution to Feature X. However, realistically, we do not always have the novelty to run an RCT, and we can never know exactly how much attribution comes from each feature.

How should we split flow-on effects in the absence of Experimentation?

We must come up with different methods of splitting attribution. A simulation was run on a dataset of 10,000 samples. The dataset contains 3 perfectly correlated features, X_1, X_2 and X_3.

$$\begin{aligned} X_1 &\sim U(0, 1000) \\ X_2 &= \frac{X_1}{2} \\ X_3 &= \frac{X_1}{4} \end{aligned} \tag{3.1}$$

It also contains 3 perfectly uncorrelated features, Z_1, Z_2 and Z_3.

$$\begin{aligned} Z_1 &\sim N(0, 200) \\ Z_2 &\sim N(0, 200) \\ Z_3 &\sim N(0, 200) \end{aligned} \tag{3.2}$$

The causal structure of the target variable, Y, is the summation of these 6 features.

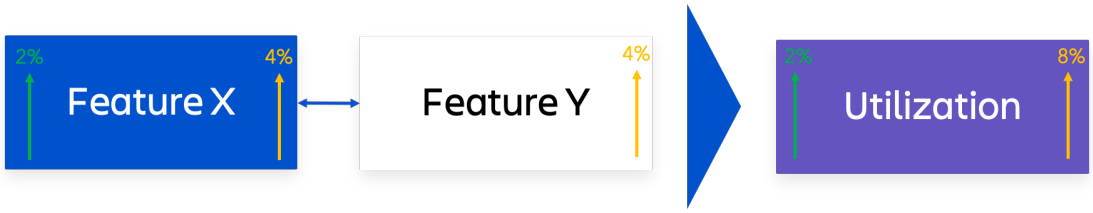
$$Y = X_1 + X_2 + X_3 + Z_1 + Z_2 + Z_3 \tag{3.3}$$

Three sets of models and explanatory methods were trained on this dataset to determine how different models would estimate the causal structure.

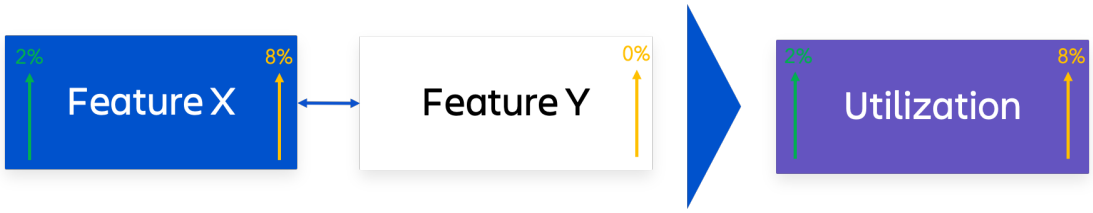
- Random Forest ALE
- XGBoost ALE
- Explainable Boosting Machine

The resultant explanatory plots of X_1, X_2 and X_3 were examined in Appendix A revealing two distinct options.

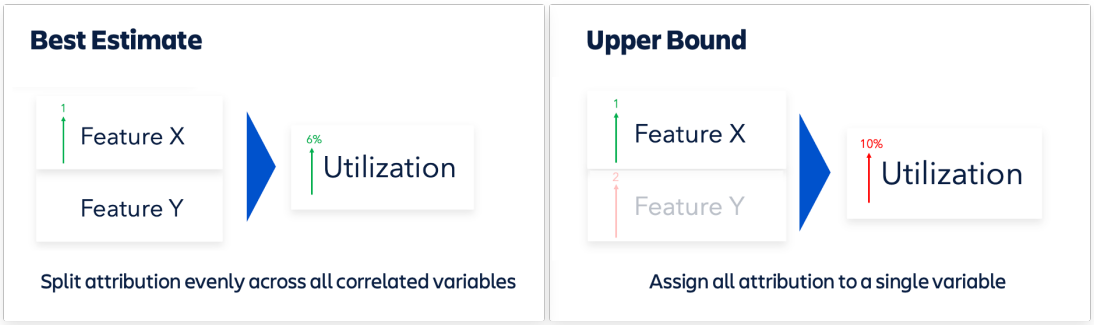
Option 1 (Random Forest ALE and Explainable Boosting Machine) is to split attribution evenly across all correlated variables. We only assign the full attribution when we consider all the other correlated features too. All explanatory plots for X1, X2 and X3 were linear lines with positive gradients proportional to their relative size, implying that attribution had been evenly split in accordance with the true causal structure.



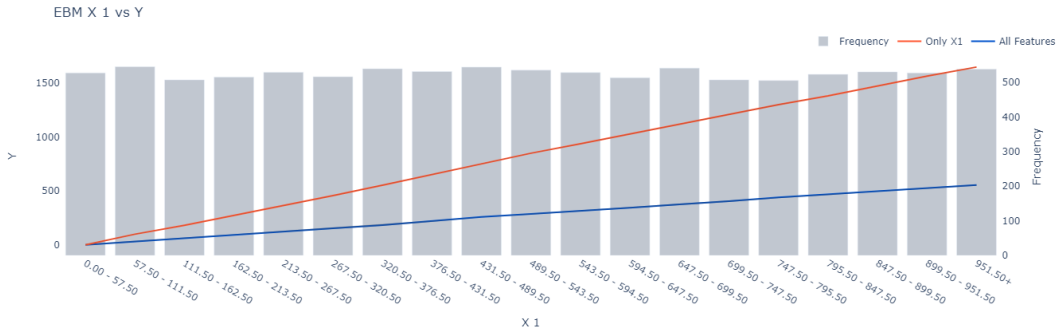
Option 2 (XGBoost ALE) is to give all the attribution to a single variable. This is identical to saying that increasing one variable will drive all flow-on effects of other variables. Only the explanatory plots of X1 had a positive gradient, while X2 and X3 were flat, implying that X1 had received all the attribution. This is the same as only including X1 in a Random Forest ALE and Explainable Boosting Machine and incorrectly estimates the true causal structure.



Knowing this, the proposed approach is to provide a range of attribution where Option 1, splitting it evenly is our Best Estimate, and Option 2, giving it all to the target is the Upper Bound. The best estimate assumes that causality is split evenly while the upper bound assumes that causality is attributed to one. These can be constructed through varying the features included in the model.



The following graph demonstrates how this method can be used to estimate X_1. As a Best Estimate, an Explainable Boosting Machine is trained on the full set of features, $X_1, X_2, X_3, Z_1, Z_2, Z_3$. As an Upper Bound, an Explainable Boosting Machine was trained only on X_1, Z_1, Z_2, Z_3. The features included in the model are those that are assumed to be part of the causal process with the model splitting attribution evenly in the presence of collinearity.



FastExplain was used to perform the Machine Learning attribution in this paper. It provides an out-of-the-box tool for researchers to quickly model and explore data, with flexibility to fine-tune if needed.

- Automated cleaning and fitting of machine learning models with hyperparameter search
- Aesthetic display of explanatory methods ready for display
- Connected interface for all data, models and related explanatory methods

Follow the following link to a Google Colab notebook that contains the code to run the simulations in this paper: `https://colab.research.google.com/github/felixzhu17/FastExplain/blob/main/demos/Model%20Attribution%20Comparison.ipynb`

For more information about FastExplain, visit Github to learn how to get started for yourself: `https://github.com/felixzhu17/FastExplain`

Bibliography

Apley, D. W. & Zhu, J. (2020), 'Visualizing the effects of predictor variables in black box supervised learning models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**(4), 1059–1086.

Basu, I. & Maji, S. (2022), Multicollinearity correction and combined feature effect in shapley values, *in* 'Australasian Joint Conference on Artificial Intelligence', Springer, pp. 79–90.

Cartwright, N. (2010), 'What are randomised controlled trials good for?', *Philosophical studies* **147**(1), 59–70.

Nori, H., Jenkins, S., Koch, P. & Caruana, R. (2019), 'Interpretml: A unified framework for machine learning interpretability', *arXiv preprint arXiv:1909.09223* .

Pearl, J. (1993), '[bayesian analysis in expert systems]: comment: graphical models, causality and intervention', *Statistical Science* **8**(3), 266–269.

Zainodin, H. & Yap, S. (2013), Overcoming multicollinearity in multiple regression using correlation coefficient, *in* 'AIP Conference Proceedings', Vol. 1557, American Institute of Physics, pp. 416–419.

Zhao, Q. & Hastie, T. (2021), 'Causal interpretations of black-box models', *Journal of Business & Economic Statistics* **39**(1), 272–281.

Zidek, J. V., Wong, H., Le, N. D. & Burnett, R. (1996), 'Causality, measurement error and multicollinearity in epidemiology', *Environmetrics* **7**(4), 441–451.

Appendix A

Comparison of Attribution Under Multicollinearity

A.1 Explainable Boosting Machine

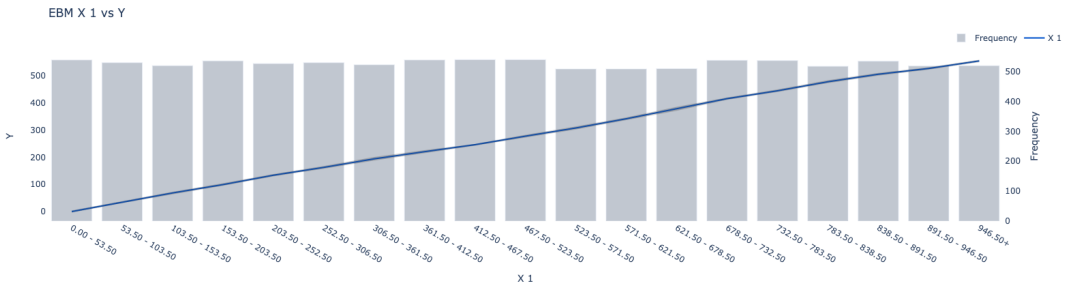


Figure A.1: Explainable Boosting Machine X1

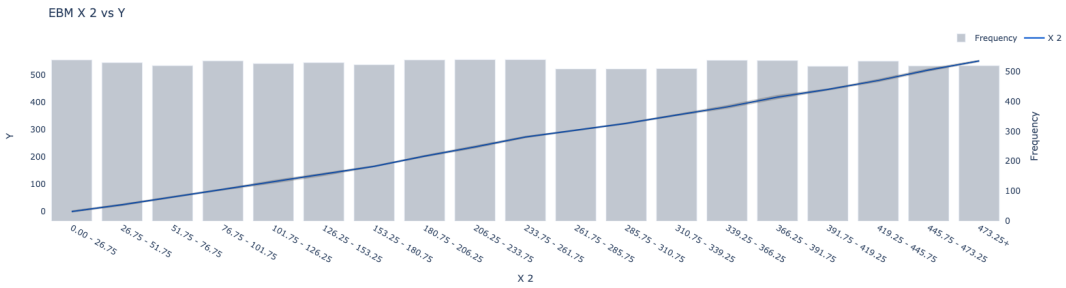


Figure A.2: Explainable Boosting Machine X2

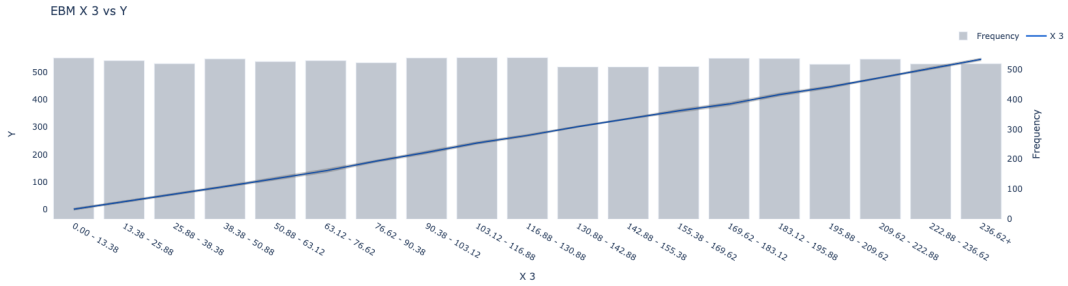


Figure A.3: Explainable Boosting Machine X3

A.2 Random Forest ALE

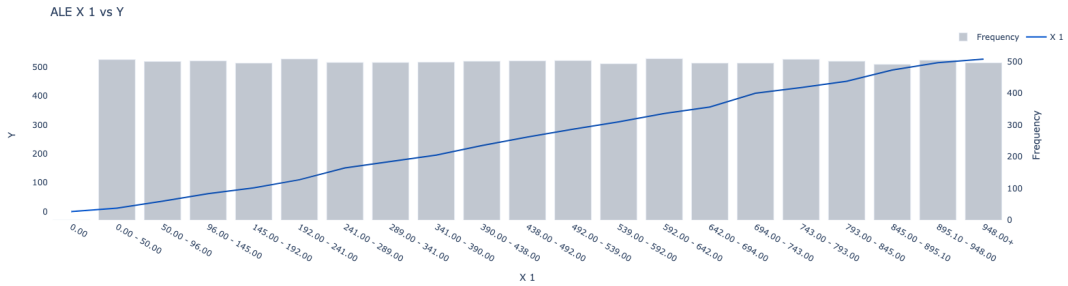


Figure A.4: Random Forest ALE X1

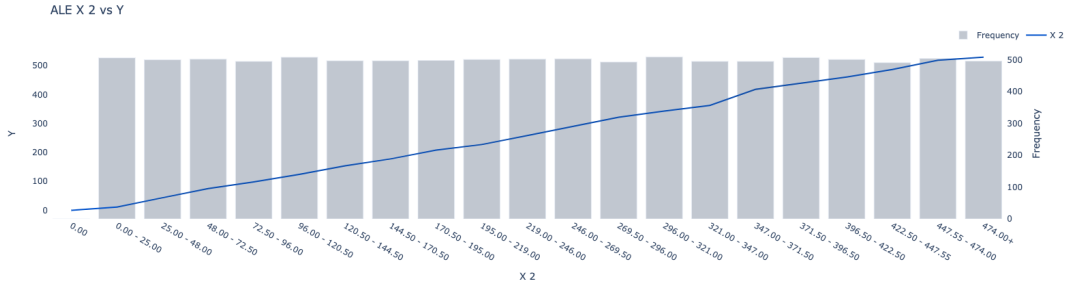


Figure A.5: Random Forest ALE X2

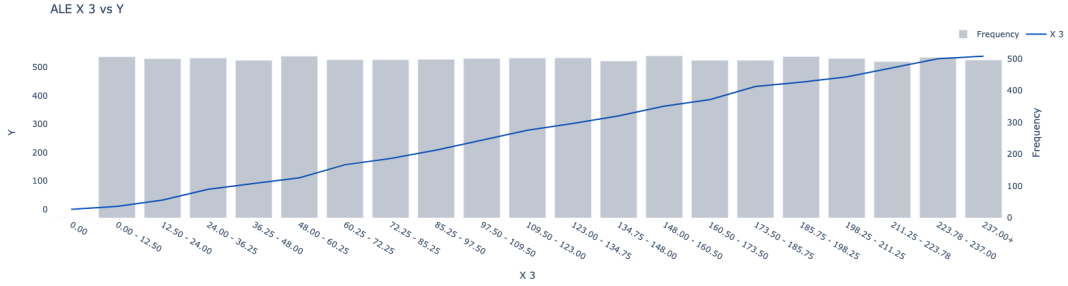


Figure A.6: Random Forest ALE X3

A.3 XGBoost ALE

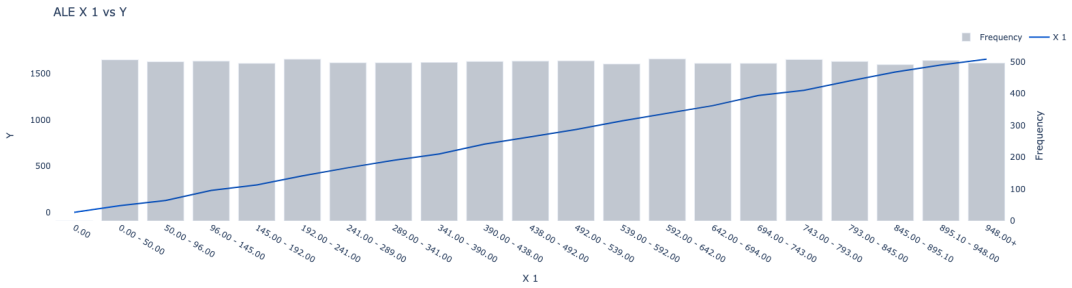


Figure A.7: XGBoost ALE X1

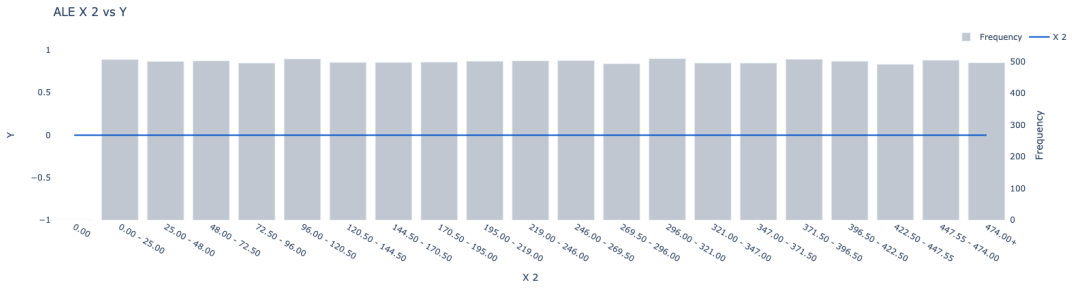


Figure A.8: XGBoost ALE X2

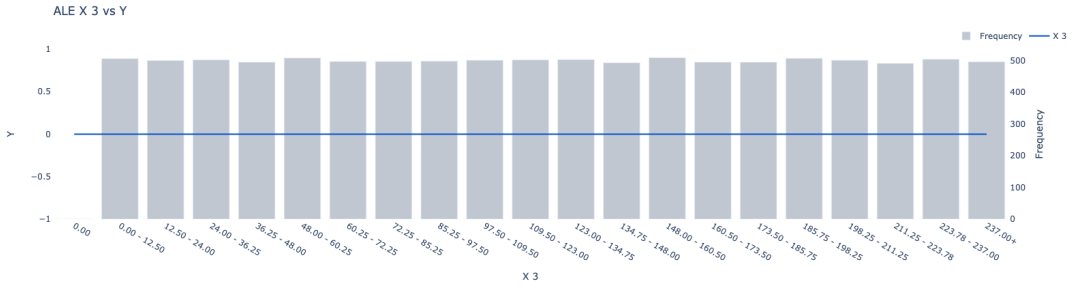


Figure A.9: XGBoost ALE X3