UNIVERSIDAD
TORCUATO DI TELLA

Departamento de Economía
Tipo de documento: Tesis de maestría

CC BY NC SA

Maestría en Econometría

Uso de herramientas de machine learning para la estimación de efectos de tratamiento heterogéneos

Autoría: Paletta, Martín Ezequiel
Fecha: 2025

¿Cómo citar este trabajo?

Paletta, M. (2025). "Uso de herramientas de machine learning para la estimación de efectos de tratamiento heterogéneos". [Tesis de maestría. Universidad Torcuato Di Tella]. Repositorio Digital Universidad Torcuato Di Tella
https://repositorio.utdt.edu/handle/20.500.13098/13567

El presente documento se encuentra alojado en el Repositorio Digital de la **Universidad Torcuato Di Tella** bajo una licencia Creative Commons Atribución-No Comercial-Compartir Igual 4.0 Internacional
Dirección: https://repositorio.utdt.edu

Biblioteca Di Tella

UNIVERSIDAD TORCUATO DI TELLA

DEPARTAMENTO DE ECONOMÍA

Maestría en Econometría

Uso de herramientas de machine learning para la estimación de efectos de tratamiento heterogéneos.

Alumno: Martín Ezequiel Paletta .....

Tutor: Gabriel Martos Venturini .....

Fecha: 5 de mayo de 2025 .....

Índice general

Resumen:

La investigación se centra en la presentación, análisis y comparación de diversos meta-algoritmos, los cuales permiten modelar de forma flexible, usando algoritmos de machine learning tradicionales, los efectos individuales de un determinado tratamiento.

A través de simulaciones controladas que buscan representar distintas estructuras de efectos heterogéneos, se evalúa el rendimiento de estos algoritmos usando distintas métricas de riesgo. Se concluye que, si bien algunos modelos muestran mejor desempeño predictivo, esto no implica una mejor estimación causal.

Como aplicación empírica, se implementa la metodología en un caso real aplicado a una empresa de préstamos personales, donde se analiza la sensibilidad de los clientes a la tasa de interés en función de múltiples variables. Los resultados permiten segmentar clientes según su propensión a aceptar distintas condiciones crediticias, lo cual representa una herramienta estratégica para la personalización de ofertas con el fin de maximizar la rentabilidad del negocio. El trabajo concluye mencionando posibles extensiones metodológicas, incluyendo múltiples tratamientos - sean estos discretos o continuos- y el uso de variables instrumentales.

Palabras Clave: Machine Learning, Inferencia causal, Efectos de tratamiento heterogéneos, Econometría, Big Data.

Capítulo 1

Introducción

1.1. Contexto y motivación

La noción de causalidad hace referencia a una relación entre dos variables en la que una de ellas genera un efecto observable en la otra. En pos de una mayor claridad expositiva, hacemos énfasis en la palabra *genera* para distinguir esta relación de otras formas de asociación estadística, como la correlación, la cual simplemente señala que dos variables tienden a variar conjuntamente, sin implicar necesariamente una relación causal. Para que una relación sea considerada causal, es requisito fundamental que modificaciones en una variable provoquen cambios sistemáticos en la otra.

En las últimas décadas, la disponibilidad masiva de datos ha dado lugar a un fenómeno ampliamente conocido por el término en inglés de *big data*. Bajo esta denominación no solo se alude al volumen de información disponible, sino también a su variedad y complejidad: datos estructurados y no estructurados, provenientes de fuentes como texto, imágenes, audio, video, entre otros. Este nuevo paradigma ha planteado desafíos y oportunidades para el análisis estadístico y, en particular, para la identificación de relaciones causales en contextos complejos.

Tradicionalmente, tanto la estadística como muchas corrientes del pensamiento económico han trabajado con datos estructurados, es decir, observaciones organizadas en filas y columnas que representan variables cuantificables. Si, por ejemplo, a un estadístico del siglo XX se le pidiera construir un modelo para tasar obras de arte, lo más probable es que recurriera a una regresión sobre variables como la antigüedad de la obra, su estado de conservación, las dimensiones, el origen geográfico o incluso la presencia de ciertos colores. Esta aproximación requiere una etapa previa de ingeniería de atributos, en la que el conocimiento experto del analista guía la selección y transformación de la información relevante.

Sin embargo, en los últimos años han emergido nuevas metodologías basadas en otra clase de algoritmos -que posteriormente identificaremos como *machine learning*- que abordan este tipo de problemas desde una perspectiva radicalmente distinta. Por dar un ejemplo, algoritmos como las redes neuronales convolucionales (CNN) permiten procesar directamente imágenes sin necesidad de definir manualmente las variables relevantes. En cierto sentido, la tarea de crear variables ha sido delegada al algoritmo, el cual es capaz de aprender representaciones complejas y no triviales a partir de los datos crudos. Esto tiene el potencial de reducir el sesgo introducido por el juicio del analista y abrir la puerta a la identificación de patrones más sofisticados que podrían pasar desapercibidos en un análisis convencional.

No obstante, este avance no está exento de desafíos. Entre ellos destaca la creciente preocupación por la explicabilidad (o interpretabilidad) de los modelos, dado que muchos de estos métodos funcionan como cajas negras cuya lógica interna resulta difícil de desentrañar.

A pesar de los notables avances que han traído consigo los métodos de machine learning, es impor-

tante reconocer que su fortaleza principal reside en tareas de predicción. Es decir, estos algoritmos están diseñados para minimizar el error al anticipar valores de una variable de interés dados ciertas covariables. Esto los convierte en herramientas sumamente efectivas para aplicaciones tan distintas como la clasificación de imágenes médicas o la predicción de default de un préstamo personal (caso que ocupará una sección del presente trabajo.)

Sin embargo, predecir correctamente no equivale a identificar una relación causal. Un modelo puede exhibir una excelente capacidad predictiva y, aun así, estar capturando asociaciones espurias o correlaciones que no reflejan mecanismos causales genuinos. Por ejemplo, un modelo que predice el rendimiento académico de los estudiantes a partir de variables como el número de libros en el hogar o la marca de la computadora que utilizan podría alcanzar una buena capacidad predictiva y ser realmente útil en algunos contextos. Pero esto no implica que tales variables causen directamente mejores resultados académicos.

La inferencia causal tiene una naturaleza epistemológica distinta: no basta con anticipar lo que va a suceder, sino que se busca entender qué pasaría si intervenimos activamente en el sistema alterando ciertas variables. En otras palabras, interesa construir modelos que permitan responder preguntas contrafactuales del tipo ¿qué hubiera pasado si...?. Esta clase de razonamiento demanda un marco conceptual distinto, basado en supuestos más fuertes y en una lógica que no puede deducirse únicamente de los datos observados.

Esa naturaleza contrafactual -que luego definiremos formalmente como un problema de datos faltantes- es la razón por la que no es directa la aplicación del *machine learning* en la inferencia causal, aún en un contexto de *big data*.

Ante este problema, ha surgido una creciente literatura (Caron et al. 2021, Künzel et al. 2019, Chernozhukov et al. 2018) que busca complementar los métodos de machine learning con herramientas provenientes de la inferencia causal, ya sea a través de la incorporación de meta-algoritmos o mediante la integración de supuestos estructurales. Estas estrategias permiten explotar el poder predictivo del *machine learning*, pero enmarcado dentro de una arquitectura que posibilite identificar —bajo ciertas condiciones— relaciones realmente causales.

1.2. Preguntas e hipótesis de investigación

Este trabajo busca responder una pregunta que puede considerarse más general que lo directamente enunciado en la sección anterior. Si bien allí se discutió el desafío general de identificar relaciones causales, el foco específico de esta investigación estará puesto en la detección de efectos causales heterogéneos.

En otras palabras, no solo nos interesa determinar si existe un efecto causal entre dos variables, sino también cómo varía ese efecto entre distintos individuos. Este tipo de heterogeneidad es especialmente relevante en ámbitos como la economía, la educación o la salud, donde los tratamientos o intervenciones pueden tener impactos significativamente distintos dependiendo de características observables (como género, nivel socioeconómico o nivel educativo) o incluso no observables de los individuos.

La hipótesis central de este trabajo es que el uso combinado de técnicas de machine learning con marcos teóricos de inferencia causal permite no solo identificar efectos causales promedio, sino también estimar de manera robusta estas variaciones en el efecto del tratamiento (heterogeneidad). En particular, se explorará en qué medida ciertos métodos modernos permiten usar la capacidad predictiva de los algoritmos de *machine learning* para la identificación de parámetros causales y bajo qué condiciones pueden considerarse válidos para este fin.

1.3. Objetivos generales y específicos

Se prentende estudiar cómo es posible utilizar la capacidad predictiva de los algoritmos *modernos* de machine learning -repasando su especificidad respecto a los algoritmos econométricos tradicionales- para poder identificar efectos causales en un contexto de datos multidimensionales.

Para ello, resulta necesario hacer una revisión crítica de la literatura reciente para asi identificar cuales son las distintas metodologías existentes, establecer criterios para compararlas en función del cumplimiento o no de los supuestos subyacentes a las mismas y analizar sus propiedades en lo que a la inferencia estadística respecta.

1.4. Contribuciones esperadas del trabajo

Idealmente, se espera que este trabajo sirva como una guía práctica para la identificación de efectos causales heterogéneos usando machine learning. No resulta esperable que exista una solución metodológica que funcione mejor para todos los problemas pero si parece razonable y útil delinear un conjunto de buenas prácticas que orienten la selección de modelos, el análisis y la interpretación de los resultados en función de las características del problema a resolver, tales como el tamaño muestral, la dimensionalidad del conjunto de variables, la presencia de factores no observables, y la necesidad de interpretabilidad.

También se busca clarificar la relación entre las métricas de naturaleza estrictamente predictiva (comúnmente utilizadas para evaluar modelos de *machine learning*) y aquellas métricas orientadas a valorar la calidad de la inferencia causal. Se discute hasta qué punto un buen desempeño predictivo se traduce en una estimación precisa de efectos causales.

1.5. Organización del trabajo

El presente trabajo estará organizado del siguiente modo: Primero se hará un repaso del modelo causal de Rubin donde se definirá el herramental teórico y la notación que se usará en los siguientes capítulos. Además, se discutirá cuál es la diferencia epistemológica entre un problema que puede ser resuelto sin mayores suspicacias usando *machine learning* en contraposición a la estimación de un conjunto de parámetros que llamaremos causales.

Luego se realizaran una serie de simulaciones con distintos tamaños de muestra y especificaciones para probar la robustez y las ventajas y desventajas de distintos algoritmos (que luego llamaremos *meta-algoritmos*) presentes en la literatura. Esta será la parte central del trabajo y la que tenga una mayor extensión.

Usando datos reales provenientes de un *lender* se estimará la sensibilidad a la tasa de interés que tienen sus distintos prospectos con el fin de encontrar una heterogeneidad no observada que tenga una interpretación económica relacionada a un perfil cuya propensión a tomar o no un producto crediticio es poco elástica ante variaciones en la tasa de interés.

Finalmente, se resumirán las conclusiones obtenidas y se mencionarán posibles extensiones al presente trabajo, donde se mencionen tipos de tratamiento más generales o la relajación de algunos supuestos.

Capítulo 2

Inferencia causal

2.1. Introducción al Modelo Causal de Rubin

En la introducción se mencionó que la causalidad puede entenderse como un tipo particular de relación entre dos variables aleatorias. En general, los problemas que nos interesan en el presente trabajo se formulan en términos de estimar cuál es el efecto de aplicar un tratamiento W sobre una variable de resultado Y.

Con pérdida de generalidad, supondremos que W sigue una distribución Bernoulli donde, para una observación i, $W_i \in \{0, 1\}$, con $W_i = 0$ para el caso de que la unidad no haya sido tratada y $W_i = 1$ en el evento complementario. Además, para cada observación i le observamos un vector de covariables X_i.

Denotemos a la distribución conjunta de las covariables, la variable de respuesta y la asignación del tratamiento como D. Es decir, $(Y, X, W) \sim D$.

Con el objetivo de formalizar esta discusión, adoptaremos el Modelo Causal de Rubin (Donald B. Rubin 1974) - RCM, por sus siglas en inglés-. Este enfoque busca dar una definición rigurosa del fenómeno causal, *matematizando* la clásica pregunta de: "¿Qué habría ocurrido si...". Para ello, introduce el concepto de resultados potenciales, que refiere a dos nuevas variables aleatorias. Cada una de estas variables representa el valor que tomaría la variable de resultado bajo distintos escenarios de tratamiento. Formalmente, para la unidad i:

$$Y_i(0) : \text{Valor de Y para el caso de que la unidad i no haya sido tratada}$$

$$Y_i(1) : \text{Valor de Y para el caso de que la unidad i haya sido tratada}$$

Luego, resulta de interés el efecto individual del tratamiento (ITE) $Y_i(1) - Y_i(0)$. Claramente, para el caso en el que el tratamiento no haya surtido ningún efecto en la unidad i, sus resultados potenciales serán idénticos teniendo $Y_i(1) - Y_i(0) = 0$.

Poblacionalmente, podemos definir el efecto promedio del tratamiento (ATE) como:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

Una generalización del efecto promedio del tratamiento (ATE) es el efecto promedio condicional (CATE, *Conditional Average Treatment Effect*). Este mide el efecto esperado del tratamiento para una subpoblación caracterizada por un conjunto de covariables observadas $X = x$:

$$\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$$

Esta última función será la que nos interesará estimar en las secciones siguientes. Una notación conveniente consiste en definir la respuesta esperada para el grupo de control y de tratamiento y expresar a τ como:

$$\begin{aligned} \mu_0(x) &:= \mathbb{E}[Y(0) \mid X = x] \\ \mu_1(x) &:= \mathbb{E}[Y(1) \mid X = x] \end{aligned} \quad \Rightarrow \quad \tau(x) := \mu_1(x) - \mu_0(x)$$

Decimos que el efecto causal es heterogéneo cuando la funcion τ no es constante a lo largo de todas las observaciones.

Además, podemos escribir - lo que será bastante práctico cuando querramos hacer selección de modelos causales- a la variable objetivo usando una formulación en la que intervenga τ. Para esto, primero definimos las siguientes funciones:

$$\begin{aligned} &\text{(Media condicional de Y)} && m(x) \stackrel{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y \mid X = x], \\ &\text{(Propensión a recibir el tratamiento)} && e(x) \stackrel{\text{def}}{=} \mathbb{P}[W = 1 \mid X = x], \end{aligned}$$

Para una observación i, podemos expresar sus resultados potenciales como:

$$y(w_i) = m(x) + \big(w_i - e(x)\big)\tau(x) + \varepsilon(x; a) \quad \text{con} \quad \mathbb{E}[\varepsilon(X; W) \mid X, W] = 0$$

Esta es la llamada *descomposición de Robinson* [1].

Desafortunadamente, en la práctica solo uno de los dos resultados potenciales es observable para cada unidad. Este fenómeno es conocido en la literatura como el problema fundamental de la inferencia causal y nos impide computar de forma exacta el ITE, el ATE o el CATE. Para poder estimar de manera razonable dichos efectos, necesitamos establecer algunos supuestos mínimos que discutimos a continuación.

2.1.1. Identificación de efectos causales y supuestos necesarios

La identificación de efectos causales a partir de datos observacionales requiere el cumplimiento de ciertos supuestos. Estos permiten vincular los resultados potenciales, que son en esencia contrafactuales, con las variables observadas. A continuación, se presentan los supuestos comúnmente aceptados en el modelo de resultados potenciales:

1. **Existencia de resultados potenciales**: Para cada unidad i, existen dos variables aleatorias $Y_i(0)$ y $Y_i(1)$.

2. **Asignación única de tratamiento**: Cada unidad es asignada a un solo estado de tratamiento. Es decir, para cada i, se observa únicamente $W_i \in \{0, 1\}$.

3. **SUTVA (*Stable Unit Treatment Value Assumption*)**: No existe interferencia entre unidades (el resultado de una unidad no depende del tratamiento asignado a otras unidades). Este supuesto garantiza que $Y_i(w)$ sea una función bien definida únicamente del tratamiento recibido por la unidad i. El ejemplo canónico en el que no se satisface este supuesto es cuando las distintas unidades compiten por algún recurso.

4. **Ignorabilidad (o independencia condicional)**: Se asume que, condicional a un conjunto de covariables observadas X, la asignación del tratamiento es independiente de los resultados potenciales:

$$(Y(0), Y(1)) \perp W \mid X$$

[1] Robinson 1988

Esto es equivalente a suponer que todos los factores que afectan simultáneamente a la variable de resultado y a la asignación del tratamiento fueron incluidos en el conjunto de covariables (es decir, no existen *confoundings*).

5. **Superposición (*overlap*)**: Para toda combinación de valores de las covariables X, existe una probabilidad positiva de recibir cada uno de los niveles de tratamiento. Es decir:

$$0 < P(W = 1 \mid X = x) < 1 \quad \forall x \in \text{soporte}(X)$$

Este supuesto garantiza que haya unidades tratadas y no tratadas en todas las regiones del espacio de covariables, lo que permite la comparación entre grupos similares.

Bajo el cumplimiento conjunto de estos supuestos, es posible estimar efectos causales utilizando métodos estadísticos sin requerir la observación directa de los contrafactuales.

El supuesto de ignorabilidad puede ser reformulado usando un resultado bastante útil en contextos observacionales (Rosenbaum y Donald B Rubin 1983) que dice que en realidad no es necesario condicionar por el conjunto de covariables sino que basta hacerlo por un *score* de propensión calculado usando las mismas. Es decir,

$$(Y(0), Y(1)) \perp W \mid X \Rightarrow (Y(0), Y(1)) \perp W \mid e(X)$$

Para el caso de datos multivariados donde la dimensionalidad de X puede ser bastante alta, $e(X)$ actúa como un resumen unidimensional suficiente de los datos.

En Künzel et al. 2019 se hace referencia a que los distintos estimadores del efecto condicional del tratamiento que serán desarrollados en secciones siguientes, son bastante robustos en presencia de *confoundings* siempre y cuando el supuesto de ignorabilidad se cumpla. Esto es de gran utilidad para el caso de datos observacionales.

2.1.2. Análisis de subgrupos para identificar heterogeneidad

En Foster et al. 2011 se menciona que, en un contexto de ensayos clinicos confirmatorios, además de tener resultados que sean propios de la población a estudiar, resulta de interés demostrar la existencia de una cierta heterogeneidad en el efecto del tratamiento en cuestión:

> ... it is quite plausible that there are subgroups of patients for whom the new treatment is especially effective. Likewise, there could be subgroups of patients for whom the new treatment is not effective, or less effective than the standard therapy. There is a strong desire to find such subgroups if they exist. From a statistical perspective, searching for subgroups is known to be a dangerous exercise, with the high possibility of finding false positives.

Los autores mencionan como "*a dangerous exercise*" el hecho de que el investigador pueda encontrar grupos donde el tratamiento parezca más eficaz probando distintos cortes en el espacio de covariables hasta encontrar algún resultado estadísticamente significativo. Para evitar esto, mencionan estrategias consistentes en declarar *ex-ante* las subpoblaciones candidatas usando intuiciones derivadas de la literatura médica o bien, predefinir el enfoque estadístico que se va a utilizar para encontrarlos [2] .

La metodología tradicional (revisitada en Kehl y Ulm 2006) consiste en utilizar algún modelo econométrico -como una regresión lineal o logística- e incluir términos de interacción entre las distintas

[2]Este fenómeno es conocido en la literatura estadística como "p-hacking" y se define como un problema de testeo frecuentista de hipótesis múltiple combinado con una selección de los resultados a publicar. Esta y otras malas prácticas en el análisis estadístico aplicado a trabajos de investigación pueden hallarse en Stefan y Schönbrodt 2023 (no casualmente titulado como *Big little lies*)

covariables y el tratamiento, luego la heterogeneidad en el tratamiento se determinará mediante la significatividad de los coeficientes obtenidos.

En un contexto de *big data* dicho enfoque resulta problemático ya que para conjuntos de datos con muchas dimensiones, la cantidad de posibles interacciones crece exponencialmente.

Capítulo 3

Machine Learning

> "All models are wrong, but some are useful."
> — *George E. P. Box*

El *Machine Learning* es un subcampo de la inteligencia artificial que busca desarrollar métodos algorítmicos capaces de aprender patrones a partir de los datos y eventualmente utilizarlos para hacer predicciones. A diferencia de los métodos econométricos tradicionales, cuyo enfoque principal está en la estimación e interpretación de relaciones causales bajo supuestos estructurales bien definidos, el aprendizaje automático prioriza la capacidad predictiva y de generalización.

Desde un punto de vista metodológico, asegurar la capacidad de generalización será de gran importancia ya que, como veremos, es bastante fácil encontrar una representación algorítmica perfecta de los datos sin que esta tenga algún tipo de utilidad práctica.

3.1. Tipos de aprendizaje

Los problemas en ML se pueden clasificar -de forma no exhaustiva- en dos grandes categorías: aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, se busca modelar la relación entre un conjunto de variables explicativas y una variable objetivo o de resultado, con el fin de predecir esta última en observaciones futuras. Por otro lado, el aprendizaje no supervisado se ocupa de identificar patrones latentes o estructuras emergentes en los datos sin que exista una variable objetivo explícita.

Para el presente trabajo, nos concentraremos exclusivamente en el aprendizaje supervisado ya que los herramentales existentes en la literatura existente de la inferencia causal usando estas nueva familia de algoritmo descansa exclusivamente en ellos.

3.2. Formulación del problema supervisado

Partimos de un conjunto de datos denominado *conjunto de entrenamiento*, que consiste en n observaciones de un vector de variables explicativas $x \in \mathbb{R}^p$ y una variable de resultado $Y \in \mathbb{R}$.

Además, asumiremos que existe una función desconocida f tal que:

$$Y = f(X) + \epsilon$$

donde el término $f(X)$ da cuenta de la influencia de las variables explicativas en la variable de respuesta. Sin pérdida de generalidad asumiremos que el término de error ϵ tiene media 0.

Dada una muestra $D = \{(x_i, y_i)\}$, el objetivo de un modelo de aprendizaje supervisado consiste en contruir (estimar) una función $\hat{f} : \mathbb{R}^p \to \mathbb{R}$ tal que $\hat{f}(x)$ se aproxime lo mejor posible a la esperanza

condicional $E[Y|x]$. Es decir, se busca una representación funcional que capture la relación subyacente entre los predictores y la variable objetivo.

Una de las particularidades de resolver un problema predictivo a uno causal es que si bien esta función puede tomar muchas formas, desde modelos lineales hasta arquitecturas complejas como redes neuronales profundas, en los problemas predictivos existe bastante consenso en criterios y metodologías para elegir cuál será el algoritmo más adecuado para solucionar un problema práctico.

Los métodos econométricos tradicionales están orientados principalmente al testeo de cierta hipótesis. En cambio, en *machine learning* el foco está puesto en la precisión predictiva, siendo especialmente robustos en la detección de relaciones no lineales en el contexto de *big data* mencionado en la primera sección.

En Hastie et al. 2009 se presenta una descripción rigurosa de los algoritmos -y sus hiperparámetros- comúnmente utilizados, por lo que se asumirán conocidos a lo largo de este trabajo, ya que nos enfocaremos en los aspectos metodológicos.

3.3. Selección de modelos y validación cruzada

La selección de modelos consiste en elegir un único modelo ganador entre los distintos binomios conformados por los distintos algoritmos existentes y sus hiperparámetros que los gobiernan [1]. Para ello, ex-ante, es necesario especificar un criterio que determine objetivamente cuándo un algoritmo funciona mejor que otro.

Si bien nos interesa el modelo que pueda aproximar mejor los datos en toda la población, como la misma en general no es accesible debemos estimar la métrica de interés en alguna muestra. Para tener una mejor estimación del error de predicción poblacional, es necesario estimarlo en un conjunto de datos distinto al que se usó para entrenar [2]. A esta técnica se la denomina validación cruzada.

A modo de resumen, un algoritmo simple para realizar selección de modelos en un contexto predictivo puede ser:

Algorithm 1 Selección de modelos en un contexto predictivo

Require: Conjunto de datos $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, conjunto de modelos candidatos $\mathcal{M} = \{M_1, M_2, \ldots, M_k\}$, métrica de evaluación ℓ
Ensure: Modelo seleccionado M^*

1: Dividir $\mathcal{D}$ en dos subconjuntos: entrenamiento $\mathcal{D}_{\text{train}}$ y testeo $\mathcal{D}_{\text{test}}$
2: **for** cada modelo $M_j \in \mathcal{M}$ **do**
3: Entrenar M_j usando $\mathcal{D}_{\text{train}}$
4: Calcular error de testeo: $E_j = \ell(M_j, \mathcal{D}_{\text{test}})$
5: **end for**
6: Seleccionar el modelo con menor error: $M^* = \arg\min_{M_j} E_j$
7: **return** M^* =0

Existen otras variantes donde el conjunto de datos es dividido en K particiones y los modelos son entrenados en K-1 de ellas para ser validado con la restante, repitiendo el procedimiento K veces para que cada una de las particiones actue en algún momento como conjunto de testeo. Esto resulta computacionalmente más costoso ya que requiere ajustar varias veces el modelo.

[1] Si bien el concepto de que no se puede analizar *in abstracto* un algoritmo en ausencia de sus hiperparámetros es correcto, por brevedad nos referiremos exclusivamente al algoritmo

[2] Una idea intuitiva del por qué de esta decisión metodológica es que los modelos de *machine learning*, a diferencia de los modelos econométricos tradicionales, pueden contener cientos de miles de parámetros y esto les permite memorizarse el conjunto de datos de entrenamiento, lo que puede llevar a un sobreajuste y, por ende, a una pobre capacidad de generalización.

Uno de los aspectos más importantes radica en la elección de la métrica con la que se hará la selección de modelos. Esta puede tener una naturaleza más *matemática* como el error cuadrático medio para el caso de una regresión o del área bajo la curva ROC para una clasificación. Paralelamente, es interesante tomar -por ejemplo, en un contexto de negocios- alguna métrica que tenga una interpretación económica en base al problema a resolver.

3.4. Econometría y *machine learning*

En Athey y Imbens 2019 se discute que la adopción de esta nueva clase de algoritmos resultó más tardía (en comparación a las ciencias naturales) en el campo de la investigación económica empírica ya que esta última estuvo más enfocada en estimadores a los que se les pueden atribuir de forma teórica propiedades como la consistencia, normalidad o eficiencia -fundamentalmente, para poder realizar inferencia estadística-, mientras que la producción científica asociada al *machine learning* se concentró en optimizar la tasa de error en las estimaciones puntuales de los mismos.

Otra dificultad radica en que, si bien existen aplicaciones económicas para las que las estimaciones puntuales son suficientes, la complejidad interna de estos algoritmos (a menudo, cientos de miles de parámetros) dificulta la imposición de cierta estructura o de restricciones con contenido ecónomico -que, en general, fue aprendida en investigaciones previas- ya que solo buscan aprender patrones presentes en los datos con los que se entrenaron.

3.5. Desafíos en la aplicación de *machine learning* para inferencia causal

Los modelos de *machine learning* suelen seleccionarse en función de su capacidad predictiva, es decir, por su habilidad para minimizar el error esperado en nuevas observaciones. Sin embargo, esta propiedad —si bien resulta valiosa en numerosos contextos— no garantiza que el modelo sea capaz de recuperar relaciones causales a partir de los datos. En otras palabras, un buen desempeño predictivo no implica, necesariamente, una correcta identificación de efectos causales.

La presencia de correlación entre una variable explicativa y la variable de respuesta puede ser suficiente para que una *feature* tenga poder predictivo. No obstante, dicha correlación no debe interpretarse como evidencia de causalidad. Esta distinción resulta fundamental en el análisis empírico, particularmente cuando el objetivo es estimar el efecto de una intervención o política.

Por otro lado, la notable capacidad de los algoritmos modernos de *machine learning* para capturar relaciones complejas y no lineales entre múltiples variables ha motivado su uso en el ámbito de la inferencia causal. En conjuntos de datos altamente multidimensionales, donde los métodos tradicionales enfrentan limitaciones, estos modelos ofrecen herramientas prometedoras para identificar efectos causales, siempre que se combinen con supuestos adecuados y técnicas específicas de inferencia.

Capítulo 4

Meta-algoritmos y sus propiedades

"If your learning algorithm is based on correlation rather than causation, it will struggle with overfitting. To understand something is to identify its minimal sufficient causal mechanisms. Parsimony isn't just elegance, it's generalization robustness."
— François Chollet[1]

4.1. Descripción de meta-algoritmos considerados

Un meta-algoritmo es un procedimiento donde, para una variable de tratamiento binaria, se quieren modelar las superficies de respuesta de forma separada para luego usarlas -de forma más o menos directa- para recuperar la forma funcional del efecto heterogéneo que eventualmente pueda causar el tratamiento.

A continuación, se desarollarán algunos métodos existentes en la literatura:

4.1.1. S-learner

El primer estimador del CATE consiste en tratar a la variable indicadora del tratamiento como si fuese una covariable indistinguible de las demás a los fines de ajustar una superficie de respuesta de la variable de resultado. Es decir, definiremos una función poblacional $f(x, w)$ como:

$$f(x,w) = \begin{cases} \mu_0(x) = \mathbb{E}[Y(0) \mid X = x] & \text{si } w = 0 \\ \mu_1(x) = \mathbb{E}[Y(1) \mid X = x] & \text{si } w = 1 \end{cases}$$

Luego, el CATE estará dado por:

$$\tau(x) = f(x,1) - f(x,0)$$

Como la función f no es conocida, deberemos estimarla usando algún algoritmo econométrico o de *machine learning* al que llamaremos estimador base. Así, el primer estimador del CATE estará dado por:

$$\widehat{\tau(x)} = \widehat{f(x,1)} - \widehat{f(x,0)}$$

El supuesto fundamental es que los resultados potenciales condicionales promedio para cada grupo (tratado y no tratado) provienen de un mismo modelo subyacente, con una única función de media condicional $f(\cdot)$ y un término de error ε_i. Esto se expresa como:

[1] Creador del paquete informático Keras, uno de los más utilizados en *deep learning*.

$$Y_i = f(X_i, W_i) + \varepsilon_i$$

donde:

- X_i representa las covariables del individuo i,
- $W_i \in \{0, 1\}$ es el indicador de tratamiento,
- $f(\cdot)$ es una función común que modela la media condicional de Y,
- ε_i es un término de error con $\mathbb{E}[\varepsilon_i \mid X_i, W_i] = 0$.

En Caron et al. 2021 se muestra que uno de los principales inconvenientes del S-learner es que no suele tener un buen desempeño en el caso de que la complejidad de las funciones de respuesta esperada para el grupo de tratamiento y control sea muy distinta. De otro modo, este estimador puede resultar adecuado cuando se tenga algún conocimiento ex-ante del efecto de tratamiento.

4.1.2. T-learner

En el *T-learner*, se entrenan dos modelos separados para estimar los resultados potenciales de los grupos tratados y no tratados. Se define:

$$f_0(x) = \mathbb{E}[Y(0) \mid X = x], \quad f_1(x) = \mathbb{E}[Y(1) \mid X = x]$$

Estas funciones representan la media condicional del resultado para cada nivel del tratamiento.

El efecto causal condicional (CATE) para una unidad con características x se puede escribir como:

$$\tau(x) = f_1(x) - f_0(x)$$

solo que ahora debemos estimar por separado $\widehat{f_0(x)}$ y $\widehat{f_1(x)}$, usando en cada caso únicamente las observaciones correspondientes al grupo de control y experimental respectivamente.

El supuesto subyacente en el T-learner es que los resultados potenciales para cada grupo pueden ser modelados por funciones separadas:

$$Y_i = \begin{cases} f_0(X_i) + \varepsilon_{i0} & \text{si } W_i = 0 \\ f_1(X_i) + \varepsilon_{i1} & \text{si } W_i = 1 \end{cases}$$

Tener que particionar los datos para ajustar dos modelos puede ser perjudicial en algunas situaciones ya que cada modelo debe encontrar de forma independiente patrones que son realmente comunes a los dos grupos.

4.1.3. X-learner

Este estimador puede considerarse en cierto sentido como una extensión del *T-learner* (Künzel et al. 2019). Consiste en una estimación en tres etapas:

Primero estimamos las funciones de respuesta para los grupos control y experimental.

$$\mu_0(x) = \mathbb{E}[Y(0) \mid X = x], \quad \mu_1(x) = \mathbb{E}[Y(1) \mid X = x]$$

A continuación, como segunda etapa, empezamos estimando los efectos individuales del tratamiento *de forma cruzada*. Es decir, usando para las observaciones del grupo de control la estimación de la respuesta hecha con el grupo experimental y viceversa:

$$\tilde{D}_i^1 := Y_i(1) - \hat{\mu}_0(x), \quad \text{y} \quad \tilde{D}_i^0 := \hat{\mu}_1(x) - Y_i(0),$$

Llamemos a estas cantidades como los efectos de tratamiento imputados y ajustaremos dos nuevos modelos -*base learners of the second stage*- usando como variables explicativas a las covariables X y como variable de resultado a los efectos imputados. Sean $\hat{\tau_0}$ y $\hat{\tau_1}$ dichas estimaciones. Claramente, tanto $\hat{\tau_0}$ como $\hat{\tau_1}$ son estimadores de τ.

Finalmente, el estimador *X-learner* de τ será el promedio ponderado de los estimadores de la segunda etapa.

$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x),$$

donde $g(x) \in [0, 1]$ indica el peso que se le da a cada estimador. En Künzel et al. 2019 se sugiere usar un score de propensión.

4.1.4. Domain Adaptation Learner

Existe un tipo de problemas en *machine learning* llamado *domain adaptation* donde se intenta ajustar un modelo en un conjunto de datos con determinada distribución (*source domain*) pero que finalmente va a ser utilizado en otro conjunto de datos con otra distribución similar (*source domain*). La motivación de esta tarea consiste en que ciertos patrones aprendidos en un conjunto de datos pueden ser aplicables para otro conjunto de datos ligeramente distinto.

Para el caso de que exista un sesgo de selección entre los distintos grupos, en particular que la distribución de las covariables para ambos grupos sea tal que $P(X|W = 0) \neq P(X|W = 1)$, podemos ponderar -usando una estimación del score de propensión- a las observaciones del grupo de control según que tan similares sean a las del grupo experimental para así aplicar un procedimiento análogo al *X-learner*.

Las funciones de respuesta son estimadas -usando cualquier método que admita ponderadores [2]- como:

$$\mu_0(x) = \mathbb{E}[Y(0) \mid X = x; \frac{\hat{e}(x)}{1 - \hat{e}(x)}], \quad \mu_1(x) = \mathbb{E}[Y(1) \mid X = x; \frac{1 - \hat{e}(x)}{\hat{e}(x)}]$$

Del mismo modo se estiman *de forma cruzada* los efectos individuales del tratamiento:

$$\tilde{D}_i^1 := Y_i(1) - \hat{\mu}_0(x), \quad \text{y} \quad \tilde{D}_i^0 := \hat{\mu}_1(x) - Y_i(0),$$

La tercera etapa donde se estima el τ puede resolverse como un promedio ponderado por un *score* de propensión como en el *X-learner* o de forma más general usando un nuevo modelo de *machine learning* cuyo *label* es la concatenación de $\tilde{D}_i^1$ y $\tilde{D}_i^0$.

4.2. Otras formas de estimar efectos heterogéneos.

La particularidad del enfoque de los meta-algoritmos frente a otros posibles es que no hacen grandes supuestos estructurales sobre el efecto heterogéneo poblacional, permitiendo una gran flexibilidad en los estimadores a utilizar.

Métodos que si suponen, por ejemplo, linealidad en el efecto causal tienen la ventaja de tener una teoría asintótica que permita computar intervalos de confianza con una mayor facilidad, entre ellos

[2] La notación utilizada para la esperanza condicional con ponderadores explicita a estos últimos después del signo ; .

podemos destacar la metodología de *Orthogonal/Double Machine Learning* presentada en Chernozhukov et al. 2018.

Otro enfoque -menos agnóstico respecto al algoritmo de base pero sin supuestos estructurales demasiado fuertes- es el de los *Honest Random Forest*, que se trata de una variación del clásico *Random Forest* donde los datos de entrenamiento son nuevamente divididos al utilizar un subconjunto de ellos para definir la estructura de los árboles de decisión mientras que el resto es utilizada para recuperar los efectos causales en cada una de las hojas de cada estimador. En Athey y Wager 2017 se muestra que con esta metodología es posible obtener un estimador consistente y asintóticamente normal del CATE poblacional.

4.3. Selección de modelos

Para cada $\hat{f} : X \times W \to Y$ estimador de la variable resultado, podemos calcular una estimación del CATE que denotaremos como τ_f.

Sea $\mathcal{F} = \{\hat{f} : X \times W \to Y\}$ una familia de posibles estimadores. La misma puede estar compuesta por distintas variaciones de los meta-estimadores como por diferencias en sus estimadores de base (o los hiperparámetros de los mismos).

Realizar una selección de modelos sobre la familia $\mathcal{F}$ consiste en resolver el siguiente problema de minimización:

$$f_\ell^* = \arg\min_{\hat{f} \in \mathcal{F}} \ell(\hat{f}, O)$$

Donde O hace referencia al conjunto de datos observado y ℓ es una función de riesgo que hace de criterio de decisión para la elección del estimador.

Idealmente, y como mostramos en la selección de modelos de *machine learning*, querríamos ver cuánto difiere nuestra estimación del CATE respecto al verdadero valor de τ, por ejemplo, usando como criterio al error cuadrático medio de la estimación:

$$\tau\text{-risk}(\hat{f}) = \mathbb{E}_{X \sim p(X)} \left[\left(\tau(X) - \hat{\tau}_f(X) \right)^2 \right]$$

Lamentablemente -y como repetimos varias veces a lo largo del presente trabajo- esta métrica llamada *Precision in Estimation of Heterogeneous Effect* no es calculable con la información observada. Sin embargo, nos servirá de *benchmark* en un contexto simulado.

A continuación, definiremos una serie de métricas[3] de riesgo que son computables en la práctica. Las mismas se enunciarán en su versión poblacional, sobreentendiendo que en un escenario real se usarán sus análogos muestrales naturales:

Definition 1 (μ-risk)**.** *Se trata del error cuadrático medio cuando se intenta estimar Y como en cualquier problema de aprendizaje supervisado.*

$$\mu\text{-}risk(f) = \mathbb{E}_{(Y,X,W) \sim \mathcal{D}} \left[\left(Y - f(X; W) \right)^2 \right]$$

Esta fórmula busca medir el grado en el que los modelos base utilizados por los meta-algoritmos para estimar τ logran efectivamente representar estilizadamente a la variable de resultado

Para las siguientes funciones de riesgo, sean $\hat{m}$ y $\hat{e}$ sendas estimaciones de la media condicional de Y y la propensión a recibir el tratamiento respectivamente. Estas estimaciones pueden realizarse usando métodos de *machine learning* o econométricos tradicionales.

[3] Una discusión más extensa sobre las métricas útiles en la selección de modelos causales puede encontrarse en Doutreligne 2023. De todos modos, por completitud reproduciremos las definiciones de las mismas en el presente trabajo.

Claramente, las métricas definidas a continuación deben pensarse como funciones de los métodos con los que se estimen $\hat{m}$ y $\hat{e}$, lo que suma una nueva capa de complejidad al análisis ya que una mala estimación de las mismas puede afectar los resultados y quitarle robustez metodológica al modo en el que se selecciona el mejor modelo.

Definition 2 (μ-risk$^{\star}_{IPW}$)**.** *Usando como ponderador a la inversa de la probablidad de recibir el tratamiendo para corregir el sesgo de selección en la asignación del tratamiento:*

$$\mu\text{-}risk^{\star}_{IPW}(f) = \mathbb{E}_{(Y,X,W)\sim\mathcal{D}}\left[\left(\frac{W}{e(X)} + \frac{1-W}{1-e(X)}\right)(Y - f(X;W))^2\right]$$

Es fácil ver que estas dos métricas tendrán el mismo valor en el caso de que estemos comparando un *T-learner* con un *X-learner* ya que los modelos de base resultarán exactamente idénticos.

Definition 3 (τ-risk$^{\star}_{IPW}$)**.** *Usando como ponderador a la inversa de la probabilidad de recibir el tratamiento:*

$$\tau\text{-}risk^{\star}_{IPW}(f) = \mathbb{E}_{(Y,X,W)\sim\mathcal{D}}\left[\left(Y\left(\frac{W}{e(X)} - \frac{1-W}{1-e(X)}\right) - \tau_f(X)\right)^2\right]$$

Definition 4 (U-risk*)**.** *Usando la descomposición de Robinson:*

$$U\text{-}risk^{\star}(f) = \mathbb{E}_{(Y,X,W)\sim\mathcal{D}}\left[\left(\frac{Y - m(X)}{W - e(X)} - \tau_f(X)\right)^2\right]$$

Definition 5 (R-risk*)**.** *Usando las estimaciones de m y e:*

$$R\text{-}risk^{\star}(f) = \mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left((Y - m(X)) - (A - e(X))\,\tau_f(X)\right)^2\right]$$

El supuesto de *overlap* garantiza que las expresiones que tengan alguna combinación lineal del *score* de propensión estén bien definidos. Sin embargo, valores muy cercanos a 1 o 0 pueden comprometer la estabilidad numérica de los resultados por lo que en Doutreligne 2023 se recomienda su truncamiento.

La principal diferencia entre todas estas métricas radican en que las primeras consideran únicamente la capacidad de que el modelo logre reproducir de la forma más precisa posible el comportamiento de la variable de resultado, como si de un problema de predicción usual se tratase, mientras que las últimas tres incluyen alguna formulación del efecto individual del tratamiento. Adicionalmente, todas -con la excepción del μ-risk- buscan ajustar el riesgo de los estimadores por el sesgo de selección en la asignación del tratamiento.

En el siguiente capítulo se utilizarán todas estas métricas para evaluar en qué medida permiten identificar la metodología que, en teoría, debería conducir a un menor error cuadrático medio, aún en un contexto en el que dicho error no es realmente computable.

A modo de resumen, la metodología propuesta para la selección de modelos causales consiste en:

Algorithm 2 Selección de modelo en un contexto causal

Require: Conjunto de datos $\mathcal{D} = \{(x_i, w_i, y_i)\}_{i=1}^{n}$, conjunto de modelos candidatos $\mathcal{F} = \{f_1, f_2, \ldots, f_k\}$, medida de evaluación ℓ

Ensure: Modelo causal seleccionado f^*

1: Dividir $\mathcal{D}$ en subconjuntos: entrenamiento $\mathcal{D}_{\text{train}}$ y testeo $\mathcal{D}_{\text{test}}$
2: Ajustar los estimadores auxiliares $\hat{m}(x)$ y $\hat{e}(x)$ en $\mathcal{D}_{\text{train}}$ (usando *cross-validation* si es necesario)
3: **for** cada modelo $f_j \in \mathcal{F}$ **do**
4: Estimar $\hat{\tau}_j(x)$ con f_j sobre $\mathcal{D}_{\text{test}}$
5: Calcular el riesgo estimado: $R_j = \ell(\hat{\tau}_j, \mathcal{D}_{\text{test}})$
6: **end for**
7: Seleccionar el modelo con menor riesgo: $f^* = \arg\min_{f_j} R_j$
8: **return** f^* =0

Capítulo 5

Simulaciones

5.1. Justificación del uso de simulaciones

Como se mencionó anteriormente, en los problemas prácticos donde se intenta estimar un efecto causal no se cuenta con el verdadero valor del efecto (*ground truth*) para poder evaluar la precisión de las estimaciones.

El uso de simulaciones permite definir *a priori* la estructura del CATE poblacional para así descubrir las ventajas y desventajas de los distintos meta-algoritmos en diversos escenarios controlados, así como la robustez de los mismos ante violaciones más o menos graves de sus supuestos subyacentes.

Resultarán de particular interés las situaciones donde el efecto causal sea realmente inexistente, homogéneo o con distintos grados de complejidad en la forma que el efecto varía en función de las covariables observadas. Esto nos permitirá obtener conclusiones lo más generales posibles.

5.1.1. Diseño experimental de la simulación

Se simularon distintos *data generating process* (DGP) con el objetivo de representar una amplia gama de posibles estructuras de efectos heterogéneos que podrían hallarse en un escenario real.

Para ello, fue necesario especificar la distribución de las variables, el mecanismo de asignación a los grupos de tratamiento y control y los resultados potenciales mediante la forma funcional de los μ_i.

Las distintas formas de la -posible- heterogeneidad en el efecto del tratamiento están dadas por:

1. **DGP1: Ausencia de efecto causal**, es decir considerando que $\mu_0 = \mu_1$.
2. **DGP2: Efecto homogéneo**, donde para una constante fija α se tenga que $\mu_0 = \mu_1 + \alpha$.
3. **DGP3: Efecto causal lineal** donde este dependa linealmente de una única variable explicativa.
4. **DGP4: Efecto causal no lineal** con relaciones más complejas entre las distintas variables explicativas.

Todas estas configuraciones fueron evaluadas usando los cuatro meta-algoritmos mencionados en la sección anterior atendiendo a dos posibles variaciones en los estimadores de base de cada uno de ellos: Se comparará el uso de métodos lineales y de *machine learning*, más especificamente regresiones lineales estimadas por mínimos cuadrados ordinarios y un Light Gradient Boosting Machine (LGBM) [1].

Se consideraron 20 variables explicativas y tres tamaños muestrales: 1.000, 10.000 y 100.000 observaciones, con el objetivo de representar escenarios con muestras pequeñas, medianas y grandes, siempre

[1] https://lightgbm.readthedocs.io/en/stable

considerando un contexto de *big data*. El mecanismo de asignación al tratamiento fue completamente aleatorio y balanceado entre los grupos, lo que implica ausencia de *confounding*.

Cada muestra fue dividida en un conjunto de entrenamiento (*train*, 70 %) y uno de prueba (*test*, 30 %). Para cada combinación de meta-algoritmo, algoritmo base y tamaño muestral, se realizaron 20 repeticiones independientes. En cada una de ellas se calcularon todas las métricas de evaluación mencionadas en la sección de selección de modelos sobre el conjunto de prueba, y se reportaron los promedios obtenidos a modo de valor representativo.

El *mejor modelo* -para cada tamaño muestral- será definido como aquel que minimice el τ-*risk*. Dado que esta métrica no será observable en aplicaciones reales, las demás métricas se evaluarán en términos de su capacidad para seleccionar el mismo modelo ganador.

DGP 1: Sin efecto causal

$$
\begin{aligned}
X_i &\sim \mathcal{N}(0, I_d) \\
\mu_0(X_i) &= X_i^\top \beta \\
\mu_1(X_i) &= \mu_0(X_i) \\
Y_i(0) &= \mu_0(X_i) + \varepsilon_{0i}, \quad \varepsilon_{0i} \sim \mathcal{N}(0,1) \\
Y_i(1) &= \mu_1(X_i) + \varepsilon_{1i}, \quad \varepsilon_{1i} \sim \mathcal{N}(0,1) \\
W_i &\sim \text{Bernoulli}(e), \quad \text{con } e = 0{,}5 \\
Y_i &= W_i Y_i(1) + (1 - W_i) Y_i(0) \\
\tau(X_i) &= \mu_1(X_i) - \mu_0(X_i) = 0
\end{aligned}
$$

DGP 2: Efecto causal constante

$$
\begin{aligned}
X_i &\sim \mathcal{N}(0, I_d) \\
\mu_0(X_i) &= X_i^\top \beta \\
\mu_1(X_i) &= \mu_0(X_i) + \alpha \\
Y_i(0) &= \mu_0(X_i) + \varepsilon_{0i}, \quad \varepsilon_{0i} \sim \mathcal{N}(0,1) \\
Y_i(1) &= \mu_1(X_i) + \varepsilon_{1i}, \quad \varepsilon_{1i} \sim \mathcal{N}(0,1) \\
W_i &\sim \text{Bernoulli}(e), \quad e = 0{,}5 \\
Y_i &= W_i Y_i(1) + (1 - W_i) Y_i(0) \\
\tau(X_i) &= \mu_1(X_i) - \mu_0(X_i) = \alpha
\end{aligned}
$$

DGP 3: Efecto lineal en una covariable

$$
\begin{aligned}
X_i &\sim \mathcal{N}(0, I_d) \\
\mu_0(X_i) &= X_i^\top \beta \\
\mu_1(X_i) &= \mu_0(X_i) + \alpha \cdot X_{i1} \\
Y_i(0) &= \mu_0(X_i) + \varepsilon_{0i}, \quad \varepsilon_{0i} \sim \mathcal{N}(0,1) \\
Y_i(1) &= \mu_1(X_i) + \varepsilon_{1i}, \quad \varepsilon_{1i} \sim \mathcal{N}(0,1) \\
W_i &\sim \text{Bernoulli}(e), \quad e = 0{,}5 \\
Y_i &= W_i Y_i(1) + (1 - W_i) Y_i(0) \\
\tau(X_i) &= \mu_1(X_i) - \mu_0(X_i) = \alpha \cdot X_{i1}
\end{aligned}
$$

DGP 4: Efecto causal no lineal

$$
\begin{aligned}
X_i &\sim \mathcal{N}(0, I_d) \\
\mu_0(X_i) &= \sin(X_{i1}) + X_i^\top \beta \\
\mu_1(X_i) &= \mu_0(X_i) + \exp(-X_{i2}^2) + X_{i3}^2 - 2 \cdot \sin(X_{i4}) \\
Y_i(0) &= \mu_0(X_i) + \varepsilon_{0i}, \quad \varepsilon_{0i} \sim \mathcal{N}(0,1) \\
Y_i(1) &= \mu_1(X_i) + \varepsilon_{1i}, \quad \varepsilon_{1i} \sim \mathcal{N}(0,1) \\
W_i &\sim \text{Bernoulli}(e), \quad e = 0{,}5 \\
Y_i &= W_i Y_i(1) + (1 - W_i) Y_i(0) \\
\tau(X_i) &= \mu_1(X_i) - \mu_0(X_i) = \exp(-X_{i2}^2) + X_{i3}^2 - 2 \cdot \sin(X_{i4})
\end{aligned}
$$

5.2. Implementación

Para llevar a cabo las estimaciones, se usará la librería en Python EconML [2]. Esta se trata de un proyecto de código abierto hecho por el proyecto *Automated Learning and Intelligence for Causation and Economics* de Microsoft.

Una de las ventajas de esta librería es que su contribución radica en la implementación de los meta-algoritmos, siendo totalmente agnóstica en la elección de los estimadores de base siempre y cuando estos sean consistentes con la interfaz de *Scikit-Learn*, que se trata de un estandard en el análisis de datos en Python.

Desafortunadamente, la librería no tiene una implementación directa de los modelos auxiliares para la variable de resultado -necesario para calcular algunas de las métricas que se mencionaron en el apartado de selección de modelos- por lo que se debió modificar el código fuente de los meta-estimadores (bajo un paradigma llamado *monkey patching*). Adicionalmente se realizó un *pull request* al repositorio de código del proyecto de Microsoft para que estos cambios sean incorporados en versiones futuras de la librería.

[2] https://econml.azurewebsites.net

5.3. Resultados de la simulación Monte Carlo

Si bien para evaluar la calidad de la estimación podríamos teóricamente usar cualquier métrica propia de un problema de regresión -como el error medio absoluto o el coeficiente de determinación-, en el presente trabajo nos decantaremos exclusivamente por el error cuadrático medio (que, como vimos en secciones previas, es mejor conocido en la literatura causal como τ-risk), ya que es la métrica utilizada en toda la bibliografía de referencia utilizada.

A continuación, se presenta la distribución del τ-risk para las distintas combinaciones de estimadores y tamaños muestrales.

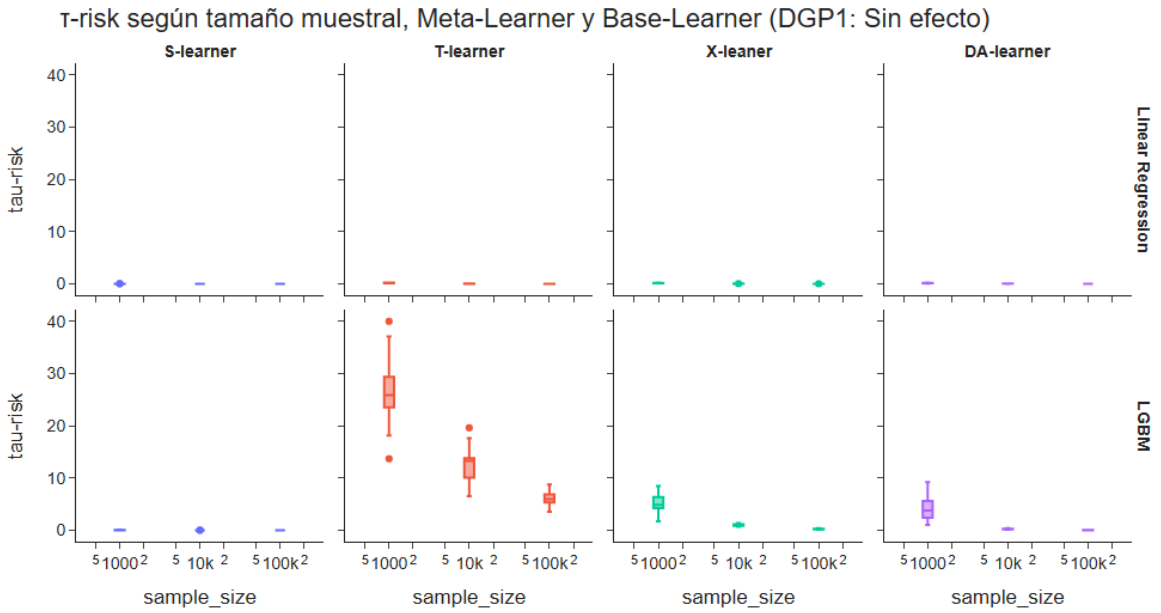

Figura 5.1: Distribución de τ-risk para DGP1

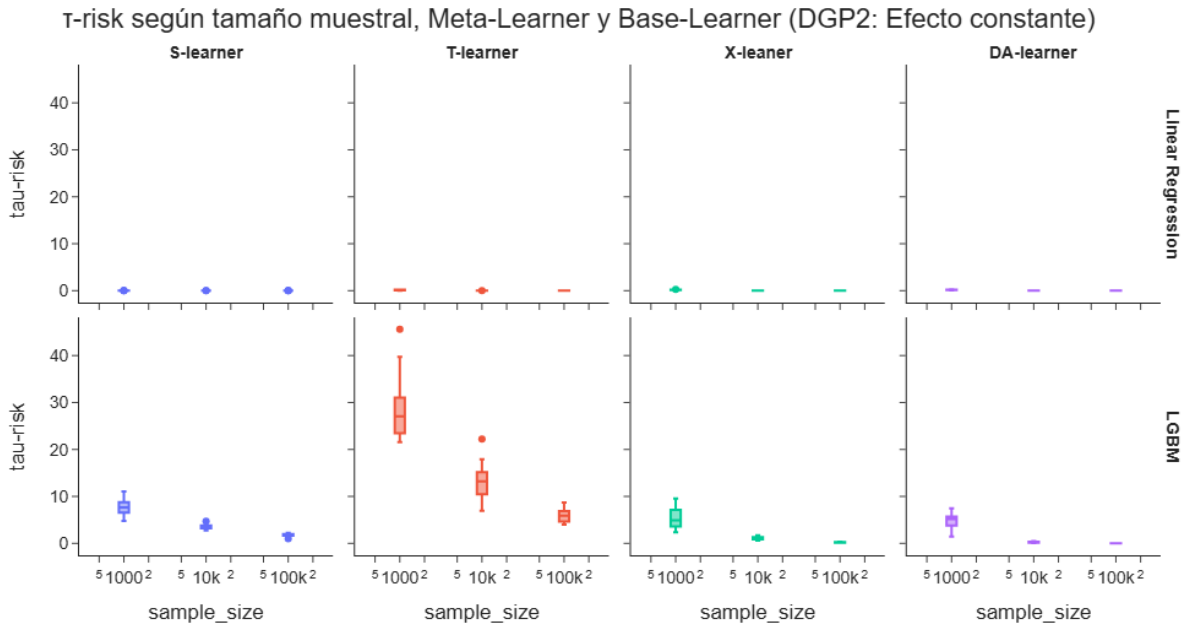

Figura 5.2: Distribución de τ-risk para DGP2

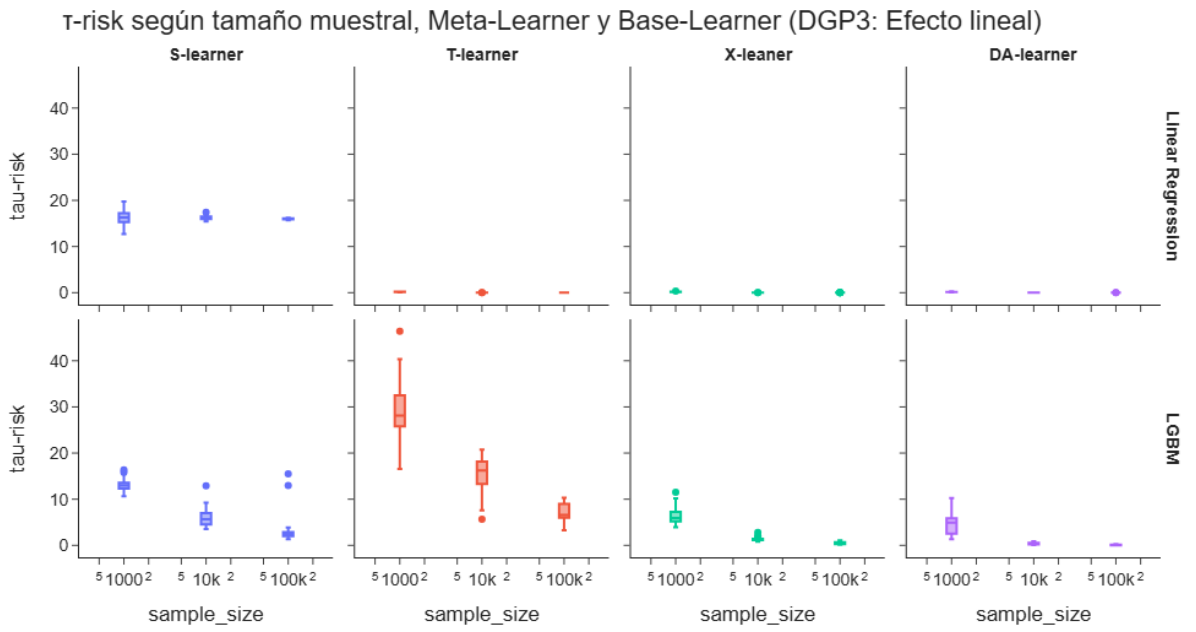

Figura 5.3: Distribución de τ-risk para DGP3

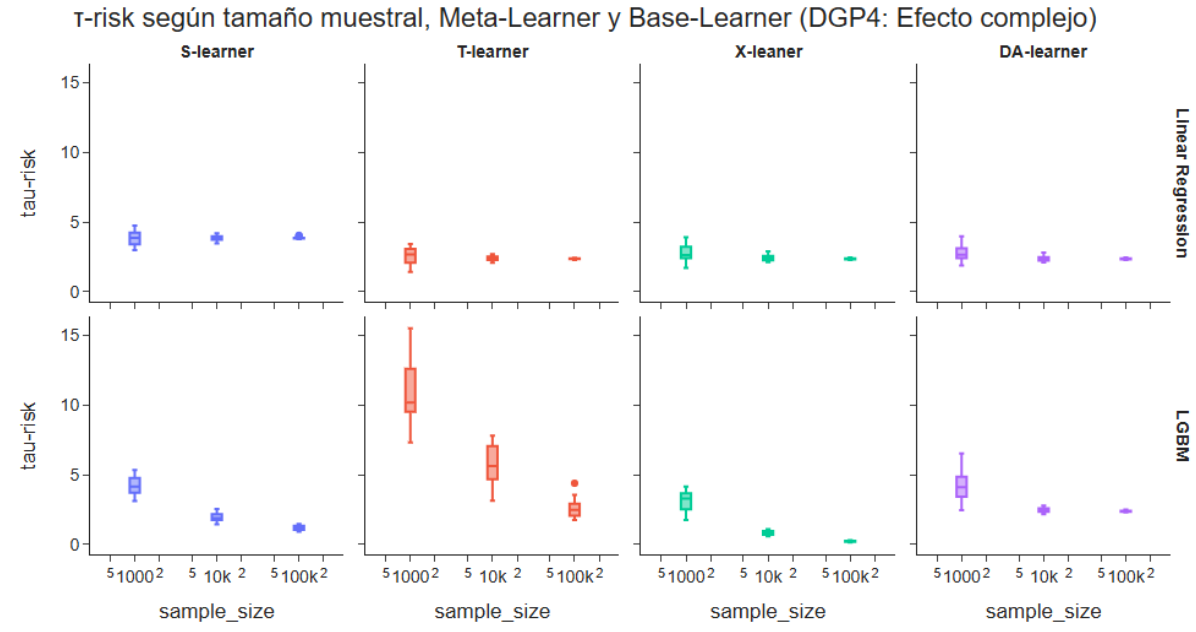

Figura 5.4: Distribución de τ-risk para DGP4

A modo de resumen, se presenta la media del estimador en cada caso.

DGP	Meta Learner	N / Base Learner	1000	10000	100000
DGP1: Sin efecto	DA-learner	LGBM	4.239584	0.212724	0.012321
		Linear Regression	0.142655	0.014545	0.001458
	S-learner	LGBM	0.029009	0.000058	0.000000
		Linear Regression	0.004085	0.000574	0.000067
	T-learner	LGBM	26.799783	12.404488	6.079128
		Linear Regression	0.148294	0.014502	0.001443
	X-leaner	LGBM	5.192335	1.008914	0.222663
		Linear Regression	0.150434	0.015501	0.001292
DGP2: Efecto constante	DA-learner	LGBM	4.627355	0.220148	0.012162
		Linear Regression	0.163394	0.014376	0.001381
	S-learner	LGBM	7.723422	3.542523	1.756170
		Linear Regression	0.010588	0.000592	0.000081
	T-learner	LGBM	28.355715	12.961184	5.944975
		Linear Regression	0.150649	0.013592	0.001372
	X-leaner	LGBM	5.433787	1.129834	0.209373
		Linear Regression	0.169281	0.012564	0.001433
DGP3: Efecto lineal	DA-learner	LGBM	4.649620	0.391156	0.079259
		Linear Regression	0.157505	0.012453	0.001365
	S-learner	LGBM	13.058453	6.098091	3.590333
		Linear Regression	16.357474	16.245137	15.984187
	T-learner	LGBM	29.085037	15.439281	7.174609
		Linear Regression	0.157202	0.013068	0.001432
	X-leaner	LGBM	6.549169	1.386162	0.491159
		Linear Regression	0.156598	0.013479	0.001513
DGP4: Efecto complejo	DA-learner	LGBM	4.289621	2.494824	2.394569
		Linear Regression	2.837907	2.393979	2.374275
	S-learner	LGBM	4.245686	1.975507	1.191782
		Linear Regression	3.841279	3.863077	3.869284
	T-learner	LGBM	10.819659	5.727072	2.589065
		Linear Regression	2.585081	2.400484	2.372153
	X-leaner	LGBM	3.131785	0.832452	0.227141
		Linear Regression	2.737120	2.419619	2.367542

5.3.1. Rendimiento de los meta-algoritmos según complejidad del efecto heterogéneo

Para los dos primeros diseños, donde poblacionalmente no existe ningún efecto causado por el tratamiento o este es realmente homogéneo, casi todos los meta-estimadores lograron buenos resultados (con la excepción del T-learner usando un algoritmo de boosting como *base learner*). Para el caso de $N = 1000$, el LGBM fue más propenso a encontrar algún tipo de efecto espúrio.

Cuando el efecto es heterogéneo, en caso de que este sea lineal en una covariable, la regresión lineal se muestra suficiéntemente expresiva como algoritmo de base, con la excepción del S-learner, donde ajustar un único modelo para el grupo de control y experimental parece ser perjudicial. Para este último caso, el LGBM funcionó mejor al tener la posibilidad de encontrar un patrón entre la variable indicadora del

tratamiento y la que causa la heterogeneidad en el efecto.

Cuando τ es una función no lineal, el S-learner, X-learner, DA-learner y el T-learner (para este último, solo con muestras grandes) funcionaron mucho mejor cuando se usó el LGBM como *base learner* ya que este puede representar mucho mejor la no linealidad del efecto.

Cuadro 5.1: Resumen del desempeño según la complejidad del efecto poblacional

Contexto	**Meta-estimadores que mejor funcionan**	**Comentarios**
Efecto nulo o homogéneo	Casi todos (excepto T-learner con boosting)	LGBM puede detectar efectos espurios con $N = 1000$
Efecto lineal en una covariable	Todos menos S-learner	S-learner se ve perjudicado al ajustar un único modelo para ambos grupos, aunque mejora con algoritmos flexibles como LGBM
Efecto no lineal	S-, X-, DA-, T-learner (este último solo con muestras grandes)	Los algoritmos más flexibles como LGBM capturan mejor la no linealidad del fenómeno.

5.3.2. Impacto del tamaño muestral en el desempeño

Los resultados obtenidos muestran que cuando se usa una regresión lineal en la primera etapa de las estimaciones, no se observa un incremento significativo en la precisión de las mismas a medida que más observaciones son tenidas en cuenta. Esta estabilidad en el desempeño puede atribuirse a que, incluso el tamaño muestral más pequeño considerado en el experimento ($N = 1000$), es suficientemente grande en relación con la cantidad de parámetros a estimar (del orden de 20), lo cual permite una estimación eficiente sin necesidad de mayor información.

Por el contrario, el uso de algoritmos de machine learning más complejos si muestra mejorías en términos de performance, tanto en términos de sesgo como de varianza en la métrica de interés. Estos modelos más flexibles requieren mayores cantidades de datos para evitar sobreajuste y capturar de manera adecuada relaciones no lineales o interacciones complejas entre las variables.

Estos resultados son también identificables en problemas de aprendizaje supervisado usuales.

Cuadro 5.2: Resumen del impacto del tamaño muestral en el desempeño de todos los meta-algoritmos

Base learner	**Desempeño**	**Comentarios**
Regresión lineal	Resultados estables para todo N.	$N = 1000$ es suficiente dada la dimensionalidad del espacio de variables considerado.
LGBM	Mejora al aumentar N	Los modelos flexibles se benefician de más datos para reducir sobreajuste y capturar relaciones complejas

5.3.3. Evaluación sin acceso al efecto causal real

En un escenario donde los resultados contrafactuales no son observados, se vuelve necesario recurrir a alguna de las métricas computables en la práctica presentadas anteriormente.

Más precisamente, lo que realmente nos interesa es evaluar la capacidad de dichas métricas para identificar correctamente al mejor modelo en términos de τ-risk.

Con ese objetivo, se analizó la relación entre el τ-risk y el resto de las métricas a través del estudio de las distribuciones bivariadas de las métricas de resultado en las distintas replicaciones del experimento. Este enfoque permite observar hasta qué punto las métricas alternativas se alinean con el τ-risk y, por lo tanto, cuán confiables resultan como criterios de selección de modelos.

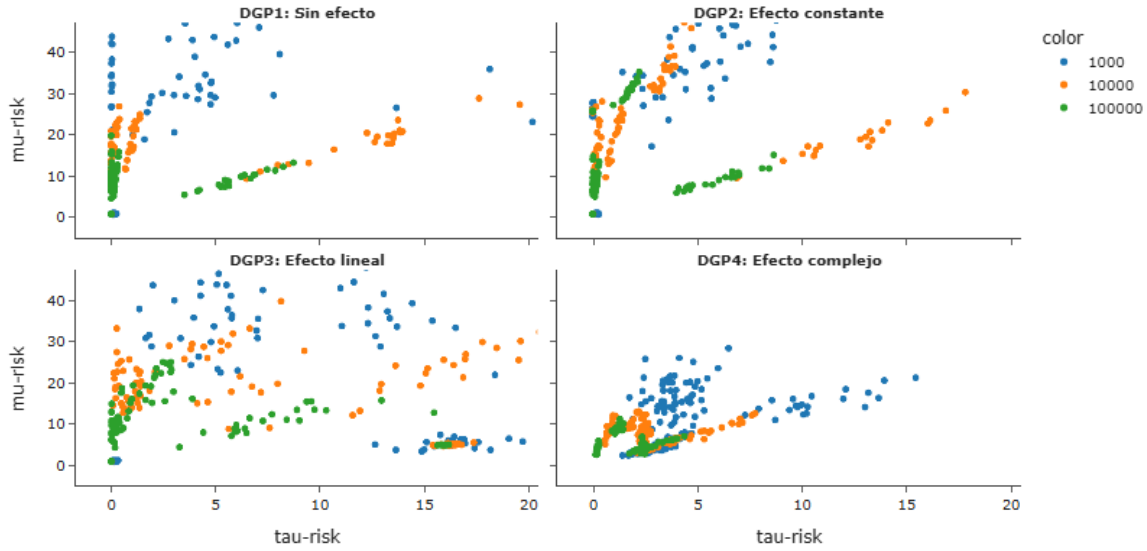

Figura 5.5: μ-risk

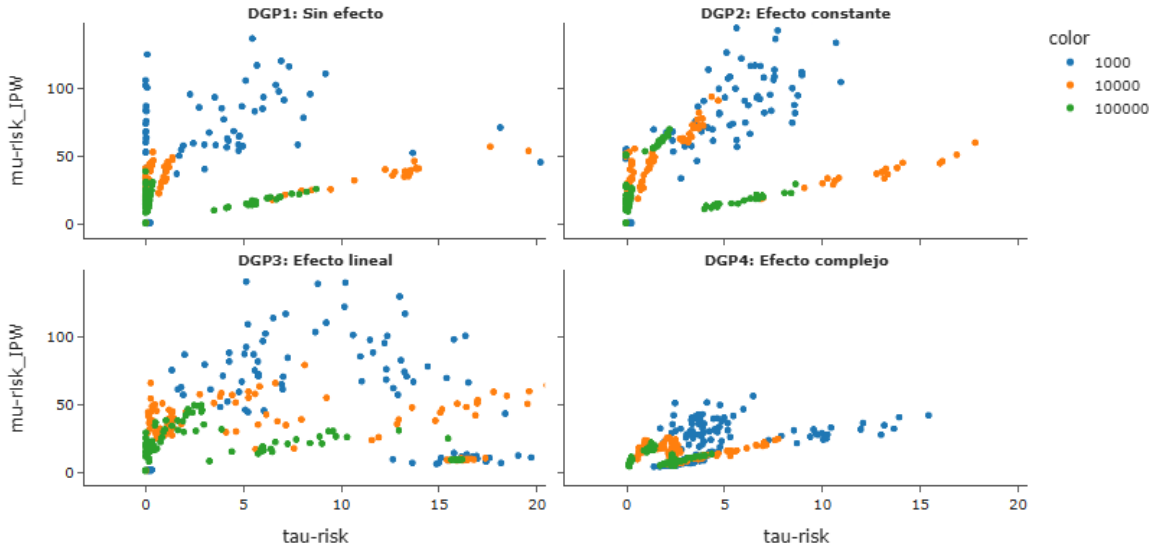

Figura 5.6: μ-risk$_{IPW}$

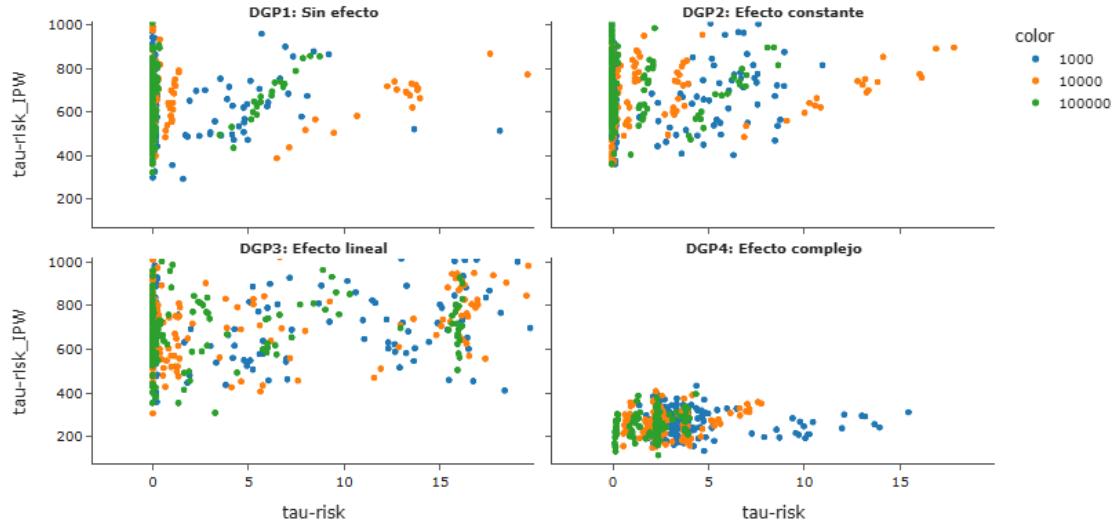

Figura 5.7: τ-risk$_{IPW}$

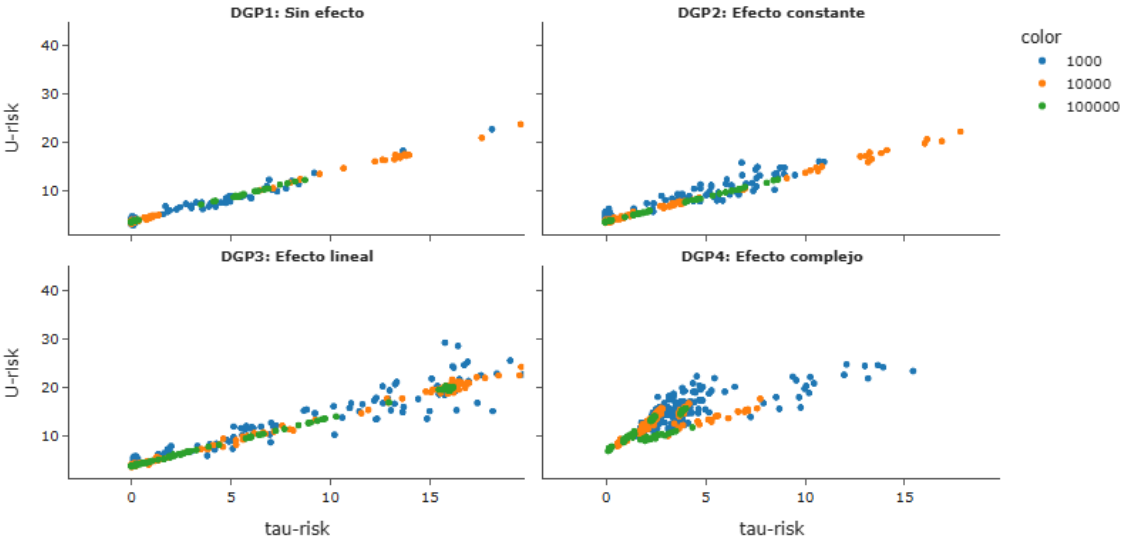

Figura 5.8: U-risk

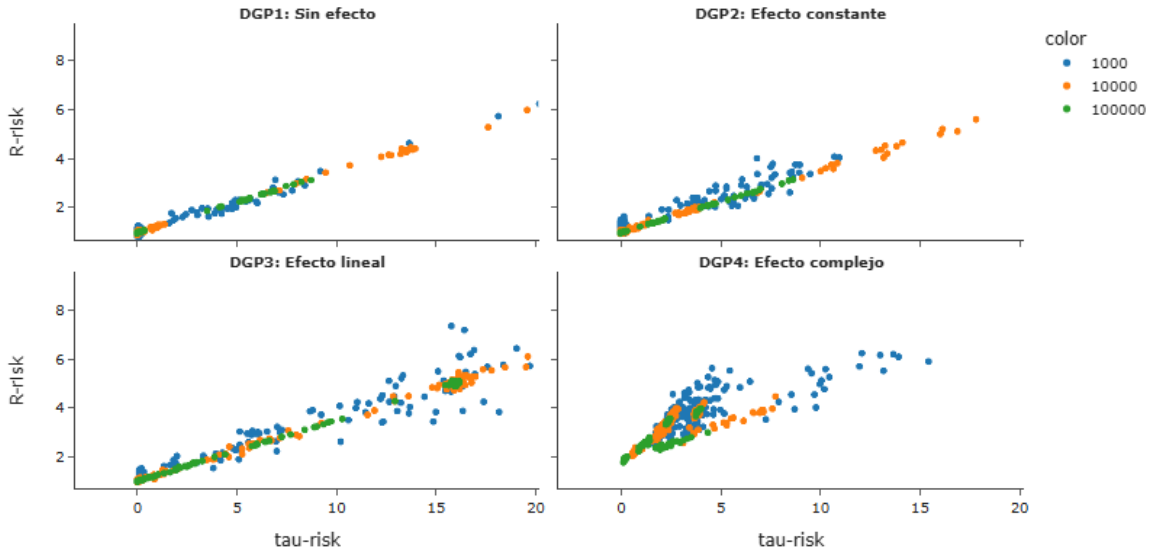

Figura 5.9: R-risk

Si bien resultaría deseable que exista una correlación entre las distintas métricas, a los fines de la selección de modelos nos es suficiente con que los valores más pequeños τ-risk se correspondan con los valores más pequeños de la métrica con la que se está comparando, fenómeno menos exigente que ni siquiera se observa en los tres primeros casos.

De hecho, estas primeras métricas están más relacionadas a tareas de predicción de la variable de resultado y esto muestra también por qué la *performance* en este tipo de problemas no es necesariamente indicadora de precisión en la estimación de efectos causales.

Por otro lado, U-risk y R-risk -ambas derivadas de la descomposición de Robinson mencionada en la segunda sección- muestran una gran correlación con τ-risk para todos los tamaños muestrales [3], siendo las elegidas para la selección de modelos de la siguiente sección. Más aún, como todos los DPGs cumplen con el supuesto de *unconfoundedness* por construcción y la asignación de tratamiento resultó aleatoria, ambas medidas son realmente múltiplos escalares.

[3] con un aumento de la variabilidad en muestras más chicas

Capítulo 6

Aplicación a un caso real: optimización de un producto financiero.

6.1. Fundamentos del *lending*

Con el término *lending* nos referimos al acto por el que una empresa ofrece un préstamo -sea éste en dinero o en especie- a alguna persona física o jurídica con la esperanza de recibir un determinado interés en un plazo pre-establecido según algún tipo de contrato legal.

En particular se menciona el caso de los préstamos personales donde el destinario es de forma excluyente un individuo que tendrá libre disponibilidad sobre los fondos desembolsados.

En los últimos años, buena parte de las empresas dedicadas a este negocio devinieron en un formato digital, donde la recepción de solicitudes, su eventual aprobación por modelos de riesgo y los desembolsos de dinero se realizan de forma totalmente automática sin intervención manual.

La calificación crediticia o *scoring* -entendida como la capacidad de medir el riesgo financiero en operaciones de crédito- es una de las primeras aplicaciones a gran escala del análisis de datos en la industria financiera (Thomas (2000)). Refiere a un conjunto de técnicas estadísticas que ayudan a distintas organizaciones a decidir la conveniencia de otorgar o no una determinada cantidad de dinero a potenciales prospectos, así como -en una gestion más moderna de los productos financieros- sus condiciones de otorgamiento, es decir, la tasa de interés, el plazo en el que debe ser devuelto o el propio monto otorgado.

En Thomas (2000) se menciona que un adulto en el Reino Unido o Estados Unidos es calificado -en general, sin saberlo- crediticiamente en promedio una vez por semana.

Un fenómeno central en el riesgo crediticio es el de la selección adversa que en Phillips y Rafard (2011) se define como una sensibilidad al precio (que en este caso, corresponde a las condiciones de otorgamiento del producto) diferencial entre distintos segmentos, en particular, entre los buenos y malos pagadores, donde los malos pagadores son mucho más propensos a aceptar un préstamo a una tasa de interés elevada.

Éste fenómeno no es exclusivo de las economías modernas. Ya en 1776, Adam Smith mencionaba:

> *If the legal rate of interest in Great Britain, for example, was fixed so high as eight or ten per cent, the greater part of the money which was to be lent would be lent to prodigals and projectors, who alone would be willing to give this high interest. Sober people, who will give for the use of money no more than a part of what they are likely to make by the use of it,*

would not venture into the competition.

Es por esto que, en búsqueda de maximizar la rentabilidad del negocio, un *lender* busque personalizar lo máximo posible su oferta de crédito. Esta personalización consiste en ajustar las condiciones del préstamo —principalmente la tasa de interés— al perfil específico de cada prospecto. Así, se procura evitar ofrecer una tasa inferior a la máxima que el cliente estaría dispuesto a aceptar, lo cual implicaría una pérdida de rentabilidad en el producto financiero. A su vez, también se busca evitar ofrecer una tasa superior a dicha máxima, ya que esto podría derivar en la no aceptación de la oferta por parte del cliente y, por ende, en la pérdida de la oportunidad de colocación del crédito.

Esta reactividad a la tasa de interés puede pensarse como un efecto heterogéneo ante una variación de la misma y en consiguiente puede identificarse con la metodología desarrollada en las secciones previas.

6.1.1. Datos

Los datos utilizados provienen de un *lender* argentino dedicado al otorgamiento de préstamos personales a un segmento post-bancarizado, entendiendo a ésta última tipología como un grupo que en general tiene un historial crediticio en instituciones bancarias, pero este resulta negativo. Es decir, son personas que en algún momento tuvieron acceso a algún producto crediticio -sea tarjeta de crédito o algún otro tipo de préstamo personal o prendario- pero no lograron cumplir sus compromisos incurriendo en un default.

Los mismos fueron recolectados en el marco de un ensayo controlado aleatorizado (mejor conocido como *randomized controlled trial* o RCT) donde se realizaron ofertas de crédito de forma aleatoria a una tasa considerablemente menor a la política de crédito vigente en la empresa en ese momento. Este diseño nos asegura que se cumpla el supuesto de ignorabilidad y de superposición. No obstante, el cumplimiento del *Stable Unit Treatment Value Assumption* (SUTVA) podría no estar plenamente asegurado, dado que la asignación aleatoria fue realizada a nivel individual. Es posible que existan vínculos entre solicitantes (por ejemplo, compañeros de trabajo o familiares) que hayan recibido ofertas a tasas distintas, lo que podría haber generado efectos de interferencia donde la decisión de aceptación del crédito por parte de uno de ellos haya sido influenciada por la oferta que recibió el otro. De todos modos, no se espera que este fenómeno sea suficientemente frecuente o sistemático como para comprometer la validez de los resultados.

Variable	Descripción	Tipo de dato
demographic_age	Edad del individuo	continua
origin	Canal de márketing con el que se originó la solicitud	categórica
cendeu_amount_sum_24M_ripte	Suma total de deuda en 24 meses según CENDEU y ajustada por RIPTE	continua
cendeu_amount_sum_6M_ripte	Suma total de deuda en 6 meses según CENDEU y ajustada por RIPTE	continua
cendeu_monthswithdebt_countdistinct_24M	Cantidad de meses con deuda en los últimos 24 meses	continua
monthly_commitments_ripte	Compromisos mensuales estimados ajustados por RIPTE	continua
AP_12m_Empleado_Pagos_Cant	Cantidad de aportes como empleado en los últimos 12 meses	continua
NSE_percentil	Percentil del Nivel Socioeconómico (NSE)	continua
AP_6m_Empleado_AportesObraSocial_Pagos_Cant	Cantidad de pagos de aportes a obra social como empleado en los últimos 6 meses	continua
querys_banca_12M	Cantidad de consultas al bureau de crédito por parte de un banco en los últimos 12 meses	continua
income_predictor	Ingreso presunto	continua
afip_employer_business_name_agrup_cat	Categoría agrupada del nombre del empleador según AFIP	categórica
afip_days_as_employee_current	Días en el empleo actual según registros AFIP	continua
cendeu_sit_1_qty_6m	Cantidad de veces en situación 1 (normal) en los últimos 6 meses según CENDEU	continua
cendeu_entity_maxdebt_type_cat_6M	Categoría del tipo de entidad con mayor deuda en 6 meses	categórica
demographic_gender_cat	Género del individuo	categórica
cendeu_creditHistory_M	Meses de historial crediticio según CENDEU	continua
querys_3M	Cantidad de consultas en los últimos 3 meses	continua
risk_score	Score de riesgo	continua
is_converted	Flag indicador de conversión	categórica
group	Grupo de tratamiento	categórica

Cuadro 6.1: Diccionario de variables utilizadas en el modelo

Estos datos provienen tanto de un *bureau* de crédito como de fuentes internas de la compañia o de bases de datos públicas, como la central de deudores del Banco Central de la República Argentina (CENDEU).

Al ser Argentina un país que experimentó ciclos de alta inflación en los últimos años, las variables monetarias fueron deflactadas convenientemente.

Las variables categóricas fueron convertidas en variables dummies siguiendo una nomenclatura donde la categória específica aparece como sufijo en el nombre de la variable correspondiente. Para evitar la multicolinealidad, se descartó una categoría de cada variable.

En total, se realizaron 119177 ofertas a una tasa alta de las cuales 4689 fueron aceptadas en contraposición a 11880 ofertas en el grupo experimental que originaron 837 ventas. Esto representarían tasas

de conversión de 3.9 % para el grupo de control y de 7 % para el experimental.

Luego de dividir al conjunto de datos en un subconjunto de entrenamiento (70 %) y de testeo, se procedió a ajustar los mismos modelos que en las secciones previas.

6.2. Resultados

6.2.1. Estimación de efectos heterogéneos sobre conversión

Luego de ajustar los modelos correspondientes, se estimó la medida de riesgo de los estimadores a partir de 100 remuestras bootstrap de los conjuntos de entrenamiento y prueba. Este procedimiento permitió evaluar la variabilidad de dicha medida y obtener una estimación más robusta de su estabilidad.

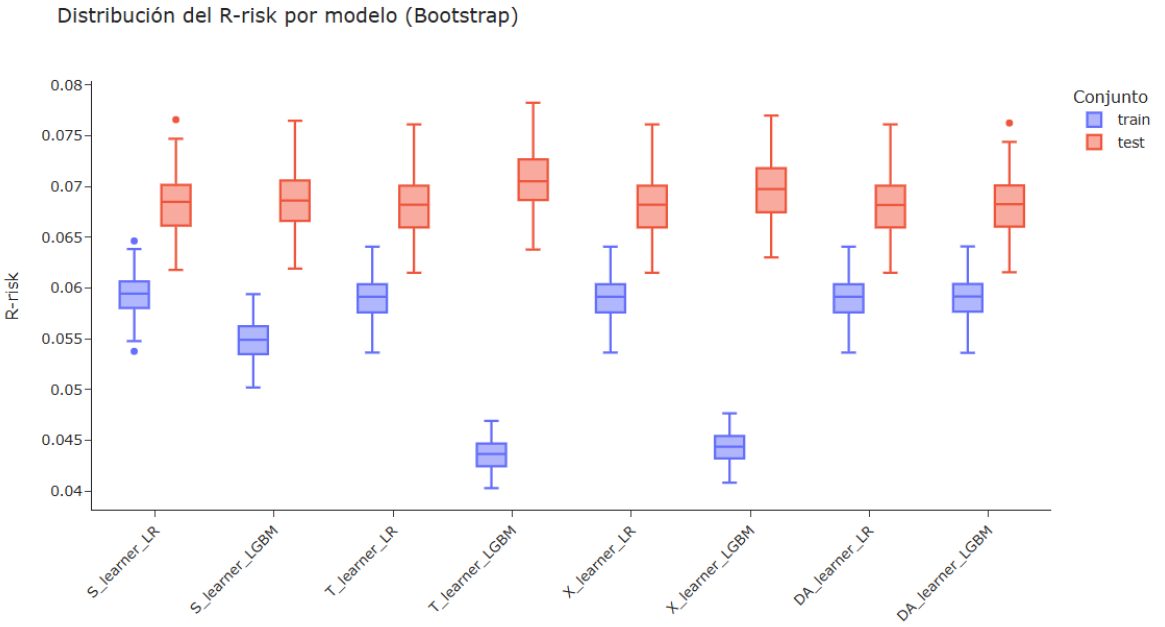

Figura 6.1: R-risk por conjunto

Se observa que si bien en general los meta-algoritmos que usan el método de boosting como *base learner* muestran un mejor desempeño en los datos de entrenamiento,los patrones identificados no parecen generalizarse adecuadamente al conjunto de prueba. Este es un fenómeno totalmente análogo al overfitting en tareas de predicción.

El algoritmo elegido por tener el mejor desempeño en *test* es el DA-learner con la regresión lineal. Como se observó en el gráfico anterior, esta diferencia no es muy significativa respecto a los otros candidatos.

Si bien no conocemos la forma funcional del efecto heterogéneo en cuestión, esta paridad en la performance de todos los meta-algoritmos, con el posible sobreajuste en algunos casos cuando se utiliza un método de *boosting* como *base learner*, el ejercicio con datos simulados nos da la pauta de que es probable que poblacionalmente este efecto no sea muy complejo.

Luego, se volvió a ajustar el modelo con todos los datos disponibles.

6.2.2. Identificación de perfiles de clientes sensibles a la tasa de interés

Con el fin de explicitar los patrones identificados por el modelo, se dividió la estimación del CATE en quintiles.

A continuación, se presentan las tasas de conversión para los grupos de control y tratamiento, así como la diferencia absoluta entre ellas en cada uno de los quintiles.

Quintil (CATE)	Control	Tratamiento	Diferencia Absoluta
(0.001, 0.018]	1.50 %	2.98 %	1.48 %
(0.018, 0.022]	1.62 %	3.81 %	2.19 %
(0.022, 0.030]	1.99 %	4.34 %	2.35 %
(0.030, 0.047]	4.04 %	8.03 %	3.99 %
(0.047, 0.573]	10.68 %	16.75 %	6.07 %

Cuadro 6.2: Tasa de conversión por quintil de CATE estimado.

Se observa que los grupos de mayor *score* efectivamente muestran un diferencial de conversión más elevado, con el detalle de que estos ya parten de una tasa de conversión -en ausencia de tratamiento- mayor.

A continuación, se muestra la media de las variables utilizadas para los distintos quintiles:

bins	(0.001, 0.018]	(0.018, 0.022]	(0.022, 0.03]	(0.03, 0.047]	(0.047, 0.573]
origin_third_party	0.000	0.000	0.000	0.212	0.435
cendeu_entity_maxdebt_type_cat_6M_Niguna	0.179	0.013	0.007	0.053	0.024
cendeu_entity_maxdebt_type_cat_6M_Compañía financiera	0.044	0.011	0.013	0.023	0.016
cendeu_entity_maxdebt_type_cat_6M_Otro	0.121	0.201	0.123	0.153	0.144
querys_3M	1.351	1.847	2.703	3.346	3.903
afip_employer_business_name_agrup_cat_SRL	0.059	0.053	0.057	0.072	0.080
cendeu_entity_maxdebt_type_cat_6M_Fintech	0.107	0.176	0.126	0.127	0.119
cendeu_entity_maxdebt_type_cat_6M_Banco privado - principales	0.020	0.296	0.462	0.283	0.404
AP_6m_Empleado_AportesObraSocial_Pagos_Cant	0.006	0.024	0.117	1.039	2.132
monthly_commitments_ripte	0.063	0.109	0.154	0.167	0.194
cendeu_amount_sum_6M_ripte	5.072	7.387	10.200	11.025	12.708
cendeu_amount_sum_24M_ripte	32.011	42.300	52.751	53.490	56.714
origin_outbound	1.000	1.000	1.000	0.788	0.389
income_predictor	0.444	0.511	0.561	0.591	0.628

Los prospectos del grupo de mayor CATE resultan ser aquellos de mayores ingresos, con un mayor endeudamiento [1] en instituciones bancarias privadas. Además, existen diferencias en el modo en el que dichas solicitudes fueron creadas: Si en el primer quintil predominan solicitudes generadas a partir de una base de datos interna, en el último predominan las creadas por fuentes de tráfico externas.

Estos hallazgos son consistentes investigaciones previas. En Céspedes (2017) se considera un escenario cuasi-experimental donde un lender define la tasa de interés según quantiles de su *score* de riesgo. Así, en el límite de los segmentos se da la situación en la que pequeñas variaciones en el puntaje producen saltos en la tasa de interés ofertada [2]. De este modo, los autores logran identificar que factores como el ingreso del prospecto, scores genéricos del mercado (FICO) y el historial crediticio son fuentes de heterogeneidad en la sensibilidad a la tasa de interés.

[1] Esto -en este segmento de clientes- es una característica positiva ya que indica que se trata de individuos a los que otros *lenders* aceptaron otorgarles crédito con anterioridad.

[2] Este tipo de diseño se conoce como Regression Discontinuity Design o RDD

Capítulo 7

Conclusiones

El leitmotiv del presente trabajo consistió en mostrar que la buena capacidad predictiva de un modelo para representar una variable de resultado no es prognóstica de una buena capacidad para identificar efectos causales cuando se aplica un tratamiento que potencialmente afecte a la misma.

Si bien esto en términos teóricos no es más que una derivación de la famosa frase de que correlación no implica causalidad, de forma práctica se mostró que existen metodologías que efectivamente permiten operar con algoritmos que buscan correlaciones para poder identificar realmente una relación causal.

Además, se presentó un paquete informático -EconML- que permite no solo realizar un análisis exploratorio de un conjunto de datos sino también la posibilidad de implementar dichos modelos de forma rápida en un ámbito productivo, para eventualmente ser usados en la toma de decisiones en tiempo real. Así, un prestamista podría calcular la sensibilidad a la tasa de interés de los prospectos y usar este nuevo número de forma conjunta con el score de riesgo para la asignación de políticas de crédito.

7.1. Interpretación general de resultados

Cuando se discutió sobre la selección de modelos causales, se planteó una metodología que permite elegir entre distintas familias de algoritmos de forma automática, sin necesidad de un conocimiento a priori de la complejidad (o existencia) del efecto causal ante un tratamiento, en el ejercicio con datos simulados obtuvimos una serie de pautas acerca de qué modelos son más razonables en cada contexto.

En caso de que el efecto poblacional sea más bien sencillo u homogéneo, casi todos los meta-algoritmos logran identificar muy bien dicho comportamiento. De forma contraria, para poder encontrar patrones más complejos, resulta necesario contar con una mayor cantidad de observaciones que permitan el uso de algoritmos más expresivos.

7.2. Posibles extensiones

En las secciones anteriores se tomaron una serie de decisiones tendientes a acotar el alcance del mismo. Entre ellas, se pueden mencionar el haber considerado únicamente tratamientos binarios y el haber asumido *unconfoundedness*.

El tipo de tratamiento puede ser generalizado tanto considerando múltiples tratamientos que se aplican de forma conjunta o como si estos fuesen variables continuas en lugar de asignaciones binarias. Esto trae nuevas dificultades metodológicas ya que es posible que aparezcan efectos heterogéneos emergentes cuando los tratamientos interaccionen entre sí.

En caso de que se sospeche la existencia de un *confounder* no observado en el conjunto de covariables

pero se cuente con alguna variable instrumental, en Hartford et al. 2017 se presenta un algoritmo basado en redes neuronales que permite estimar el CATE con una cierta pérdida en la eficiencia. En Newey y Powell 2003 aparece un procedimiento en dos etapas que también permite aprovechar un vector de instrumentos. Ambos procedimientos se encuentran implementados en EconML.

Bibliografía

Athey, S., & Imbens, G. W. (2019). Machine Learning Methods That Economists Should Know About. *Annual Review of Economics*, *11*(1), 685-725. https://doi.org/10.1146/annurev-economics-080217-053433

Athey, S., & Wager, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests [Version July 11, 2017]. *arXiv preprint arXiv:1510.04342*.

Caron, A., Baio, G., & Manolopoulou, I. (2021). *Estimating Individual Treatment Effects using Non-Parametric Regression Models: a Review* [Department of Statistical Science, University College London]. https://arxiv.org/abs/2009.06472

Céspedes, J. (2017). *Heterogeneous Sensitivities to Interest Rate Changes: Evidence from Consumer Loans* [Job Market Paper].

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters [Version v7, 3 Nov 2024]. *arXiv preprint arXiv:1608.00060*.

Doutreligne, M. (2023). How to Select Predictive Models for Causal Inference? *Preprint*.

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data [Published in final edited form as: Stat Med. 2011 October 30; 30(24). Available in PMC 2014 January 05]. *Stat Med*, *30*(24). https://doi.org/10.1002/sim.4322

Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction. *Proceedings of the 34th International Conference on Machine Learning*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2.ª ed.). Springer.

Kehl, V., & Ulm, K. (2006). Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis*, *50*(5), 1338-1355. https://doi.org/10.1016/j.csda.2004.11.015

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning. *arXiv preprint arXiv:1706.03461*. https://arxiv.org/abs/1706.03461

Newey, W. K., & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, *71*(5), 1565-1578.

Phillips, R., & Rafard, R. (2011). Price-Driven Adverse Selection in Consumer Lending. http://www.cprm.columbia.edu

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, *56*(4), 931-954. https://doi.org/10.2307/1912705

Rosenbaum, P. R., & Rubin, D. B. [Donald B]. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rubin, D. B. [Donald B.]. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Stefan, A. M., & Schönbrodt, F. D. (2023). Big little lies: a compendium and simulation of p-hacking strategies. *Royal Society Open Science*, *10*(2). https://doi.org/10.1098/rsos.220346

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149-172.