

## ECONOMETRÍA APLICADA AVANZADA

### CONTROL DE LECTURA: Causal Forests Method

#### Control de Lectura: *Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs*

Responde las siguientes preguntas que se desprenden del texto de Jonathan M.V. Davis and Sara B. Heller (2017) Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs

#### 1) ¿Cuál es el objetivo principal del paper? ¿Qué problema intenta resolver?

El objetivo principal del paper es explorar cómo los **Causal Forests** pueden ser utilizados para identificar y analizar la heterogeneidad en los efectos del tratamiento en experimentos controlados aleatorizados (RCTs). En particular, el estudio busca demostrar la capacidad de esta técnica para predecir variaciones en el impacto de un programa social, el "**One Summer Plus**" (OSP), sobre los arrestos por crímenes violentos y el empleo futuro.

El problema que intenta resolver es cómo estimar los efectos de tratamiento promedio condicionales (CATEs) para identificar qué subgrupos dentro de una población responden de manera más efectiva a un tratamiento o intervención específica. Esto es especialmente útil para optimizar el diseño y la focalización de políticas públicas.

#### 2) ¿Qué métodos utiliza el paper para abordar este problema?

El paper utiliza los siguientes métodos para abordar el problema:

- **Causal Forests:**
  - Una extensión de los algoritmos de árboles de regresión y bosques aleatorios diseñada para estimar efectos de tratamiento promedio condicionales (CATEs).
  - Este enfoque permite modelar la heterogeneidad en los efectos del tratamiento al segmentar la población en subgrupos basados en características observables.
- **Validación Cruzada:**
  - Se dividen los datos en conjuntos de entrenamiento y prueba para garantizar que las predicciones de los CATEs sean robustas y no se ajusten excesivamente a los datos.
- **Análisis de Experimentos Controlados Aleatorizados (RCTs):**
  - El estudio se basa en datos de dos RCTs del programa "One Summer Plus" (OSP) para garantizar un diseño experimental riguroso y minimizar el sesgo.
- **Pruebas de Robustez:**

- Se realizan comparaciones adicionales entre los resultados estimados y los datos observados para validar la precisión del método.
- **Visualización de Resultados:**
  - Se presentan gráficos y análisis descriptivos para ilustrar la heterogeneidad de los efectos identificados por los Causal Forests.

Estos métodos en conjunto permiten al paper abordar el problema de identificar subgrupos de población con respuestas diferenciadas a una intervención, facilitando la optimización de programas sociales.

### **3) ¿Cómo se aplica el método de causal forest en el paper?**

En el paper, el método de Causal Forest se implementa para analizar la heterogeneidad en los efectos de tratamiento del programa One Summer Plus (OSP), un programa social diseñado para reducir arrestos por crímenes violentos y mejorar el empleo futuro en jóvenes vulnerables de Chicago. La aplicación del método se realiza en varias etapas cuidadosamente estructuradas para garantizar la validez de los resultados.

Primero, los autores recopilaron datos de dos experimentos controlados aleatorizados (RCTs) realizados en 2012 y 2013. Estos datos incluyen variables de tratamiento (participación en OSP), resultados de interés (arrestos por crímenes violentos y empleo futuro) y covariables que describen características individuales como demografía y antecedentes.

Luego, el modelo de Causal Forest fue entrenado utilizando estos datos. Este método combina árboles de regresión y bosques aleatorios, pero con un enfoque en la estimación de efectos de tratamiento promedio condicionales (CATEs). Específicamente, los árboles dividen la población en subconjuntos homogéneos en cuanto al impacto del tratamiento, lo que permite identificar subgrupos con respuestas diferenciadas.

Para evitar sobreajuste y garantizar la robustez de los resultados, los datos se dividieron en conjuntos de entrenamiento y prueba. El modelo se entrenó con un subconjunto de datos y sus predicciones se validaron utilizando un conjunto de prueba separado, una práctica estándar en el aprendizaje automático para evaluar el desempeño predictivo.

Finalmente, los resultados de los Causal Forests fueron analizados e interpretados para identificar patrones significativos de heterogeneidad. Estos hallazgos no solo destacaron subgrupos específicos con respuestas positivas al programa, sino que también ofrecieron insights prácticos sobre cómo focalizar recursos en futuras iteraciones del programa OSP.

### **4) ¿Qué datos se utilizaron en el estudio? ¿Cómo se recopilaron y procesaron estos datos?**

En el estudio se utilizaron datos provenientes de dos experimentos controlados aleatorizados (RCTs) realizados en 2012 y 2013 sobre el programa **"One Summer Plus" (OSP)**. Los datos se recopilaron y procesaron de la siguiente manera:

- **Diseño de los Experimentos:**

- Los participantes fueron seleccionados mediante un proceso aleatorio, asegurando que cualquier diferencia en los resultados pudiera atribuirse al programa y no a características preexistentes.
- Se incluyeron 1,634 participantes en el experimento de 2012 y 5,216 en el de 2013.
- **Variables Recopiladas:**
  - **Variables de Tratamiento:** Participación en el programa OSP.
  - **Resultados:** Arrestos por crímenes violentos y datos sobre empleo futuro.
  - **Covariables:** Características demográficas, antecedentes escolares, y registros de arrestos previos.
- **Fuentes de Datos:**
  - Los datos fueron recopilados de registros administrativos (como antecedentes penales y laborales) y formularios de inscripción al programa.
- **Procesamiento:**
  - Los datos fueron limpiados y organizados para su análisis, eliminando duplicados y registros incompletos.
  - Las covariables se estandarizaron para facilitar el análisis comparativo y reducir el sesgo.
  - Se dividieron los datos en conjuntos de entrenamiento y prueba para implementar el modelo de Causal Forests y validar los resultados.

Este proceso asegura la calidad y validez de los datos utilizados en el análisis. Así se permite que las estimaciones de los efectos del tratamiento sean confiables y precisas.

## 5) ¿Cuáles fueron los hallazgos clave del paper?

Uno de los hallazgos más destacados del estudio es la identificación de una heterogeneidad significativa en los efectos del tratamiento. El programa OSP no tuvo el mismo impacto en todos los participantes; en cambio, mostró ser particularmente beneficioso para ciertos subgrupos dentro de la población objetivo. Por ejemplo, los jóvenes con antecedentes más altos de vulnerabilidad, como un historial de arrestos previos, experimentaron reducciones más notables en los arrestos por crímenes violentos. Este hallazgo destaca la importancia de reconocer y analizar las diferencias individuales dentro de una población, en lugar de asumir que un programa tiene un impacto uniforme.

Además, los resultados confirman la utilidad del programa OSP para reducir arrestos por crímenes violentos, uno de sus objetivos principales. Sin embargo, este efecto fue más pronunciado en subgrupos específicos, lo que sugiere que los recursos del programa podrían ser dirigidos de manera más eficiente hacia aquellos que tienen una mayor probabilidad de beneficiarse. Este hallazgo tiene implicaciones importantes para la política pública, ya que permite optimizar el diseño de programas y maximizar su impacto social.

El estudio también valida la capacidad predictiva de los **Causal Forests**, demostrando que estos modelos pueden estimar con precisión los efectos de tratamiento promedio condicionales (CATEs) y proporcionar información útil para decisiones prácticas. Esto refuerza la importancia de los métodos de aprendizaje automático no solo como herramientas académicas, sino también como instrumentos prácticos para mejorar la efectividad de las intervenciones sociales.

## 6) ¿Cómo interpretas los resultados presentados en el paper?

Un aspecto central de la interpretación es la oportunidad de lograr una **focalización más efectiva** de los recursos. Los hallazgos muestran que los programas sociales, como el programa **One Summer Plus (OSP)**, pueden diseñarse de manera que beneficien más a aquellos grupos que tienen una mayor probabilidad de obtener resultados positivos. Este enfoque optimiza el uso de recursos, asegurando que los esfuerzos se concentren en quienes tienen más que ganar, como jóvenes con antecedentes de alta vulnerabilidad.

Además, la identificación de subgrupos con respuestas diferenciadas permite la **adaptación de políticas** a nivel individual. En lugar de adoptar un enfoque universal para las intervenciones, es posible personalizar programas basándose en características específicas de los participantes, maximizando los beneficios para cada grupo. Este nivel de precisión en el diseño de políticas puede ser crucial para abordar problemas complejos en contextos sociales diversos.

Desde un punto de vista metodológico, los resultados también destacan la versatilidad de los **Causal Forests**. Este enfoque no solo valida su eficacia en el contexto del programa OSP, sino que también sugiere que puede ser aplicado en otros sectores, como la educación, la salud pública o la justicia penal. La capacidad de los Causal Forests para estimar efectos de tratamiento heterogéneos en diversos dominios amplía las posibilidades de su uso en múltiples contextos.

Por último, los resultados subrayan la **importancia de reconocer la heterogeneidad** en los efectos de las intervenciones. No todos los participantes se benefician de la misma manera, y este tipo de análisis permite abordar estas diferencias de manera explícita. Al hacerlo, se pueden diseñar programas que no solo maximicen los resultados generales, sino que también reduzcan desigualdades al garantizar que los beneficios lleguen a aquellos que más los necesitan.

## 7) ¿Qué limitaciones tiene el estudio? ¿Cómo podrían superarse estas limitaciones en futuras investigaciones?

Se destacan varias limitaciones que podrían abordarse en futuras investigaciones para mejorar la metodología y ampliar su aplicabilidad. Una de las principales limitaciones es su dependencia de las covariables observables. Si factores no medidos influyen en los resultados, los modelos podrían estar sesgados. Para abordar esto, se recomienda recolectar datos más completos e incluir técnicas avanzadas, como instrumentos o variables latentes, que permitan considerar variables no observadas.

Otra limitación significativa es la generalización restringida de los hallazgos. Al estar basado en un programa específico (OSP) en un contexto urbano en Chicago, su aplicabilidad a otras poblaciones o regiones puede ser limitada. Replicar el análisis en diferentes contextos geográficos y culturales sería esencial para validar la robustez y universalidad de los resultados.

La **complejidad computacional** es también un desafío, ya que los **Causal Forests** requieren grandes cantidades de recursos computacionales. Este problema podría mitigarse mediante optimizaciones

algorítmicas o aprovechando tecnologías de paralelización para mejorar la eficiencia en el procesamiento de grandes conjuntos de datos.

Adicionalmente, aunque los **Causal Forests** identifican patrones de heterogeneidad, la interpretación y traducción de estos hallazgos en recomendaciones prácticas puede ser compleja. Para superar esto, se podrían desarrollar herramientas más intuitivas de visualización que faciliten la comunicación de los resultados a audiencias no técnicas.

Finalmente, el modelo se valida principalmente con el conjunto de datos disponible, lo que podría limitar su capacidad predictiva en otros contextos. Probar el modelo en escenarios futuros o conjuntos de datos independientes sería clave para garantizar su generalización.

#### **8) ¿Cómo contribuye este paper a la literatura existente sobre causal forest?**

Este paper representa una contribución significativa a la literatura existente sobre **Causal Forests**, destacándose por su enfoque práctico y metodológico. Una de sus principales aportaciones es su aplicación empírica en el análisis de políticas públicas. Demuestra cómo los **Causal Forests** pueden utilizarse para identificar la heterogeneidad en los efectos de un programa social, en este caso, el programa **OSP**, proporcionando un ejemplo relevante y práctico en un contexto real.

Otra contribución clave es el uso de un diseño experimental riguroso, basado en experimentos controlados aleatorizados (RCTs). Esto refuerza la credibilidad de los resultados, ya que muchos estudios previos sobre **Causal Forests** se habían centrado principalmente en simulaciones o datos no experimentales.

Además, el paper destaca el enfoque en la heterogeneidad de los efectos del tratamiento. Subraya cómo los **Causal Forests** pueden identificar patrones de variación en los efectos según las características individuales, un aspecto que no siempre se aborda con profundidad en la literatura. Esto enriquece la comprensión de cómo los tratamientos pueden impactar de manera diferencial a distintos subgrupos de población.

Desde un punto de vista metodológico, el estudio proporciona una guía detallada para la implementación de los **Causal Forests**, desde la preparación de los datos hasta la validación de los resultados. Esta guía no solo facilita su adopción por parte de otros investigadores, sino que también amplía las posibilidades de aplicación de este enfoque en diversas áreas.

Por último, el trabajo tiene una relevancia práctica significativa, ya que demuestra cómo los resultados obtenidos con **Causal Forests** pueden informar decisiones reales, como la focalización de programas sociales. Al cerrar la brecha entre la investigación académica y su aplicación en el mundo real, este estudio establece un modelo valioso para futuras investigaciones.

#### **9) ¿Cómo podría aplicarse la investigación de este paper en la práctica? ¿Existen aplicaciones potenciales específicas que sean relevantes?**

Al utilizar **Causal Forests**, los formuladores de políticas y diseñadores de programas pueden identificar subgrupos específicos dentro de una población que probablemente se benefician más de una intervención. Este enfoque permite una focalización más eficiente de recursos, maximizando el impacto de las intervenciones.

Otra aplicación clave es el diseño de intervenciones personalizadas. Los resultados generados por los **Causal Forests** pueden utilizarse para adaptar las políticas a las necesidades de diferentes subgrupos, aumentando la efectividad de las intervenciones. Esto es especialmente útil en contextos donde los recursos son limitados y es fundamental maximizar los beneficios para las poblaciones objetivo.

El método también tiene aplicaciones más allá de las políticas sociales. En educación, puede ayudar a identificar estudiantes que se beneficiarían más de tutorías o becas específicas. En el ámbito de la salud, se podrían focalizar intervenciones preventivas o tratamientos en poblaciones de alto riesgo basándose en características demográficas o clínicas.

La optimización de políticas públicas es otra área donde este enfoque podría ser transformador. Los **Causal Forests** permiten predecir cómo las políticas planificadas afectarían a diferentes segmentos de la población, lo que facilita el ajuste de estrategias para maximizar su impacto. Incluso en el sector privado, los principios presentados en el paper podrían aplicarse para personalizar servicios o productos según las características de los clientes, mejorando la eficiencia y la satisfacción.

#### **10) ¿Qué preguntas quedan sin respuesta después de leer este paper? ¿Cómo podrían ser abordadas en trabajos futuros?**

Aunque el paper ofrece importantes avances en la aplicación de los **Causal Forests**, quedan varias preguntas abiertas que podrían ser abordadas en futuros estudios para maximizar la utilidad de esta metodología y ampliar su impacto.

Uno es el impacto de las **variables no observadas** en los efectos estimados. Si bien el estudio se basa en datos observables, las variables omitidas podrían introducir sesgos significativos en los resultados. Futuras investigaciones podrían emplear técnicas como análisis de sensibilidad o el uso de variables instrumentales para mitigar este problema, asegurando una representación más completa de los factores que influyen en los efectos de tratamiento.

Otra interrogante clave es la **generalización de resultados**. Los hallazgos están basados en un programa específico y en un contexto urbano en Chicago, lo que plantea la duda de si los mismos patrones de heterogeneidad se observarían en otras regiones, programas o poblaciones. Replicar el análisis en diversos contextos podría proporcionar una comprensión más global y validar la robustez de los resultados.

Los **mecanismos subyacentes** que explican la heterogeneidad en los efectos del tratamiento también requieren mayor exploración. Aunque los **Causal Forests** identifican subgrupos con diferentes respuestas al tratamiento, no explican por qué ocurre esta variación. Estudios cualitativos o experimentos complementarios podrían arrojar luces sobre las dinámicas que impulsan estos resultados.

Asimismo, sería valioso analizar la **interacción con otras políticas públicas**. Las sinergias o conflictos entre diferentes programas pueden influir en los resultados de manera significativa. Evaluar estas combinaciones podría mejorar la integración de políticas y optimizar su impacto colectivo.

Finalmente, es esencial abordar los **aspectos éticos y de implementación**. El uso de hallazgos para focalizar intervenciones podría, inadvertidamente, estigmatizar a ciertos subgrupos. Diseñar lineamientos éticos y estrategias de comunicación que minimicen estos riesgos será crucial para garantizar un uso justo y responsable de la metodología.

