Técnicas de ciencia de datos para la modelización de la respuesta incremental en problemas de clasificación binaria

escrito por

JAVIER CARRASCO SERRANO

Tutor: Dr. Jorge Martín Arevalillo

Facultad de Ciencias
UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Trabajo presentado para la obtención del título de
Máster Universitario en Matemáticas Avanzadas de la UNED.
Especialidad en Estadística e Investigación Operativa

JUNIO 2022

ABSTRACT

Abstract en español:
En este trabajo se estudian diferentes enfoques para la modelización de la respuesta incremental en problemas de clasificación binaria, así como los métodos para evaluarlos. Estos enfoques son útiles para definir la población a incluir en campañas de marketing directo, especialmente porque cambian el paradigma frente a modelos predictivos tradicionales y ponen el foco en los clientes persuasibles, los que se espera que respondan positivamente debido al tratamiento. Se estudia además un caso de uso en banca donde se estiman y comparan distintos modelos.

Abstract in English:
In this work, we study different approaches to model the incremental response or uplift in the binary classification problem as well as their evaluation methods. These approaches are useful to define the population to be included in a direct marketing campaign, especially since there is a paradigm change with respect to traditional predictive models: they focus on the persuadable individuals, the ones who are expected to respond positively due to the treatment. It is also studied a business case in banking where different models are estimated and compared.

Keywords: uplift modelling, true-lift modelling, net lift, prescriptive analytics, persuadable, business analytics, direct marketing, causality, literature survey.

DEDICATORIA Y AGRADECIMIENTOS

A la gente que sobrelleva mis ausencias.

Agradecimientos. La realización de este trabajo hubiera sido imposible sin la motivación y el apoyo del Dr. Jorge Martín Arevalillo: por la presentación de un problema que desconocía y que ha resultado ser muy interesante y con bastante utilidad práctica; por la cantidad, calidad y rigor de sus comentarios, que ayudan a que la lectura de la memoria por parte de un tercero sea mucho más agradable, pues contenía alguna explicación verdaderamente farragosa; y por aguantar estoicamente la irregularidad con la que me he podido dedicar a este trabajo.

TABLA DE CONTENIDOS

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

FIGURA **Página**

CAPÍTULO 1

INTRODUCCIÓN Y MOTIVACIÓN: ENFOQUE PREDICTIVO FRENTE AL ENFOQUE DE RESPUESTA INCREMENTAL

El 4 de noviembre de 2008, Barack Obama cambió el transcurso de la historia política en Estados Unidos al convertirse en el primer presidente afrodescendiente de este país, gracias a su victoria sobre el republicano John McCain. Con una incuestionable capacidad oratoria y carisma, los analistas políticos señalaban como claves en su victoria el rechazo a las políticas de George Bush (en especial a las guerras de Irak y Afganistán), la crisis financiera… y una campaña de comunicación que fue capaz de movilizar masivamente a "sus" votantes, y de persuadir a abstencionistas, votantes indecisos o incluso tradicionales votantes republicanos. ¿Responde dicha persuasión únicamente a la capacidad retórica de Barack Obama y al inolvidable "yes, we can"? Además, en 2012 revalidaría su mandato al imponerse a Mitt Romney en un entorno de caída del apoyo popular a su figura y con las encuestas pronosticando un empate técnico.

El sistema electoral estadounidense otorga todos los representantes de cada estado al partido ganador en esa circunscripción, por lo que si, por ejemplo, Mitt Romney hubiera ganado en el estado de California (55 representantes en las elecciones de 2012) y en otro estado que le reportara al menos 9 representantes, como Nueva York (29) o Florida (29), tres de los que más representantes otorgan, el resultado hubiera sido favorable a los republicanos, que habrían pasado de 206 representantes a 290, mientras que los demócratas habrían bajado de 332 a 248. El equipo de Obama se enfrentaba así a un problema de optimización en el que tenían que distribuir los recursos de la campaña (inversión en marketing, actos del candidato en cada estado…) entre los distintos estados para alcanzar, no sólo el mayor número de votos, sino el mayor número de representantes[1].

[1] Por ejemplo, en las elecciones de 2016, Hillary Clinton consiguió el 48% de los votos frente al 46% de Donald Trump, que sin embargo se alzaría con la victoria gracias a sus 304 representantes por los 227 de la demócrata.

Pero, ¿es razonable una victoria demócrata en un estado como Texas (38)? La historia reciente dice que no: la última victoria que cosecharon allí fue en 1976. Los republicanos, por el contrario, no ganan en California desde 1984. En cambio, se dice que quien gana en Ohio (20), gana las elecciones, por haber sido así en 29 de las 32 últimas ocasiones.

Dicha asignación de recursos debería ser apoyada por modelos estadísticos que sirvieran para predecir el apoyo en cada estado, aunque si no se tuviera el cuidado suficiente en su construcción, estos modelos podrían recomendar hacer una inversión demasiado alta en California para garantizar sus 55 representantes (donde la diferencia de apoyo fue de 23 puntos porcentuales para los demócratas), o en Texas para intentar conseguir sus 38 (con 16 puntos de diferencia a favor de republicanos). La peculiaridad de estos dos estados es que, posiblemente, por muchos recursos que se destinen en campaña a ellos, no van a cambiar su voto, porque hay una masa de votantes fieles a cada uno de los partidos a los que no es posible persuadir para cambiar su papeleta.

Yendo a un caso extremo y suponiendo que se dispone de presupuesto en un estado para incluir en una campaña de marketing al 10% del censo, si se elige a este 10% de manera aleatoria (o estratificada utilizando variables que se consideran fundamentales para determinar el voto, como la edad, los ingresos, el género o el nivel de estudios) sobre el censo, se estarían incluyendo en la campaña en torno al 10% de los afiliados al partido demócrata y al 10% de los afiliados al partido republicano en ese estado, lo que no tendría ningún efecto en el voto de estos individuos, ya que votarían por el partido al que están afiliados en cualquier caso, tanto si son incluidos en la acción de marketing como si no, malgastando así parte del presupuesto de la campaña que podría ser utilizado para persuadir a potenciales votantes. Una campaña en la que sólo se incluyeran a afiliados del propio partido podría parecer un éxito si se pudiera observar que todos ellos acaban votando a dicho partido, pero no lo harían por el efecto de la campaña.

La asignación de recursos no se limita sólo a la cantidad de dinero que gastar en publicidad o las horas que el candidato debe invertir allí, sino también al diseño de las propias políticas incluidas en el programa electoral: por ejemplo, podría haber un estado muy preocupado por las políticas medioambientales, y en cambio otro que fuera contrario a dichas políticas por las posibles restricciones a determinada industria y su potencial impacto negativo en el empleo. Es decir, también existen votantes o estados para los que determinada acción genera un impacto negativo en el sentido del voto. Así, es posible identificar a distintos estados, o a distintas subpoblaciones dentro de cada uno de los estados y, por ejemplo, se podría dividir un estado entre votantes fieles al partido republicano, que no cambiarán su voto por muchos recursos que inviertan los demócratas allí; votantes fieles al partido demócrata, que les darán su apoyo en cualquier caso, sin necesidad de ser incluidos en ninguna campaña; votantes a los que es posible persuadir para ganar su voto; y votantes a los que es posible disgustar para perderlo o hacer que voten al partido contrario. En los dos últimos grupos encontraríamos por ejemplo a abstencionistas o votantes que son propensos a cambiar su voto en las distintas elecciones.

Algunos estudios señalan que, de hecho, una de las claves de la victoria de Obama en 2012 fue el uso de modelos predictivos y, en particular, unos modelos que eran capaces de señalar qué individuos debían ser incluidos en las campañas por ser votantes persuasibles. Por ejemplo, en Samuelson (2013) se destaca el uso de analítica avanzada en ambas campañas, especialmente en la de 2012, y cómo fueron capaces de identificar pequeños grupos de individuos que responderían positivamente ante mensajes adaptados a unos intereses similares, así como el uso de redes sociales para recoger información de los votantes y enviarles estos mensajes diseñados de acuerdo a sus preferencias particulares. En Siegel (2013) se destaca:

- el paso de predicciones a nivel estado a predicciones a nivel individuo, que permitieron tomar decisiones (diseñar campañas) a ese nivel;

- que en lugar de predecir qué electores votarían por Obama si fueran incluidos una campaña, predicen qué electores votarían por Obama influidos por dicha campaña (es decir, son persuasibles);

- el uso de distintos tratamientos: estudiaron a quién sería posible persuadir para votar por Obama gracias a una llamada telefónica, una visita puerta a puerta, un correo postal o un anuncio televisivo;

- que se intenta identificar a los electores que votarían por Obama aunque no fueran incluidos en la campaña, y a los electores que cambiarían su voto y votarían a Romney si fueran incluidos en la campaña demócrata;

- la construcción de distintos modelos de persuasión para cada estado o grupos de estados similares, en lugar de un único modelo para todo el país, de especial interés en los estados "bisagra", que son los que acaban determinando el ganador de las elecciones.

En Scherer (2012) y en Issenberg (2013) se enfatizan aspectos similares: la importancia del trabajo de los responsables de modelos estadísticos en la campaña de Obama de 2012, Dan Wagner y Rayid Ghani, que por ejemplo tenía trabajos previos tan diversos como determinar qué tratamiento produciría una mejor respuesta en cada paciente o qué descuento conceder a cada cliente de un supermercado para alcanzar el mayor beneficio, y que estaban a cargo de un equipo de más de 50 científicos de datos, aproximadamente cuatro veces más que en la campaña de 2008.

De hecho, Daniel Porter, director de modelización estadística para la campaña de Obama en 2012, expone en Porter (2013) cómo abordaron el diseño de las campañas de marketing en estas elecciones con el uso de modelos estadísticos. Es necesario destacar en primer lugar que en las elecciones de congresistas y senadores de 2010 sufrieron una dura derrota al perder el control del congreso y ver reducida su ventaja en el senado, lo que dificultó la gobernabilidad de la segunda mitad de la primera legislatura de Obama, y anticipaba la posibilidad de una

derrota en las elecciones de 2012. Identificaron que habían perdido a votantes que habían votado por Obama en las elecciones presidenciales de 2008, y consideraban que para repetir la victoria era necesario persuadir a esos votantes demócratas que en 2010 optaron por la abstención o un candidato republicano.

Para definir una metodología que les permitiera alcanzar dicho propósito, se centraron en cuatro cuestiones: "¿Cómo se puede convencer a los votantes de que Obama es una mejor elección que Mitt Romney? ¿Cómo elegir los mensajes en los que centrar la campaña? ¿Cómo asegurarse de que dichos mensajes no producen el efecto contrario al buscado en algunos votantes? ¿Cómo determinar a qué votantes incluir en las campañas?"

Hasta las elecciones de 2012, el diseño de las campañas electorales de publicidad se hacía en base a suposiciones y se incluía en ellas a supuestos votantes independientes, los que no estaban afiliados a ningún partido, que eran a los que entendían se podía persuadir. Se dieron cuenta del potencial de introducir un cambio de paradigma: no se centraron en predecir quién podría votar a Obama, o estimar quién estaba indeciso, o a quién le preocupaba determinado aspecto de la campaña (economía, trabajo, medio ambiente…), sino en predecir quién era propenso a cambiar su voto de Romney a Obama como resultado de ser incluido en una campaña, pero no lo haría sin la existencia de dicha campaña.

En principio, parecería imposible poder construir modelos de predicción de voto a nivel individual, porque no se conoce el voto a este nivel, pero es cierto que después de las elecciones se realizan encuestas con amplias muestras para estudiar los perfiles de voto de cada uno de los partidos en las distintas regiones (en España, por ejemplo, las lleva a cabo el CIS). Igualmente, también es complejo conocer el efecto de campañas anteriores en los votantes por el mismo motivo, pero los partidos tienen mucho interés en medir dicha efectividad e incluyen en este tipo de encuestas cuestiones para estudiar si los votantes han cambiado la intención de voto por la campaña electoral o una campaña de publicidad. Pero también es posible recabar este tipo de información en otro momento: al principio de su trabajo para esas elecciones, llevaron a cabo experimentos en diversos estados para medir la potencialidad de campañas aleatorias como punto de partida. Por ejemplo, en Illinois definieron una muestra con público objetivo al que consideraban potencialmente persuasible, y lo dividieron aleatoriamente en un grupo de control y otro de tratamiento, al que llamaría un voluntario del partido para convencer de que Obama era el candidato al que debían votar. Después midieron mediante encuestas en esta muestra la intención de voto de cada uno de los candidatos: en el grupo de control la intención de voto era de un 47.8%-52.2% para Romney, mientras que en el grupo de tratamiento la intención de voto era del 52.7-46.9% para Obama. El resultado final en el estado de Illinois, donde llevaron a cabo este experimento, fue de 57.60% para Obama frente al 40.73% para Romney.

Así, construyeron modelos para cada estado basándose en cómo se persuadió en elecciones anteriores para definir los distintos tratamientos que se podían aplicar, y utilizando variables como los ingresos, el nivel educativo, el historial de voto, comportamiento online, variables de consumo, la raza, el género o el estado civil. Los modelos utilizados fueron regresiones

logísticas con las habituales variables explicativas, pero además incluyendo variables de tratamiento y la interacción entre las variables de tratamiento y las variables independientes, para poder capturar así los efectos de tratamiento heterogéneos. Por ejemplo, el modelo con un único tratamiento sería:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n + \gamma T + \sigma_1 X_1 T + \cdots + \sigma_n X_n T$$

donde p es la probabilidad de que un individuo vote por Obama, $\beta_1, \dots, \beta_n$ los coeficientes de las variables explicativas, γ el coeficiente que mide el efecto de incluir a un individuo en la campaña, $\sigma_1, \dots, \sigma_n$ los coeficientes que miden el efecto heterogéneo del tratamiento respecto a las variables explicativas, y β_0 el intercepto. Al calcular la diferencia en la estimación de la respuesta de un individuo cuando era incluido ($T = 1$) o no ($T = 0$) en la campaña, obtenían una métrica para determinar si incluirlo o no en dicha campaña.

Algunas de las distintas campañas incluidas fueron una llamada telefónica por parte de voluntarios del partido, el contacto por correo postal explicando la posición del partido respecto a un tema relevante, el contacto por redes sociales o anuncios de televisión. Pero, además, dentro de cada una de estas tipologías no hay un único tratamiento, sino que además se ajustaban a nivel individuo (a conjuntos de individuos con características e intereses similares), intentando centrar el mensaje en contenidos que acabaran persuadiendo a los votantes.

Aparte de la regresión logística, se probaron otro tipo de modelos que acabaron descartados; por ejemplo, el enfoque de dos modelos se descartó porque se acumulaban los errores y enfatizaba en medir los efectos principales en lugar de los efectos incrementales, y los enfoques basados en árboles por presentar altos sobreajustes. Estos enfoques serán explicados en el Capítulo 2.

Daniel Porter concluye que estos modelos sirvieron para dos propósitos: identificar qué votantes eran más propensos a cambiar su voto o acudir a votar, y medir la influencia de una campaña o tratamiento en dicha decisión. Al combinar estos resultados, el modelo les permitió maximizar el retorno de la inversión en marketing. Habían transformado un problema predictivo en cada uno de los estados de EEUU (o la definición del público objetivo de una campaña), en el que debían predecir el apoyo al candidato en función de diversas variables que fueran manipulables por la campaña (inversión de recursos), en un problema en el que se predecía el impacto diferencial de la campaña en esas poblaciones, es decir, el aumento del voto debido exclusivamente a la campaña, que podría ser medido comparando el porcentaje de voto[2] de la gente incluida en una campaña frente al resto de la población. Es lo que se conoce como modelos de respuesta incremental, en contraste con los modelos predictivos tradicionales. Se trata así de un caso en el que el problema de diseño de una

[2] Como ya se ha mencionado, no se puede observar a nivel individuo si se ha votado por un determinado partido, pero las encuestas que se realizan después de las elecciones pueden ayudar a estimar la efectividad de estas campañas.

campaña de marketing pasa de utilizar modelos predictivos tradicionales a modelos de respuesta incremental, y en el que además se hace uso del multitratamiento.

Obviamente, no se limitaron a explotar la potencialidad de estos modelos únicamente para persuadir a posibles votantes, sino que lo extendieron a otros problemas similares en los que tenían que asignar una cantidad de recursos limitada a distintas campañas con objeto de maximizar la respuesta incremental. Por ejemplo, también utilizaron modelos de respuesta incremental para la recaudación que sirvió para financiar la campaña, donde se recogen recursos de afiliados y simpatizantes del partido. Dicha recaudación, de 722 millones de dólares, se mantuvo en niveles similares a los conseguidos en 2008, con 779 millones, a pesar de la caída de popularidad de Obama. Por otro lado, Romney sólo fue capaz de recaudar 467 millones.

1.1. Enfoque tradicional

Como se ha ilustrado en el ejemplo anterior, los problemas de asignación de recursos para motivar una respuesta positiva se han tratado tradicionalmente como un problema de predicción en el que modelizar la respuesta de un problema de clasificación binaria.

En estos problemas, se dispone un presupuesto limitado para una acción comercial o campaña de marketing directo, es decir, a nivel individuo, y se debe delimitar la población objetivo de dicha campaña con el fin de maximizar el retorno de la inversión, para incluir en la campaña por ejemplo a los que son más propensos a comprar un determinado producto, si esa fuera la situación que se está abordando.

Si no se dispone de un histórico de respuesta ante campañas similares, es decir, si no ha habido campañas previas o si se desconoce si el individuo al que se ha incluido en una campaña compra finalmente el producto que se le ha ofrecido o no, en algunos casos (productos que no se contratan si no existe una acción comercial) la única manera de determinar los clientes que serán incluidos en la campaña es utilizando algún criterio experto o hacerlo de manera aleatoria. Existirán otros casos donde, a pesar de no haber existido una campaña previa, sí que se pueda construir un modelo sobre la población total que simplemente estime la probabilidad de respuesta positiva, centrando entonces los esfuerzos de la campaña en los clientes con una mayor probabilidad estimada. Si, por ejemplo, se utilizara una regresión logística, se tendrían que estimar los parámetros del modelo

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

donde p es la probabilidad de respuesta a estimar, $\beta_1, \ldots, \beta_n$ los coeficientes de las variables explicativas y β_0 el intercepto. Los individuos con probabilidades de respuesta estimada más altas serían incluidos en una campaña de marketing: una vez se ha estimado un modelo como el anterior sobre una muestra histórica, se puntúa la muestra de clientes a los que se está estudiando incluir o no en la campaña, y se segmentan las probabilidades estimadas en

grupos heterogéneos o se utilizan los deciles (Figura 1) de la probabilidad estimada para distribuir los recursos de la campaña en los tramos con mayores probabilidades estimadas. El objetivo en esta metodología sería identificar a los clientes que responden positivamente dentro de los que se pueden incluir en una campaña, para garantizar el éxito de la acción comercial.

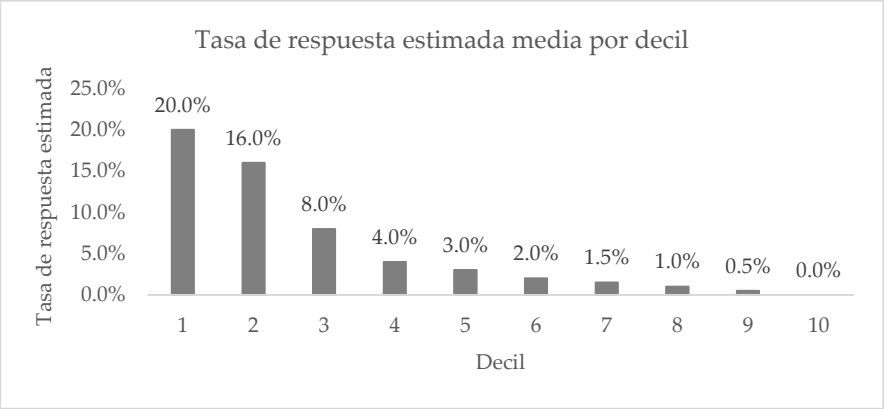


FIGURA 1

Si en el ejemplo del gráfico anterior, donde el modelo ha sido estimado sobre una población en la que no ha habido ninguna campaña previa, se decidiera incluir en la campaña a los tres primeros deciles, se tendría una tasa de respuesta estimada media del 14.67%[2] (media de los tres primeros deciles) en el grupo de tratamiento que se acaba de definir frente a un 1.71%[3] (media de los 7 últimos deciles) en el grupo de control. Si los recursos a destinar a la campaña fueran algo más limitados y sólo se pudieran incluir los dos primeros deciles en la campaña, o si la métrica de rentabilidad de la inversión en marketing que se utilizara determinara que sería más rentable actuar así al tener en cuenta el coste de incluir un individuo en la campaña y el retorno de una respuesta positiva, la tasa de respuesta estimada media sería del 18%[2] (media de los dos primeros deciles) para ese grupo de tratamiento frente a un 2.5%[2] (media de los 8 últimos deciles) para el grupo de control. La efectividad de dicha campaña o el incremento en la respuesta en la población contactada se podrá medir, una vez llevada a cabo la acción comercial, al comparar la tasa de respuesta de la población incluida en la campaña frente a la tasa de respuesta del resto de individuos, los que pertenecían al escenario base o grupo de control.

Una vez se dispone de una muestra en la que ya se ha realizado una campaña similar, tanto si la campaña previa se ha realizado de manera aleatoria como si se ha diseñado con un modelo que predecía la respuesta positiva, es posible incorporar esa información para construir un modelo sobre la población tratada que permita diferenciar aún más entre los que son propensos a responder positivamente y los que no (Figura 2), es decir, un modelo que

[3] Nótese que, como el modelo se ha estimado en una muestra en la que no había ninguna campaña previa, se deben entender estas tasas estimadas como la tasa de respuesta estimada sin efectos de campañas, por lo que es razonable esperar que las tasas de respuesta que se observen tras la campaña varíen debido al efecto de ésta.

permita incrementar la diferencia entre las tasas de respuesta de los grupos de tratamiento y control, con el objetivo de maximizar la rentabilidad de la campaña (ROMI, *return on marketing investment*) en lugar de centrado en la respuesta positiva de la población total.

Incluidos en el tratamiento (T)	TR	TN
	Responden positivamente (R)	No responden positivamente (N)

FIGURA 2

Con esta información, e igual que en el caso anterior, el modelo que estime la probabilidad de respuesta positiva cuando un individuo es incluido en una campaña podría ser un modelo estadístico tradicional como una regresión logística, o uno que utilice técnicas de *machine learning*, como los modelos basados en árboles o redes neuronales. La especificación en el caso de la regresión sería la misma que en el caso anterior, pero se estimaría únicamente sobre poblaciones incluidas en el grupo de tratamiento previo:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

o, yendo a un caso más general en el que se pueda utilizar otro tipo de modelos, se debería estimar

$$p(Y = 1|X_1, \dots, X_n)$$

donde Y es la variable de respuesta binaria a estimar, que toma el valor 1 cuando la respuesta es positiva y 0 en caso contrario, y $X_1, \dots, X_n$ son las variables explicativas del modelo.

Igualmente, se incluiría en la nueva campaña a los individuos con estimaciones de la probabilidad de respuesta más altas.

Algunos problemas en los que la definición de la población de tratamiento se puede abordar con modelos predictivos serían:

- Problemas de marketing en los que se modeliza la propensión a la compra o contratación de un producto (seguro de vida o cambio de compañía eléctrica), la suscripción a un servicio (Spotify Premium o Amazon Prime), o el incremento de la cuota o tarifa de un contrato existente (incrementar el servicio contratado con una compañía telefónica). En estos casos, la respuesta a modelizar es la compra o contratación de un producto o servicio, y la acción comercial o tratamiento a efectuar sería la oferta de dicho producto, ya sea mediante un mensaje de texto, un correo electrónico, una llamada telefónica, o un anuncio en Youtube.

- Problemas de fuga de clientes, en los que la respuesta positiva sería permanecer como

cliente y la acción comercial podría ser un descuento en el servicio contratado.

- Problemas de recursos humanos como la retención de empleados, donde la respuesta positiva es que el empleado permanezca en la empresa, y la acción podría ser un ascenso, aumento salarial o cambio o flexibilización de condiciones laborales.

- Problemas como el presentado al principio del capítulo, en el que se quiere conseguir el voto de un indeciso o simpatizante de otro partido para las próximas elecciones, o una aportación a la financiación de la campaña electoral por parte de un afiliado o simpatizando, donde la respuesta positiva sería el apoyo al partido político (ya sea con el voto o con una donación), y las acciones irían desde una llamada de un voluntario del partido animando al apoyo a ese candidato o el envío de publicidad postal, hasta el diseño de políticas que favorezcan a colectivos particulares entre los que ese individuo está incluido.

- En medicina, es posible definir a qué pacientes se les aplica determinado tratamiento utilizando como variable de respuesta la evolución positiva de la enfermedad y como acción la aplicación o no de un tratamiento particular, como por ejemplo el uso o no de radio o quimioterapia, o la extirpación o no de un tumor.

- En el ámbito financiero, también es posible definir de esta manera problemas como los recobros de clientes que impagan un préstamo, la propensión a la contratación de un fondo de inversión o planes de pensiones, la prevención de impagos o la retención de pasivo (depósitos) en la entidad.

Este enfoque tradicional presenta diversos problemas que pueden comprometer la robustez de las estimaciones y el objetivo que se pretende alcanzar al usarlos:

- En primer lugar, al estimarse contra una muestra, como en cualquier otro problema de inferencia o predicción, es necesario garantizar la representatividad de esa muestra respecto a la población en la que se va a separar entre tratamiento y control para llevar a cabo la campaña. Si se construye el modelo con una muestra histórica, o si se construye con la información de una campaña aplicada en una ciudad y se quiere extender a otra, hay que asegurar que el comportamiento de las poblaciones es similar. Es decir, la muestra en la que se estime el modelo debe ser similar a la muestra en la que se aplique.

- En estos modelos se estima la probabilidad de respuesta positiva, y se incluyen en la campaña los individuos con una mayor propensión (o estimación) a la respuesta positiva, por lo que el modelo está identificando o recomendando incluir en la campaña por igual a individuos que responderían positivamente sin necesidad de la acción comercial que a individuos que responden positivamente debido a dicha acción. Estiman una probabilidad de respuesta, que no mide el efecto incremental de la acción comercial y por tanto no son capaces de separar entre estos dos subgrupos,

lo que permitiría asignar mejor los recursos, pues se invertirían sólo en los que no responderían positivamente si no fuera porque son incluidos en la campaña.

- Podría darse la paradoja de que una acción comercial provoque el efecto contrario al deseado en algunos clientes, que el individuo contactado fuera a tener una respuesta positiva, y cambie a una respuesta negativa al ser incluido en la campaña. Por ejemplo, se podría acelerar una fuga de un cliente (o empleado) si se le "recuerda" que su contrato va a expirar (como en las empresas de telecomunicación al final de la permanencia) o se le podría estar animando a buscar ofertas en la competencia.

Parece claro que se realizaría una mejor asignación de los recursos en la definición de la campaña, y por tanto se elevaría su retorno o rentabilidad, si se pudieran construir modelos estadísticos que atajaran las dos últimas debilidades de los modelos predictivos tradicionales: por un lado, sería deseable poder identificar entre los individuos que son propensos a responder positivamente, cuáles lo son debido a la acción comercial y cuáles responderían positivamente en cualquier caso, aunque no sean incluidos en la campaña; y a su vez que no sólo se excluyeran de la campaña a individuos que responderían negativamente en cualquier caso, sino que también se evitara contactar a individuos que cambian su respuesta a una negativa debido a la campaña. Utilizar técnicas estadísticas que modelicen el incremento en la probabilidad de respuesta debido a la campaña o acción comercial, lo que se conoce en la literatura como modelos de respuesta incremental, incrementaría la efectividad de la campaña.

1.2. Modelos de respuesta incremental

Los modelos de respuesta incremental surgen con el objetivo de modelizar el incremento en la respuesta positiva debido a una campaña o acción comercial. Suponen el paso de técnicas descriptivas, en las que se describe el comportamiento que se observa en base a otras variables, o modelos predictivos, en los que se predice el comportamiento de la variable de respuesta en función de diversas variables explicativas, a modelos prescriptivos, que arrojan luz acerca de las acciones a tomar para causar una respuesta positiva en un individuo. Estos modelos deberían ser capaces entonces de encontrar o separar a los individuos que pueden ser influenciados positivamente para incluirlos en la campaña, y también a los que reaccionarían negativamente para excluirlos y reducir el efecto adverso que pudiera tener. El output del modelo debería por tanto medir el incremento en la variable respuesta al aplicar un tratamiento.

Sin que sea necesario que existan campañas previas o se pueda observar la variable de respuesta, si se va a definir un grupo en la población al que incluir en una campaña comercial, y suponiendo que hay un único tratamiento posible[4], es posible dividir de manera natural a la población en cuatro grupos:

[4] Si existen distintos tratamientos (fuera del alcance de este trabajo), un individuo puede pertenecer a distintos grupos en función del tratamiento que se plantee.

- Causas seguras: clientes que responderán positivamente aunque no se incluyan en el tratamiento.

- Causas perdidas: clientes que responderán negativamente aunque se incluyan en el tratamiento.

- Persuasibles: clientes que responderían positivamente sólo si se incluyen en el tratamiento, y que responderían negativamente en caso contrario.

- Clientes "no molestar": clientes que responderían negativamente sólo si se incluyen en el tratamiento, y que responderían positivamente en caso contrario.

El objetivo teórico de estos modelos sería entonces poder separar estas cuatro poblaciones, para invertir recursos en acciones comerciales únicamente en el colectivo de persuasibles, pues es donde se encuentra la potencial ganancia; evitando los grupos de causas seguras y causas perdidas, pues aunque no generan una variación en la variable de respuesta, y por tanto en los ingresos que la puedan acompañar, sí que generan un gasto innecesario de recursos en el tratamiento; y evitando especialmente la subpoblación de clientes a los que no molestar, porque no sólo se perderían los recursos que se invierten en la campaña, sino que también se perderían individuos que iban a responder positivamente y cambian su respuesta.

El problema de esta división teórica es que no es posible identificar en la población total a qué grupo pertenece cada uno de los individuos.

En cambio, si se han aplicado campañas previas a la población, o a una población similar, sí que es posible observar subgrupos de la población por dos ejes:

- En función de la marcación de campaña previa, es posible separar a la población de tratamiento y el grupo de control.

- En función de si los individuos responden positivamente o no.

Así, cruzando estas dos variables, los individuos de una población que ha pasado por una campaña previa se pueden dividir en las siguientes cuatro casuísticas (Figura 3):

- CR: individuos que pertenecían al grupo de control y respondieron positivamente.

- CN: individuos que pertenecían al grupo de control y no respondieron positivamente.

- TR: individuos que pertenecían al grupo de tratamiento y respondieron positivamente.

- TN: individuo que, aun perteneciendo al grupo de tratamiento, no respondieron positivamente.

Incluidos en el tratamiento (T)	TR	TN
No incluidos en el tratamiento (C)	CR	CN
	Responden positivamente (R)	No responden positivamente (N)

FIGURA 3

Lo importante de esta división es que, en general[5], si se ha realizado una campaña previa sí que se puede conocer a cuál de estos cuatro grupos pertenecía un individuo, por lo que tiene sentido plantearse si es posible relacionar estas cuatro categorías con la división teórica que persigue el modelo: causas seguras, causas perdidas, persuasibles e individuos a los que no molestar. Aunque esta relación no sea unívoca, si que es posible determinar que (Figura 4):

- CR: en el grupo de control los individuos no han sido contactados, por lo que, si han respondido positivamente, sólo pueden ser causas seguras o individuos a los que no molestar.

- CN: por la misma razón, si estos individuos no responden positivamente, al no haber sido tratados sólo pueden ser causas perdidas o **persuasibles**.

- TR: en el grupo de tratamiento los individuos han sido incluidos en una acción comercial, por lo que si responden positivamente es porque son causas seguras o **persuasibles**.

- TN: si, a pesar de haber sido contactados estos individuos no han respondido positivamente, es porque son causas perdidas o clientes a los que no se debería haber molestado.

[5] Por ejemplo, en el caso de las elecciones no es posible observar la variable respuesta, si finalmente un elector votó por un candidato u otro, aunque sí que se puede estimar mediante encuestas posteriores a las elecciones si determinado perfil de votante que fue incluido en la campaña votó, o en qué proporción lo hizo, por el candidato promovido por la campaña.

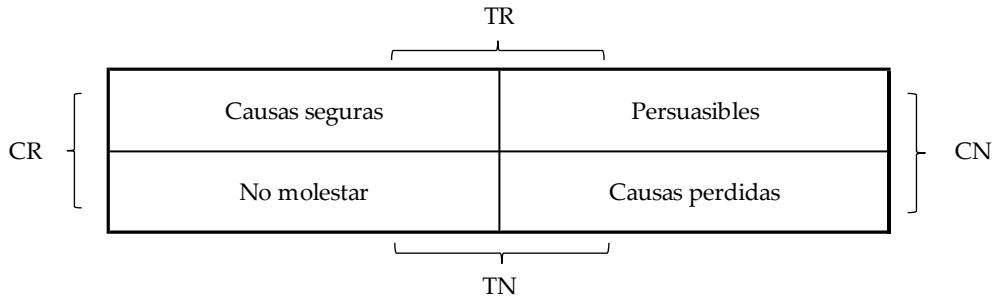


FIGURA 4

De esta manera se ha identificado que los individuos persuasibles, a los que se quiere incluir en una nueva campaña, pertenecían a los grupos CN y TR de la campaña previa, y que los individuos a los que no se debe molestar, para evitar que cambien su respuesta a una negativa, estaban en los grupos CR y TN. Se ha transformado así el objetivo teórico de identificar la subpoblación de persuasibles, o de separarla de la de "no molestar", en la identificación de la unión de los grupos CN y TR, o su separación respecto de los grupos CR y TN de la campaña anterior, todas ellas etiquetas observables en la muestra.

Se define entonces la respuesta incremental, o *uplift*[6], como la diferencia de la probabilidad (condicional o a posteriori) de responder positivamente entre la población con y sin tratamiento:

$$uplift = p(R|T) - p(R|C)$$

que para un individuo i en particular sería la diferencia entre la estimación de dichas probabilidades:

$$\widehat{uplift}(\boldsymbol{x}^i) = \hat{p}\left(y^i = 1|t^i = 1, x_1^i, \dots, x_n^i\right) - \hat{p}\left(y^i = 1|t^i = 0, x_1^i, \dots, x_n^i\right)$$

donde y^i es el valor que toma la variable de respuesta, y, para el individuo i, $x_1^i, \dots, x_n^i$ son los valores que toman las variables explicativas $x_1, \dots, x_n$ para ese individuo, y t^i es una variable binaria que toma el valor 1 si el individuo recibe el tratamiento y el valor 0 si no lo recibe. Nótese que, en una campaña previa, un individuo pertenece o bien al grupo de tratamiento, o bien al grupo de control (pertenece únicamente a uno de los grupos TR, TN, CR o CN), y por tanto no se puede observar cuál hubiera sido el efecto en su comportamiento o respuesta si se hubiera cambiado del grupo de tratamiento al de control o viceversa, pero que, al realizar una nueva campaña, se va a discutir si un individuo es incluido en un nuevo grupo de tratamiento o no, por eso se hace uso de ambas estimaciones (con $t^i = 1$ y $t^i = 0$ respectivamente). Por ejemplo, se pueden encontrar más discusiones y detalles sobre el problema de la inferencia causal en los modelos de respuesta incremental en Gutiérrez y

[6] En la literatura, y en especial en *papers* en los que se hace una revisión del estado del arte de los modelos de respuesta incremental, como puedan ser Devriendt *et al* (2018), Kane *et al* (2014), Rößler *et al* (2021) o Devriendt *et al* (2021), se hace referencia a ellos indistintamente con expresiones como modelos de *uplift, true lift, net response, net lift, true response*, modelos de persuasión o de marketing diferencial.

Gérardy (2017).

Este enfoque, y en especial la ecuación anterior, contrastan con el enfoque del modelo tradicional en el que simplemente se estimaba la probabilidad de que un individuo respondiera positivamente cuando es incluido en la acción de marketing directo:

$$\hat{p}\left(y^i = 1 | t^i = 1, x_1^i, \dots, x_n^i\right)$$

Y contrastan mucho más con enfoques primitivos en los que sólo se estima la probabilidad de respuesta positiva (por no haber disponibilidad de campañas previas o similares):

$$\hat{p}\left(y^i = 1 | x_1^i, \dots, x_n^i\right)$$

Parece entonces claro que la potencialidad de los modelos de respuesta incremental respecto a los modelos predictivos está en que pueden generar mayores tasas de respuesta, y por tanto mayores retornos, incluso reduciendo el tamaño de la población tratada, al reducir la presencia en ésta de clientes a los que no se debe molestar, causas seguras y causas perdidas. Aun así, los enfoques que se plantearán en el siguiente capítulo no tienen en cuenta totalmente el objetivo final del usuario de esos modelos (de acuerdo a Devriendt *et al* (2018)), que en un problema particular podría ser maximizar el retorno de una campaña comercial; emplear alguna métrica de rentabilidad *ad hoc* al problema, como por ejemplo el ROMI (*return on marketing investment*), permitiría formular el problema como un problema de optimización en el que se debe maximizar dicha métrica de rentabilidad, minimizar las pérdidas… dadas unas restricciones como un presupuesto de marketing limitado o un número máximo de campañas a aplicar en un cliente en el tiempo. Este problema está abierto, aunque existen publicaciones al respecto como Devriendt *et al* (2021), donde se trata el problema particular de fuga de clientes con esa visión de optimización de una métrica de rentabilidad.

Otra extensión clara del problema sería cuestionarse si es posible aplicar uno o varios tratamientos a la vez a un mismo individuo, o si es mejor utilizar un tratamiento para una subpoblación determinada y otro tratamiento para otra. Los modelos de respuesta incremental con multitratamiento optimizan, no sólo la selección de clientes (quiénes entran en la campaña), sino también los distintos tratamientos a aplicar. Estos modelos pueden mejorar la tasa de respuesta porque modifican la partición de la población total en las categorías de persuasibles, causas seguras, causas perdidas y "no molestar", ya que puede haber individuos que bajo un tratamiento sean considerados causas perdidas, y bajo otro tratamiento sí que sean persuasibles. Por ejemplo, un cliente podría no estar dispuesto a adquirir un determinado producto si el descuento sobre su precio es del 10%, pero sí lo adquiriría para un descuento mayor del 20%, o un empleado "quemado" se iría de la empresa aunque se le ofrezca una subida de sueldo del 2%, pero lo retendrían con una subida del 6%. En los problemas de recobros de impagos bancarios, los distintos tratamientos a aplicar según las casuísticas podrían ser: no llevar a cabo ninguna acción porque el cliente va a pagar en cualquier caso, o porque no (causas seguras y causas perdidas); simplemente recordar que han impagado y que deben regularizar su situación con un sms, carta o correo electrónico

(porque necesita un "pequeño empujón", donde estarían los persuasibles); o insistir en el recobro y ofrecer algún tipo de carencia o cambio en las condiciones de su producto (también clientes persuasibles, pero que necesitan un tratamiento con más ventajas porque de lo contrario serían causas perdidas). El problema multitratamiento no será abordado en este trabajo, aunque se puede leer más al respecto en Olaya *et al* (2020). Igualmente, es posible plantear el uso de estos modelos cuando la variable de respuesta no es binaria, sino continua. Por ejemplo, se podría utilizar para modelizar el gasto en una tienda (física o web) en el período de rebajas en función de si se incluye a los individuos en una (o diversas) campañas. En el presente trabajo sólo se cubre el problema de modelización de respuesta binaria, pero se puede encontrar más detalle sobre los modelos de respuesta incremental para variable de respuesta continua en Rudas y Jaroszewicz (2018), Gubela *et al* (2020) o Baier y Stöcker (2021), dando lugar por ejemplo a modelos de respuesta incremental para los ingresos debidos a una campaña de marketing.

Es conveniente resaltar algunos aspectos a tener en cuenta cuando se plantea el uso de modelos de respuesta incremental:

- Es necesaria la existencia de campañas previas porque se requiere la distinción entre población tratada y población de control, una restricción que no aparece en los modelos de predicción. El grupo de control no se utiliza sólo para medir la efectividad de la campaña, como podía pasar en los modelos predictivos, sino que también se utilizará en la modelización.

- Existen casos en los que utilizar un modelo de respuesta incremental no supone una ganancia respecto al uso de modelos predictivos. Estas situaciones se dan cuando la tasa de respuesta positiva en el grupo de control es cero (por ejemplo, ocurre en problemas sobre compañías o productos nuevos o no conocidos), o en productos reactivos en los que sea imposible obtener una respuesta positiva si no hay una acción comercial, algo que podría suceder incluso en un mercado muy competitivo en precios. Nótese que, si no hay grupo de CR, tampoco tiene sentido que puedan existir clientes a los que no molestar ni causas seguras, y por tanto sólo habría persuasibles y causas perdidas.

- Además del problema de clases no balanceadas entre clientes que responden positivamente y clientes que no, que también está presente en los modelos predictivos, hay un problema de balanceo de clases entre clientes en el grupo de tratamiento y en el grupo de control. Es más, si la separación entre el grupo de tratamiento y control de la campaña previa responde a cierto estudio o modelo, habrá otro sesgo que mitigar.

- Aunque, como se verá en el capítulo siguiente, existen diversos enfoques y técnicas de modelización de la respuesta incremental, no hay una técnica que sea "mejor que las otras" (*no-free-lunch theorem*[7]), puesto que hay dependencia de los datos y, por ejemplo,

[7] Aunque no es un "teorema" matemático, ésta es la manera en la que se referencia este principio en la literatura.

el enfoque de dos modelos podría ser el mejor en un problema con un conjunto de datos en particular, y el enfoque de estimación directa resultar más conveniente para abordar el mismo problema pero con otros datos, por ejemplo al extender el análisis a otra ciudad o a otro momento temporal. En el tercer capítulo, se presentarán además las métricas habituales que sirven para comparar los distintos modelos de respuesta incremental que se construyan.

CAPÍTULO 2

MODELIZACIÓN DE LA RESPUESTA INCREMENTAL

Aunque en la literatura existen diversas clasificaciones de los enfoques de modelización de la respuesta incremental (por ejemplo, en Devriendt *et al* (2018), Kane *et al* (2014), o Rößler *et al* (2021)), éstas tienen en común que los dividen en métodos indirectos, basados en técnicas predictivas tradicionales para modelizar la respuesta incremental, y métodos para modelizar directamente el *uplift*. En este primer grupo encontraremos el enfoque de dos modelos, el uso de una variable *dummy* de tratamiento, o las transformaciones en el *target*. En cambio, los enfoques de modelización directa modifican técnicas predictivas (árboles y ensamblados) para modelizar directamente la respuesta incremental.

A lo largo de este capítulo, se van a presentar los siguientes enfoques de modelización de la respuesta incremental:

- Estimación indirecta:
    - Enfoque de dos modelos.
    - Uso de *dummy* de tratamiento: enfoque de Lo.
    - Transformación de variable *target*/clases: enfoques de Lai, Kane, Kane generalizado, Lai ponderado, reflexivo, pesimista, y Jaskowski.
- Estimación directa:
    - Enfoques basados en árboles (adaptaciones de CART, C4.5, CHIAD).
    - Enfoques de ensamblado basados en la agregación de modelos.

El orden de aparición de estas técnicas en este trabajo no responde necesariamente al orden cronológico, sino que se presentan comenzando por el más intuitivo y terminando por el que entendemos que es más refinado.

Antes de entrar en el detalle de cada uno de estos enfoques, es necesario definir una serie de elementos que serán recurrentes cada vez que se presente una técnica.

- i es un individuo o cliente;
- $\boldsymbol{x} = (x_1, \dots, x_n)$ es un vector de n variables aleatorias predictivas o explicativas, que para el individuo i toma el valor $\boldsymbol{x}^i = \left(x_1^i, \dots, x_n^i\right)$;
- T y C son, respectivamente, los grupos de tratamiento y control dentro de la población a la que se refiera, y se definen las variables t y c para indicar si un individuo pertenece a estas poblaciones:

$$t^i = \begin{cases} 1 & i \in T \\ 0 & i \in C \end{cases}, c^i = 1 - t^i = \begin{cases} 1 & i \in C \\ 0 & i \in T \end{cases}$$

- R y N son los grupos de individuos de la población que han respondido positivamente y que no han respondido positivamente, y se definen las variables r y n para indicar si un individuo pertenece a estas poblaciones:

$$r^i = \begin{cases} 1 & i \in R \\ 0 & i \in N \end{cases}, n^i = 1 - r^i = \begin{cases} 1 & i \in N \\ 0 & i \in R \end{cases}$$

- Se definen los cuatro conjuntos $TR = T \cap R, TN = T \cap N, CR = C \cap R, CN = C \cap N$;
- y es la variable *target* (binaria) que se va a modelizar, que no necesariamente será si el cliente responde positivamente o no (variable r), puesto que en función del enfoque planteado podrá tomar ese valor u otro; para el individuo i se designará su valor con y^i.

También es importante recordar que, como se explica en la sección 1.2, es necesario estimar estos modelos en una población sobre la que se ha llevado a cabo una campaña previa similar, con una marcación preferiblemente aleatoria[8], y se conoce la respuesta a dicha campaña, por lo que para cada interviniente del conjunto de datos de entrenamiento es posible observar los valores de las variables t y r.

2.1. Enfoque de dos modelos o diferencia de *scores*

El primer enfoque de modelización de la respuesta incremental surge, de acuerdo a las revisiones sobre este problema que aparecen en Devriendt *et al* (2018), Kane *et al* (2014), Rößler *et al* (2021) o Devriendt *et al* (2021), de manera intuitiva al intentar estimar con modelos predictivos la diferencia en la tasa de respuesta entre el grupo de tratamiento y el de control como la diferencia de la tasa de respuesta estimada para cada una de las subpoblaciones. Ante las

[8] En caso de que la marcación de la campaña previa no fuera aleatoria, se estaría introduciendo un sesgo adicional en la definición del perímetro de la campaña y se debería estudiar la manera de reducir dicho sesgo.

dificultades de una estimación directa, se construye un modelo que estime la probabilidad de respuesta positiva, tal como se hacía en los modelos predictivos, pero estimando sólo sobre el grupo de tratamiento (de hecho, este es uno de los enfoques tradicionales habituales), y otro modelo que también estime la probabilidad de respuesta positiva, pero sólo sobre el grupo de control, estimando el *uplift* mediante la diferencia de ambas estimaciones.

Para utilizar este enfoque se define la variable a estimar sin realizar ninguna modificación, como la que indica simplemente la respuesta positiva:

$$y^i = \begin{cases} 1 & i \in R \\ 0 & i \in N \end{cases}$$

Y se estiman las probabilidades de respuesta positiva en las subpoblaciones de tratamiento y control:

$$p(R|T, \boldsymbol{x}), p(R|C, \boldsymbol{x})$$

o, dicho de otro modo, se estiman

$$p_T(y = 1|T, \boldsymbol{x}), p_C(y = 1|C, \boldsymbol{x})$$

Obteniendo los siguientes modelos predictivos para estas subpoblaciones:

$$\hat{p}(R|T, \boldsymbol{x}) \; o \; \hat{p}_T\left(y^i = 1|i \in T, x_1^i, \dots, x_n^i\right)$$

$$\hat{p}(R|C, \boldsymbol{x}) \; o \; \hat{p}_C\left(y^i = 1|i \in C, x_1^i, \dots, x_n^i\right)$$

En el enfoque de dos modelos se obtiene la estimación de la respuesta incremental como la diferencia entre las probabilidades, o *scores*, estimadas por estos dos modelos, que para un individuo i en particular sería

$$\widehat{uplift}(\boldsymbol{x}^i) = \hat{p}_T\left(y^i = 1|t^i = 1, x_1^i, \dots, x_n^i\right) - \hat{p}_C\left(y^i = 1|t^i = 0, x_1^i, \dots, x_n^i\right)$$

y que se puede entender como la probabilidad de respuesta positiva si se incluyera a ese individuo en una nueva campaña similar a la observada, menos la probabilidad de respuesta positiva si no se incluyera. Además, esta definición es una mera convención motivada por la intuición (la campaña debería generar un efecto positivo, por eso la diferencia se calcula en ese orden y no en el contrario), pero el *uplift* estimado puede tomar un valor negativo (individuos pertenecientes al colectivo al que no se debe molestar).

Es importante resaltar que, aunque estos modelos hayan sido construidos separadamente sobre las poblaciones de tratamiento y control, sí que es posible calcular la estimación de la respuesta incremental para un individuo i dados unos valores $x_1^i, \dots, x_n^i$, independientemente de si este individuo estuviera en el grupo de tratamiento, en el de control, o no estuviera incluido en ninguna de estas dos subpoblaciones, y a pesar de que, de estar incluido en una de ellas, no sea posible observar cuál hubiera sido su respuesta de haberse incluido en la otra.

Si, por ejemplo, se utilizara una regresión logística para estimar estos modelos, se obtendría la siguiente expresión:

$$\widehat{uplift}(\boldsymbol{x}^i) = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_n x_n)}} - \frac{1}{1+e^{-(\hat{\gamma}_0+\hat{\gamma}_1 x_1+\cdots+\hat{\gamma}_n x_n)}}$$

donde los coeficientes $\hat{\beta}_i$ se refieren al modelo estimado sobre la población de tratamiento, y los coeficientes $\hat{\gamma}_i$ se refieren al modelo sobre el grupo de control.

Nótese que, aunque en ambas estimaciones aparezcan las mismas n variables explicativas, las especificaciones de estos modelos se deben leer como modelos construidos utilizando esas n variables candidatas, pero teniendo en cuenta que posiblemente sólo haya $k \leq n$ y $m \leq n$ variables explicativas significativas (entre esas n variables disponibles) para los modelos de las subpoblaciones de tratamiento y control respectivamente.

Este enfoque es bastante primigenio a pesar de introducir la noción de respuesta incremental, ya que por ejemplo no se están eligiendo las variables que explican directamente la respuesta incremental, sino que tendrán más peso en las predicciones las que expliquen adecuadamente la respuesta positiva en cada una de las subpoblaciones. Además, como se señala en Radcliffe y Surry (2011), obtener la estimación de la respuesta incremental como la diferencia entre dos estimaciones puede dar lugar a que los errores de estos dos modelos se acumulen.

2.2. Enfoque de *dummy* de tratamiento

La idea detrás de este enfoque, introducido en Lo (2002), es similar a la del enfoque de dos modelos: estimar la probabilidad de respuesta positiva en el grupo de tratamiento y en el grupo de control, y obtener la respuesta incremental como la diferencia entre las estimaciones en las dos poblaciones. En cambio, para evitar los problemas de lidiar con dos modelos distintos, en este caso se construye un único modelo, que se estima sobre una única muestra que contendrá tanto las observaciones correspondientes a individuos incluidos en el grupo de tratamiento como a los incluidos en el grupo de control, utilizando para ello una variable *dummy* que sirve para indicar a qué grupo pertenece cada uno de los individuos, y que será usada como variable explicativa en el modelo de predicción de la respuesta positiva y también para calcular la estimación de la respuesta incremental.

En primer lugar, se vuelve a definir la variable a estimar como la que indica la respuesta positiva:

$$y^i = \begin{cases} 1 & i \in R \\ 0 & i \in N \end{cases}$$

Además, se genera una variable *dummy* que señala qué individuos pertenecían a la población de tratamiento y cuáles a la de control:

$$t^i = \begin{cases} 1 & i \in T \\ 0 & i \in C \end{cases}$$

Se construye un único modelo sobre la población total, es decir $T \cup C$, que estima la probabilidad de respuesta positiva:

$$p(R|T \cup C, \boldsymbol{x}, t)$$

o, con otra notación, se estima

$$p(y = 1|T \cup C, \boldsymbol{x}, t)$$

Obteniendo el siguiente modelo predictivo para la población total:

$$\hat{p}(R|T \cup C, \boldsymbol{x}, t) \; o \; \hat{p}\left(y^i = 1|i \in T \cup C, t^i, x_1^i, \dots, x_n^i\right)$$

Además, para la estimación de este modelo no sólo se utilizan como variables explicativas las ya disponibles $x_1, \dots, x_n$, sino que también se utiliza la variable de tratamiento t, que en algunos casos captura la diferencia en la tasa de respuesta media entre tratamiento y control, y su interacción con el resto de variables explicativas, $tx_1, \dots, tx_n$, que se incluyen con el objetivo de modelizar el distinto efecto que cada variable explicativa puede tener en cada una de las dos subpoblaciones. De hecho, se puede ver como un modelo que utiliza las variables $x_1, \dots, x_n$ para la subpoblación de control (cuando $t = 0$), y unas estimaciones distintas para estas mismas variables, $x_1, \dots, x_n$, y para el intercepto, a través de su interacción con la variable de tratamiento, t, para la subpoblación de tratamiento (cuando $t = 1$).

La respuesta incremental se calcula entonces como la diferencia en la estimación de la respuesta entre la población de tratamiento y la de control, que para un individuo i toma el valor:

$$\widehat{uplift}_{Lo}(\boldsymbol{x}^i) = \hat{p}\left(y^i = 1|t^i = 1, x_1^i, \dots, x_n^i\right) - \hat{p}\left(y^i = 1|t^i = 0, x_1^i, \dots, x_n^i\right)$$

Como en el enfoque anterior, esta estimación se puede entender como la probabilidad de respuesta positiva si se incluyera al individuo i con las características $\boldsymbol{x}^i$ en una nueva campaña similar a la observada, menos la probabilidad de respuesta positiva si no se incluyera.

Aunque se pueda utilizar cualquier técnica de modelización, el caso particular de estimación utilizando una regresión logística, que de hecho es el enfoque propuesto originalmente por Lo, pone en relieve el uso de las variables de interacción. Así, la estimación de la variable respuesta positiva sería

$$\hat{p}\left(y^i = 1|t^i, x_1^i, \dots, x_n^i\right) = \frac{1}{1 + exp\left(-\left(\hat{\alpha} + \widehat{\beta_1} x_1^i + \dots + \widehat{\beta_n} x_n^i + \hat{\delta} t^i + \widehat{\gamma_1} t^i x_1^i + \dots + \widehat{\gamma_n} t^i x_n^i\right)\right)}$$

que da lugar a la siguiente estimación de la respuesta incremental:

$$\widehat{uplift}_{Lo}(\boldsymbol{x}^i) = \hat{p}\left(y^i = 1 | t^i = 1, x_1^i, \dots, x_n^i\right) - \hat{p}\left(y^i = 1 | t^i = 0, x_1^i, \dots, x_n^i\right)$$
$$= \frac{1}{1 + \exp\left(-\left(\hat{\alpha} + \widehat{\beta_1} x_1^i + \cdots + \widehat{\beta_n} x_n^i + \hat{\delta} + \widehat{\gamma_1} x_1^i + \cdots + \widehat{\gamma_n} x_n^i\right)\right)}$$
$$- \frac{1}{1 + \exp\left(-\left(\hat{\alpha} + \widehat{\beta_1} x_1^i + \cdots + \widehat{\beta_n} x_n^i\right)\right)}$$
$$= \frac{1}{1 + \exp\left(-\left(\left(\hat{\alpha} + \hat{\delta}\right) + \left(\widehat{\beta_1} + \widehat{\gamma_1}\right) x_1^i + \cdots + \left(\widehat{\beta_n} + \widehat{\gamma_n}\right) x_n^i\right)\right)}$$
$$- \frac{1}{1 + \exp\left(-\left(\hat{\alpha} + \widehat{\beta_1} x_1^i + \cdots + \widehat{\beta_n} x_n^i\right)\right)}$$

Como se expone en Lo (2002), la principal ventaja que presenta este enfoque respecto al enfoque de dos modelos es que, al ser estimada la respuesta incremental para las dos subpoblaciones mediante un único modelo, es un enfoque *ad hoc* para predecir la respuesta incremental, y no para estimar las respuestas positivas en ambas subpoblaciones por separado (dos modelos). Además, las dos probabilidades (*scores*) están en la misma escala. Como debilidades, en Kane *et al* (2014) se destaca que igualmente existe acumulación de errores al restar dos estimaciones de probabilidades y que puede existir multicolinealidad al utilizar las mismas variables explicativas en la población de control y en la población de tratamiento (a través de las de interacción), lo que puede resultar en un sobreajuste de la estimación a los datos de entrenamiento y en inestabilidad en las predicciones.

Este enfoque es el utilizado, de acuerdo a lo expuesto en Porter (2013), en la campaña presidencial de Obama de 2012 a la que se hace referencia en la introducción de este trabajo.

2.3. Transformación del *target*

El resto de técnicas de modelización indirecta del *uplift* se agrupan bajo lo que se denomina una transformación del *target*: definir la variable objetivo a estimar no como la variable de respuesta positiva, sino como una modificación de ésta que permita estimar mejor la respuesta incremental. Destacan en este sentido los enfoques de Lai, Kane, Kane generalizado, Pesimista, y de Jaskowski.

Estos enfoques también son conocidos como modelos de 4 *outputs* por intentar separar las subpoblaciones teóricas de causas seguras, persuasibles, causas perdidas y "no molestar" a partir los datos observados en una campaña diseñada para tal fin: de si los clientes eran tratados o no, y si respondían positivamente o no (CR, CN, TR y TN). Como ya se señalaba en la sección 1.2, estas dos divisiones de la población están relacionadas como sigue:

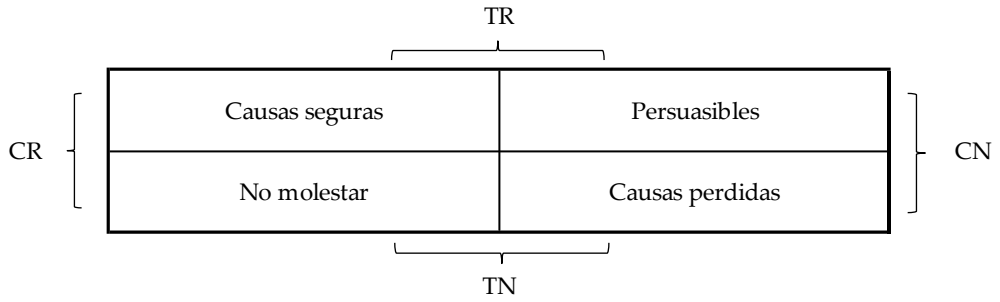


FIGURA 5

La pertenencia a cada uno de los grupos CR, CN, TR o TN servirá para definir la variable objetivo en cada enfoque.

2.3.1. Enfoque de Lai

Bajo el enfoque de Lai, expuesto en Lai (2006), se determina que, si se quiere identificar a los persuasibles para incluirlos en nuevas campañas, y a su vez a los que no se deben molestar[9] para excluirlos de ellas, se debe intentar construir una variable *target* a estimar que separe en la medida de lo posible estos dos grupos. De acuerdo con la relación expuesta en la Figura 5, todos los persuasibles están en CN y TR, que además no contienen clientes "no molestar", por lo que CN y TR definen el valor 1 de la nueva variable *target*, mientras que todos los "no molestar" están en CR y TN, por lo que se marcan con un 0 en esta variable todos los individuos de estos dos grupos. Dentro de CN y TR también hay causas seguras y causas perdidas, clientes que de hecho generan un coste si se contactan y no generan un beneficio adicional al ser contactados, aunque este coste es menor que las pérdidas que generaría un "no molestar" al ser incluido en una nueva población de tratamiento. Por el contrario, CR y TN sólo contienen clientes "no molestar", causas seguras y causas perdidas, por lo que en ningún caso incluir en una nueva campaña individuos de estas subpoblaciones generaría ingresos adicionales, y sí costes.

Así, se define la variable objetivo como:

$$y^i = \begin{cases} 1 & i \in TR \cup CN \\ 0 & i \in TN \cup CR \end{cases}$$

Una vez definida la variable a estimar, se trata simplemente de un problema de clasificación binaria que puede modelizarse con cualquier técnica de modelización para estimar la probabilidad

$$p(y = 1|\boldsymbol{x})$$

[9] En realidad, en el paper de Lai no se menciona la existencia del grupo "no molestar", pero no resulta un inconveniente, ya que en el modelo no se hace un uso explícito de esta clase, y en cualquier caso se modeliza tratando de separar $TR \cup CN$ de $TN \cup CR$, por lo que sí se tiene en cuenta, aunque no de manera intencionada, la existencia de individuos "no molestar".

El ajuste del modelo predictivo permite estimar la probabilidad a posteriori anterior, cuya estimación se denota por:

$$\hat{p}\left(y^i = 1 | x_1^i, \dots, x_n^i\right)$$

La respuesta incremental debida al tratamiento se define como la diferencia de la probabilidad estimada de pertenecer a uno u otro de estos dos grupos ($TR \cup CN$, o $TN \cup CR$):

$$\begin{aligned}\widehat{uplift}_{Lai}(\boldsymbol{x}^i) &= \hat{p}\left(y^i = 1 | x_1^i, \dots, x_n^i\right) - \hat{p}\left(y^i = 0 | x_1^i, \dots, x_n^i\right) \\ &= \hat{p}\left(i \in TR \cup CN | x_1^i, \dots, x_n^i\right) - \hat{p}\left(i \in TN \cup CR | x_1^i, \dots, x_n^i\right)\end{aligned}$$

En comparación con el enfoque de dos modelos, al igual que el enfoque que utiliza una *dummy* de tratamiento, éste no tiene los problemas de escalas distintas derivados de estimar modelos distintos para dos subpoblaciones, y las variables explican directamente la respuesta incremental en lugar de estar centradas en estimar la respuesta positiva. Respecto al modelo con *dummy* de tratamiento, este enfoque reduce los problemas de multicolinealidad por la reducción del número de variables explicativas. Por contra, estos modelos pueden ser inestables (Shaar *et al* (2016)).

2.3.2. Enfoque de Kane

Partiendo del planteamiento de Lai, si el objetivo es separar las poblaciones de persuasibles y "no molestar" a través de los grupos CR, CN, TR y TN, se plantea en Kane *et al* (2014) construir directamente un modelo de clasificación multiclase que estime la probabilidad de pertenecer a cada uno de estos cuatro grupos, por lo que se puede ver como una variación del enfoque de Lai.

Se define entonces la variable objetivo, que no será binaria sino multiclase, como:

$$y^i = \begin{cases} 0 & i \in TR \\ 1 & i \in CN \\ 2 & i \in TN \\ 3 & i \in CR \end{cases}$$

Este problema de clasificación multinomial no ordinal (la elección de los valores 0, 1, 2 y 3 es totalmente arbitraria y no establece ningún orden entre las categorías) se puede modelizar entonces a través de cualquier técnica que permita este tipo de variables de respuesta, como la regresión logística multinomial o árboles de clasificación. Es decir, se estiman conjuntamente las siguientes probabilidades:

$$p(y = 0|\boldsymbol{x}), p(y = 1|\boldsymbol{x}), p(y = 2|\boldsymbol{x}), p(y = 3|\boldsymbol{x})$$

Obteniendo, a partir de un único modelo predictivo, las estimaciones

$$\hat{p}\left(y^i = 0 | x_1^i, \dots, x_n^i\right), \hat{p}\left(y^i = 1 | x_1^i, \dots, x_n^i\right), \hat{p}\left(y^i = 2 | x_1^i, \dots, x_n^i\right), \hat{p}\left(y^i = 3 | x_1^i, \dots, x_n^i\right)$$

La respuesta incremental debida al tratamiento se define como en el enfoque de Lai, pero teniendo

en cuenta que en este caso ha sido posible separar las probabilidades de pertenecer a los grupos TR o CN, y a TN o CR. Así, se define como la probabilidad de estar en los grupos en los que hay persuasibles, TR o CN, menos la probabilidad de estar en los grupos en los que hay "no molestar", TN o CR:

$$\begin{aligned}\widehat{uplift}_{Kane}(\boldsymbol{x}^i) &= \hat{p}(y^i = 0|x_1^i, \dots, x_n^i) + \hat{p}(y^i = 1|x_1^i, \dots, x_n^i) - \hat{p}(y^i = 2|x_1^i, \dots, x_n^i) \\ &\quad - \hat{p}(y^i = 3|x_1^i, \dots, x_n^i) \\ &= \hat{p}(i \in TR|x_1^i, \dots, x_n^i) + \hat{p}(i \in CN|x_1^i, \dots, x_n^i) - \hat{p}(i \in TN|x_1^i, \dots, x_n^i) \\ &\quad - \hat{p}(i \in CR|x_1^i, \dots, x_n^i)\end{aligned}$$

En comparación con enfoques previos, aparte de presentar las mismas ventajas que el enfoque de Lai, en algunos escenarios puede ser un enfoque más preciso que éste por ser capaz de separar además entre las poblaciones TR y CN, y TN y CR. Intuitivamente, puede haber un conjunto de datos en el que haya características que sirvan para distinguir TR y CN, modelizados de manera separada, de TN y CR, pero que no sean útiles si la modelización se hace de manera conjunta; mientras que también es plausible que en otro conjunto de datos sea más fácil separar $TR \cup CN$ de $TN \cup CR$ si se modelizan bajo el enfoque de Lai. En cambio, este enfoque presenta un problema adicional de sesgo por desbalanceo entre las clases TR, CN, TN y CR, que, en general, no tendrán tamaños similares. De hecho, en el apéndice de Kane *et al* (2014) se demuestra que este enfoque es correcto sólo si las poblaciones de tratamiento y control son del mismo tamaño.

La problemática de las clases desequilibradas también se puede dar bajo el enfoque de Lai, aunque en Lai (2002) se justifica asegurando que el problema se minora debido a que en general en las campañas de marketing directo el tamaño de TR es mucho menor que el de TN, y que el tamaño de CR es mucho menor que el de CN, por lo que al unir un grupo pequeño y uno grande en ambos casos, se estaría mitigando el problema.

2.3.3. Enfoque de Kane generalizado

En Kane *et al* (2014) se destaca que la principal debilidad del enfoque de Kane es el desbalanceo de clases, pero en ese mismo trabajo se propone una solución: el enfoque de Kane generalizado.

Puesto que se quiere corregir el sesgo introducido por el desbalanceo entre el grupo de tratamiento y el de control, se corrigen cada uno de los términos que conforman la expresión de la respuesta incremental con los pesos adecuados: la probabilidad de pertenencia de un individuo al grupo de tratamiento para las expresiones en las que éste interviene, $\hat{p}(i \in TR|x_1^i, \dots, x_n^i)$ y $\hat{p}(i \in TN|x_1^i, \dots, x_n^i)$, y la probabilidad de pertenecer al grupo de control para las estimaciones afectadas por dicho grupo, $\hat{p}(i \in CR|x_1^i, \dots, x_n^i)$ y $\hat{p}(i \in CN|x_1^i, \dots, x_n^i)$. Así, la respuesta incremental se obtendría como:

$$\widehat{uplift}_{Kane\ generalizado}(\boldsymbol{x}^i)$$
$$= \frac{\hat{p}\left(i \in TR|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in T|x_1^i, \dots, x_n^i\right)} - \frac{\hat{p}\left(i \in TN|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in T|x_1^i, \dots, x_n^i\right)} + \frac{\hat{p}\left(i \in CR|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in C|x_1^i, \dots, x_n^i\right)}$$
$$- \frac{\hat{p}\left(i \in CN|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in C|x_1^i, \dots, x_n^i\right)}$$
$$= \frac{\hat{p}\left(y^i = 0|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in T|x_1^i, \dots, x_n^i\right)} - \frac{\hat{p}\left(y^i = 2|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in T|x_1^i, \dots, x_n^i\right)} + \frac{\hat{p}\left(y^i = 1|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in C|x_1^i, \dots, x_n^i\right)}$$
$$- \frac{\hat{p}\left(y^i = 3|x_1^i, \dots, x_n^i\right)}{\hat{p}\left(i \in C|x_1^i, \dots, x_n^i\right)}$$
$$= \frac{1}{\hat{p}\left(i \in T|x_1^i, \dots, x_n^i\right)}\left[\hat{p}\left(i \in TR|x_1^i, \dots, x_n^i\right) - \hat{p}\left(i \in TN|x_1^i, \dots, x_n^i\right)\right]$$
$$+ \frac{1}{\hat{p}\left(i \in C|x_1^i, \dots, x_n^i\right)}\left[\hat{p}\left(i \in CN|x_1^i, \dots, x_n^i\right) - \hat{p}\left(i \in CR|x_1^i, \dots, x_n^i\right)\right]$$

Además, en el escenario ideal en el que la marcación de la campaña hubiera sido aleatoria, se tendrá que la probabilidad de que un individuo pertenezca al grupo de tratamiento o al de control no depende de sus características, por lo que $p\left(i \in T|x_1^i, \dots, x_n^i\right) = p(i \in T)$ y $p\left(i \in C|x_1^i, \dots, x_n^i\right) = p(i \in C)$, y la mejor estimación posible de estas dos probabilidades será simplemente la proporción de estos grupos en la muestra, que se puede denotar como $\hat{p}(T)$ y $\hat{p}(C)$, quedando la expresión anterior como:

$$\widehat{uplift}_{Kane\ generalizado}(\boldsymbol{x}^i)$$
$$= \frac{1}{\hat{p}(T)}\left[\hat{p}\left(i \in TR|x_1^i, \dots, x_n^i\right) - \hat{p}\left(i \in TN|x_1^i, \dots, x_n^i\right)\right]$$
$$+ \frac{1}{\hat{p}(C)}\left[\hat{p}\left(i \in CN|x_1^i, \dots, x_n^i\right) - \hat{p}\left(i \in CR|x_1^i, \dots, x_n^i\right)\right]$$

Como se muestra a continuación, esta construcción tiene además un respaldo teórico.

Si se parte de una expresión genérica de la respuesta incremental, similar por ejemplo a la utilizada en el enfoque de Lo, se tiene que

$$uplift(\boldsymbol{x}) = p(y = 1|t = 1, \boldsymbol{x}) - p(y = 1|t = 0, \boldsymbol{x}) = p(R|T, \boldsymbol{x}) - p(R|C, \boldsymbol{x})$$

Haciendo uso de que R y N son complementarios, se tiene que la respuesta incremental se puede expresar como

$$uplift(\boldsymbol{x}) = p(R|T, \boldsymbol{x}) - \left(1 - p(N|C, \boldsymbol{x})\right)$$

o bien como

$$uplift(\boldsymbol{x}) = \left(1 - p(N|T, \boldsymbol{x})\right) - p(R|C, \boldsymbol{x})$$

Aplicando la definición de probabilidad condicionada, se tiene que

$$uplift(\boldsymbol{x}) = \frac{p(R \cap T, \boldsymbol{x})}{p(T, \boldsymbol{x})} - \left(1 - \frac{p(N \cap C, \boldsymbol{x})}{p(C, \boldsymbol{x})}\right)$$

$$uplift(\boldsymbol{x}) = \left(1 - \frac{p(N \cap T, \boldsymbol{x})}{p(T, \boldsymbol{x})}\right) - \frac{p(R \cap C, \boldsymbol{x})}{p(C, \boldsymbol{x})}$$

Pero precisamente $R \cap T$, $N \cap C$, $N \cap T$ y $R \cap C$ son las poblaciones que se han definido como TR, CN, TN y CR respectivamente. Por lo que reemplazando las expresiones y reordenando quedaría:

$$uplift(\boldsymbol{x}) = \frac{p(TR, \boldsymbol{x})}{p(T, \boldsymbol{x})} + \frac{p(CN, \boldsymbol{x})}{p(C, \boldsymbol{x})} - 1$$

$$uplift(\boldsymbol{x}) = 1 - \frac{p(TN, \boldsymbol{x})}{p(T, \boldsymbol{x})} - \frac{p(CR, \boldsymbol{x})}{p(C, \boldsymbol{x})}$$

donde se observa además que los dos primeros términos contribuyen positivamente a la respuesta incremental, y los dos segundos contribuyen negativamente. Si se suman estas dos igualdades, se obtiene la expresión de la respuesta incremental que se ha dado para el enfoque de Kane generalizado:

$$2 \cdot uplift(\boldsymbol{x}) = \frac{p(TR, \boldsymbol{x})}{p(T, \boldsymbol{x})} + \frac{p(CN, \boldsymbol{x})}{p(C, \boldsymbol{x})} - \frac{p(TN, \boldsymbol{x})}{p(T, \boldsymbol{x})} - \frac{p(CR, \boldsymbol{x})}{p(C, \boldsymbol{x})}$$

Por último, si bien este enfoque corrige el sesgo por desbalanceo entre tratamiento y control que había en el enfoque anterior, sigue pudiendo presentar resultados inestables, como se expone en las conclusiones de Kane *et al* (2014).

Es necesario resaltar también que, al igual que en el enfoque de Kane, no se trata de un problema de clasificación binaria, que es el principal objeto de estudio de este trabajo. Un posible enfoque que haga uso de modelos de clasificación binario podría ser construir cuatro modelos que estimen la probabilidad de pertenencia a cada uno de los grupos, que sí que utilizaría variables objetivo binarias, aunque añadiría problemas a los ya observados: habría que ajustar las probabilidades estimadas para que siempre sumen 1, los modelos resultantes podrían tener escalas distintas, y se acumularían los errores de los cuatro modelos.

2.3.4. Enfoque Pesimista

El enfoque pesimista surge en Shaar *et al* (2016) con el objetivo de reducir la inestabilidad que puede aparecer en el enfoque de Lai. Para ello, se define la respuesta incremental como la media entre dos enfoques: una modificación del enfoque de Lai, y el enfoque reflexivo.

Por un lado, el enfoque de Lai ponderado intenta corregir el sesgo por desbalanceo de clases de manear similar al enfoque de Kane generalizado. En este caso no se pueden ajustar las probabilidades con las proporciones de la población en los grupos de control y tratamiento,

porque no intervienen aisladamente en la expresión, sino que se debe hacer utilizando las proporciones de las poblaciones $TR \cup CN$ y $TN \cup CR$, que sí que intervienen directamente en los términos de la expresión. Así, la respuesta incremental se calcularía como

$$\begin{aligned}\widehat{uplift}_{Lai\ ponderado}(\boldsymbol{x}^i) &= \hat{p}\left(y^i = 1|x_1^i, \dots, x_n^i\right) \cdot \hat{p}(y^i = 1) - \hat{p}\left(y^i = 0|x_1^i, \dots, x_n^i\right) \cdot \hat{p}(y^i = 0) \\ &= \hat{p}\left(i \in TR \cup CN|x_1^i, \dots, x_n^i\right) \cdot \hat{p}(i \in TR \cup CN) - \hat{p}\left(i \in TN \cup CR|x_1^i, \dots, x_n^i\right) \\ &\cdot \hat{p}(i \in TN \cup CR)\end{aligned}$$

donde $\hat{p}(i \in TR \cup CN)$ es simplemente la proporción de individuos tratados que responden positivamente o no tratados que no responden (observaciones en TR o CN) en la muestra, y $\hat{p}(i \in TN \cup CR) = 1 - \hat{p}(i \in TR \cup CN)$ es la proporción en la muestra de individuos tratados que no responden positivamente o no tratados que sí lo hacen.

Este enfoque se va a ponderar con otro enfoque con el propósito de aportarle estabilidad: el enfoque reflexivo. En este caso, se construye una estimación para la respuesta incremental similar a la del enfoque de Lai, es decir, se buscan expresiones del tipo $\hat{p}_{Reflexivo}\left(i \in TR \cup CN|x_1^i, \dots, x_n^i\right)$ y $\hat{p}_{Reflexivo}\left(i \in TN \cup CR|x_1^i, \dots, x_n^i\right)$, pero utilizando dos modelos auxiliares para R y N (separando la muestra entre estas dos subpoblaciones y generando dos modelos) que modelizan la probabilidad de que un individuo haya sido tratado dado que ha respondido positivamente, y la probabilidad de que un individuo haya sido tratado dado que no ha respondido positivamente:

$$\hat{p}_{Reflexivo}\left(i \in TR \cup CN|x_1^i, \dots, x_n^i\right) = \hat{p}_R(i \in T|i \in R) \cdot \hat{p}(i \in TR) + \hat{p}_N(i \in C|i \in N) \cdot \hat{p}(i \in CN)$$

$$\hat{p}_{Reflexivo}\left(i \in TN \cup CR|x_1^i, \dots, x_n^i\right) = \hat{p}_N(i \in T|i \in N) \cdot \hat{p}(i \in TN) + \hat{p}_R(i \in C|i \in R) \cdot \hat{p}(i \in CR)$$

donde $\hat{p}(i \in TR)$, $\hat{p}(i \in CN)$, $\hat{p}(i \in TN)$ y $\hat{p}(i \in CR)$ son las proporciones de dichas subpoblaciones (TR, CN, TN y CR respectivamente) en la muestra, y $\hat{p}_R$ y $\hat{p}_N$ son, respectivamente, los modelos individuales de predicción de pertenencia a la población de tratamiento (y a su complementaria, a la de control) para las subpoblaciones de los que responden positivamente y los que no (R y N).

Una vez se obtienen estas dos estimaciones, es posible estimar la respuesta incremental reflexiva de manera similar al enfoque de Lai:

$$\widehat{uplift}_{Reflexivo}(\boldsymbol{x}^i) = \hat{p}_{Reflexivo}\left(i \in TR \cup CN|x_1^i, \dots, x_n^i\right) - \hat{p}_{Reflexivo}\left(i \in TN \cup CR|x_1^i, \dots, x_n^i\right)$$

Se define por tanto la respuesta incremental pesimista como la media de las respuestas incrementales obtenidas mediante estos dos enfoques:

$$\widehat{uplift}_{Pesimista}(\boldsymbol{x}^i) = \frac{1}{2} \cdot \widehat{uplift}_{Lai\ ponderado}(\boldsymbol{x}^i) + \frac{1}{2} \cdot \widehat{uplift}_{Reflexivo}(\boldsymbol{x}^i)$$

Así, de acuerdo a los resultados obtenidos en Shaar *et al* (2016), se obtiene una estimación de la respuesta incremental que es más precisa y robusta que la proporcionada por el enfoque de

Lai. Por último, nótese que tanto la probabilidad de Lai generalizada como la probabilidad reflexiva son nuevamente problemas de clasificación binaria que pueden modelizarse con cualquier técnica habitual, al contrario de lo que sucedía en los enfoques de Kane.

2.3.5. Enfoque de Jaskowski

Este enfoque, propuesto en Jaskowski y Jaroszewicz (2012), parte de una construcción de la variable *target* similar a la del enfoque de Lai: tomará el valor 1 para individuos en TR o en CN, donde estaban los persuasibles, y 0 en caso contrario, pues en $TR \cup CN$ se identifican, aparte de los respondieron positivamente al tratamiento previo (TR), los que no hubieran empeorado su respuesta si hubieran sido incluidos en el grupo de tratamiento en lugar de en el de control (CN).

$$y^i = \begin{cases} 1 & i \in TR \cup CN \\ 0 & i \in TN \cup CR \end{cases}$$

En cambio, se parte de la concepción inicial en la que se mide la diferencia en la respuesta positiva entre el grupo de tratamiento y de control, por lo que se quiere modelizar

$$p\left(r^i = 1 | t^i = 1, x_1^i, \dots, x_n^i\right) - p\left(r^i = 1 | t^i = 0, x_1^i, \dots, x_n^i\right)$$

donde r^i indica la respuesta positiva o no del individuo, pero con el objetivo de hacerlo a través de la variable *target* que se acaba de definir.

Para obtener la respuesta incremental mediante este enfoque, en primer lugar, se descompone la probabilidad de que esta nueva variable *target* tome el valor 1 (mediante el teorema de probabilidad total) y se utiliza la definición de la *target* para obtener una expresión en función de la variable de respuesta positiva, r:

$$\begin{aligned} &p\left(y^i = 1 | x_1^i, \dots, x_n^i\right) \\ &\quad = p\left(y^i = 1 | t^i = 1, x_1^i, \dots, x_n^i\right) \cdot p\left(t^i = 1 | x_1^i, \dots, x_n^i\right) + p\left(y^i = 1 | t^i = 0, x_1^i, \dots, x_n^i\right) \\ &\quad \cdot p\left(t^i = 0 | x_1^i, \dots, x_n^i\right) \\ &\quad = p\left(r^i = 1 | t^i = 1, x_1^i, \dots, x_n^i\right) \cdot p\left(t^i = 1 | x_1^i, \dots, x_n^i\right) + p\left(r^i = 0 | t^i = 0, x_1^i, \dots, x_n^i\right) \\ &\quad \cdot p\left(t^i = 0 | x_1^i, \dots, x_n^i\right) \end{aligned}$$

Asumiendo que la marcación de tratamiento/control de la campaña previa fue realizada de manera aleatoria, se tiene que $p\left(t^i = 1 | x_1^i, \dots, x_n^i\right) = p(t^i = 1)$ y $p\left(t^i = 0 | x_1^i, \dots, x_n^i\right) = p(t^i = 0)$, por ser estas probabilidades independientes de las características del individuo, obteniendo

$$\begin{aligned} &p\left(y^i = 1 | x_1^i, \dots, x_n^i\right) \\ &\quad = p\left(r^i = 1 | t^i = 1, x_1^i, \dots, x_n^i\right) \cdot p(t^i = 1) + p\left(r^i = 0 | t^i = 0, x_1^i, \dots, x_n^i\right) \cdot p(t^i = 0) \end{aligned}$$

Si además se asume que los grupos de tratamiento y control están balanceados (si no fuera así, habría que abordar el problema del desequilibrio de las clases utilizando alguna propuesta estándar para tal fin), es decir, que $p(t^i = 1) = p(t^i = 0) = 1/2$, se tiene que

$$p(y^i = 1|x_1^i, \dots, x_n^i) = p(r^i = 1|t^i = 1, x_1^i, \dots, x_n^i) \cdot \frac{1}{2} + p(r^i = 0|t^i = 0, x_1^i, \dots, x_n^i) \cdot \frac{1}{2}$$
$$\rightarrow 2 \cdot p(y^i = 1|x_1^i, \dots, x_n^i)$$
$$= p(r^i = 1|t^i = 1, x_1^i, \dots, x_n^i) + p(r^i = 0|t^i = 0, x_1^i, \dots, x_n^i)$$
$$= p(r^i = 1|t^i = 1, x_1^i, \dots, x_n^i) + \left(1 - p(r^i = 1|t^i = 0, x_1^i, \dots, x_n^i)\right)$$

Por lo que es posible relacionar la probabilidad de que la nueva *target* tome el valor 1, $p(y^i = 1|x_1^i, \dots, x_n^i)$, con la expresión inicial que se buscaba y que describe la respuesta incremental, $p(r^i = 1|t^i = 1, x_1^i, \dots, x_n^i) - p(r^i = 1|t^i = 0, x_1^i, \dots, x_n^i)$, mediante

$$2 \cdot p(y^i = 1|x_1^i, \dots, x_n^i) - 1 = p(r^i = 1|t^i = 1, x_1^i, \dots, x_n^i) - p(r^i = 1|t^i = 0, x_1^i, \dots, x_n^i)$$

Así, la respuesta incremental se estimaría por:

$$\widehat{uplift}_{Jaskowski}(\boldsymbol{x}^i) = 2 \cdot \hat{p}(y^i = 1|x_1^i, \dots, x_n^i) - 1$$

Por lo que este enfoque reduciría el problema a modelizar a uno de clasificación binaria en el que hay que estimar $p(y^i = 1|x_1^i, \dots, x_n^i)$.

Cabe mencionar que existe una generalización de este enfoque para el caso de clases no balanceadas que se puede consultar en Athey e Imbens (2015).

2.4. Estimación directa

Al contrario que los enfoques anteriores, el enfoque por estimación directa cubre los casos en los que se intenta estimar directamente la respuesta incremental. Una idea intuitiva sería por ejemplo intentar separar en la población total aquellas subpoblaciones en las que los ratios de respuesta incremental sean significativamente distintos, definiendo estos ratios de respuesta incremental para esas nuevas subpoblaciones como el porcentaje de respuesta positiva en los individuos del grupo de tratamiento menos el mismo porcentaje en el grupo de control. Básicamente, las técnicas aquí incluidas modificarán algoritmos existentes para modelizar directamente la respuesta incremental, por ejemplo utilizando métricas objetivo relacionadas con diferencias en el ratio que se acaba de definir entre nuevas subpoblaciones y tamaños muestrales mínimos en dichas subpoblaciones.

Los métodos de estimación directa se pueden agrupar en:

- Enfoques basados en árboles de decisión a partir de algoritmos conocidos (CART, C4.5, CHAID)[10] en los que se adaptan los criterios de división de los nodos o de poda del árbol para poder estimar la respuesta incremental directamente. En esta categoría se incluyen los enfoques de Hansotia y Rukstales, o de Rzepakowski y Jaroszewicz.

[10] Por ejemplo, se puede consultar el detalle de los algoritmos CART y C4.5 en Hastie *et al* (2001) y de CHAID en Kass (1980).

- Enfoques de ensamblado que, mediante técnicas habituales de ensamblado[11], combinan otros modelos de respuesta incremental (principalmente, árboles) para reducir la inestabilidad o el sobreajuste de estos modelos. Se encontrarían aquí los enfoques *bagging* (de empaquetado), *random forest* (bosques aleatorios) y *causal conditional inference forest* (ensamblado mediante árboles de inferencia condicional).

De acuerdo con lo expuesto en Radcliffe y Surry (2011), en general los enfoques presentados en las secciones previas son más inestables que los enfoques de estimación directa.

2.4.1. Enfoques basados en árboles

2.4.1.1. Enfoque de Hansotia y Rukstales

El enfoque propuesto en Hansotia y Rukstales (2002) es una modificación del algoritmo CHAID en la que la población total se va subdividiendo de forma recursiva en dos nodos descendientes descritos por cortes en una variable explicativa. En cada nodo, se obtienen los descendientes utilizando un corte en una variable continua o agrupaciones de una variable categórica, y para determinar el corte de subdivisión de cada nodo se maximiza una métrica objetivo. La modificación respecto al CHAID es precisamente la métrica objetivo que se utiliza.

Así, la población de cada nodo se divide maximizando la diferencia de la respuesta incremental entre los nodos hijos: en concreto, la tasa de respuesta en el grupo de tratamiento que hay en ese nodo menos la tasa de respuesta en el grupo control de dicho nodo. La métrica que se maximiza sería:

$$\Delta(\Delta p) = \Delta p_I - \Delta p_D = (p_I^T - p_I^C) - (p_D^T - p_D^C)$$

donde p es la tasa de respuesta observada, los subíndices I y D se refieren a los nodos hijos (izquierda y derecha) y los superíndices T y C se refieren a las poblaciones de tratamiento y control contenidas en esos nodos. Utilizando la notación previa, y viendo las probabilidades como tasas observadas, en cada corte se estaría maximizando la métrica:

$$\Delta(uplift) = uplift_I - uplift_D = \big(p_I(R|T) - p_I(R|C)\big) - \big(p_D(R|T) - p_D(R|C)\big)$$

Intuitivamente, en los primeros cortes se van a ir obteniendo nodos hijos en los que la presencia de individuos persuasibles e individuos "no molestar" va a ser muy desigual, y en cortes posteriores se van a separar a estos de las causas seguras y las causas perdidas. Para facilitar el entendimiento de los árboles, se puede ilustrar con un ejemplo ficticio y su Figura 6 en el que:

- En primer lugar, se dispone de una población en la que la tasa de respuesta positiva en el grupo de tratamiento es del 0.1, mientras que en el grupo de control es del 0.04. La respuesta incremental en este nodo sería por tanto del 0.06.

[11] Se puede encontrar más información sobre técnicas de ensamblado en James *et al* (2013).

- En la primera división, en el nodo 2 (izquierdo) se observa una respuesta positiva del 0.14 y 0.056 en tratamiento y control, por lo que la respuesta incremental sería del 0.084, y en el nodo 3 (derecho) del 0.06 y 0.024 en tratamiento y control, con una respuesta incremental del 0.036. El valor de $\Delta(uplift)$ sería de 0.048 para el corte del nodo raíz.

- El nodo 2 se divide en el nodo 4 (izquierdo), donde se observa una respuesta positiva del 0.154 y 0.062 en tratamiento y control, por lo que la respuesta incremental sería del 0.092, y en el nodo 5 (derecho), con respuesta positiva del 0.119 y 0.048 en tratamiento y control, con una respuesta incremental del 0.071. $\Delta(uplift)$ alcanza el valor 0.021 en esta división.

- El nodo 3 se divide en el nodo 6 (izquierdo), donde se observa una respuesta positiva del 0.068 y 0.026 en tratamiento y control, por lo que la respuesta incremental sería del 0.042, y en el nodo 7 (derecho), con respuesta positiva del 0.04 y 0.052 en tratamiento y control, con una respuesta incremental del -0.012. La métrica $\Delta(uplift)$ toma el valor 0.054 en este corte.

- En resumen, se pasa de un nodo inicial en el que la respuesta incremental era del 0.06, a cuatro nodos finales en los que las respuestas incrementales son del 0.092, 0.071, 0.042 y -0.012.

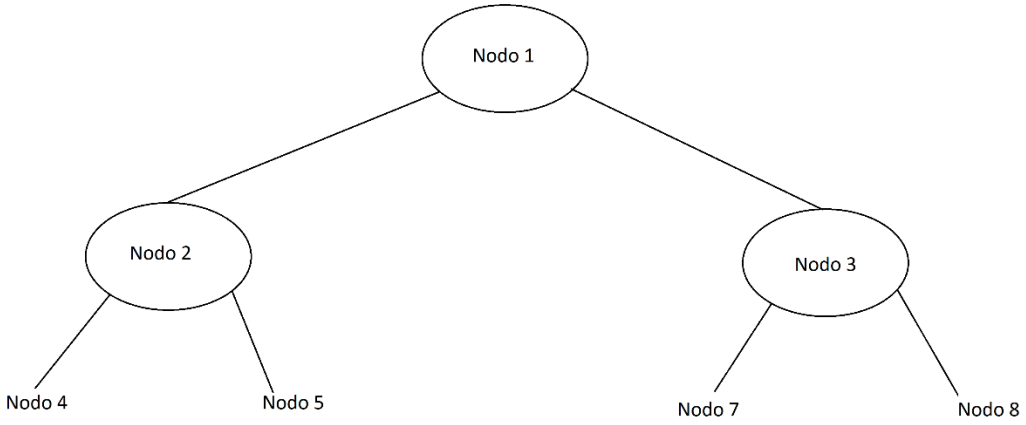


FIGURA 6

Los criterios de parada de este algoritmo son similares a los del CHAID: se utiliza la significación estadística, profundidad máxima y/o población mínima por nodo.

Más allá de poder establecer una población mínima en cada nodo como criterio de parada, la partición de cada nodo no tiene en cuenta los tamaños de los nodos resultantes, lo que puede sesgar el algoritmo hacia la separación de pequeñas subpoblaciones y acabar resultado en un modelo sobreajustado (Radcliffe y Surry (2011)).

2.4.1.2. Enfoque de Rzepakowski y Jaroszewicz

En Rzepakowski y Jaroszewicz (2012) se desarrolla un enfoque basado en la teoría de la información, similar al algoritmo C4.5 o CART cuando se usa la entropía para dividir cada nodo.

Este enfoque utiliza una métrica de divergencia, D, entre las distribuciones de la variable respuesta en los grupos de tratamiento y control como criterio de división de cada nodo entre sus nodos hijos. En primer lugar, se busca que la distribución de la variable respuesta difiera (o diverja) lo máximo posible entre grupos de control y tratamiento, pues precisamente así se captura la respuesta incremental. Posteriormente, para llevar a cabo la división de cada nodo, se maximiza la ganancia de divergencia (o cantidad de información), $D_{ganancia}$, al pasar del nodo padre a los nodos hijos (generados por la división A del nodo padre):

$$D_{ganancia}(A) = D\big(p(R|T,A):p(R|C,A)\big) - D\big(p(R|T):p(R|C)\big)$$

donde $D(p(R|T):p(R|C))$ es la divergencia en el nodo padre, y $D(p(R|T,A):p(R|C,A))$ es la media ponderada (por tamaño de los nodos) de las divergencias en los nodos hijos, a, que además no son necesariamente dos:

$$D\big(p(R|T,A):p(R|C,A)\big) = \sum_a \frac{n_a}{n} D\big(p(R|T,a):p(R|C,a)\big)$$

donde a es cada uno de los nodos hijos, n_a el número de observaciones en cada nodo hijo, y n el número de observaciones en el nodo padre.

Respecto a la métrica utilizada para medir la divergencia entre las distribuciones, se proponen tres: la divergencia KL (Kullback-Leibler), la distancia euclídea o la divergencia chi-cuadrado. Si R_t y R_c son las distribuciones de la variable de respuesta en los grupos de tratamiento y control de cualquier nodo, se definen estas tres métricas de divergencia como:

$$KL(R_t:R_c) = \sum_{i=0,1} r_{t_i} \log \frac{r_{t_i}}{r_{c_i}}$$

$$E(R_t:R_c) = \sum_{i=0,1} \left(r_{t_i} - r_{c_i}\right)^2$$

$$\chi^2(R_t:R_c) = \sum_{i=0,1} \frac{\left(r_{t_i} - r_{c_i}\right)^2}{r_{c_i}}$$

donde r_{t_i} y r_{c_i} son, respectivamente, la proporción de individuos que respondieron positivamente ($i = 1$) e individuos que no respondieron positivamente ($i = 0$) en los grupos de tratamiento y control.

El siguiente ejemplo sirve para ilustrar las métricas de divergencia y el funcionamiento del algoritmo.

- Inicialmente, se dispone de una población en la que la tasa de respuesta positiva en el grupo de tratamiento es de 0.1, mientras que en el grupo de control es de 0.04. La respuesta incremental en este nodo sería por tanto 0.06. Si, por ejemplo, se utiliza la divergencia KL:

$$D\big(p(R|T):p(R|C)\big) = \sum_{i=0,1} r_{t_i} \log \frac{r_{t_i}}{r_{c_i}} = 0.9 \log \frac{0.9}{0.96} + 0.1 \log \frac{0.1}{0.04} = 0.033544$$

- Idealmente, se ha podido encontrar una variable que genera una división, A, de la población total (nodo 1) en los siguientes tres nodos hijos, en los que claramente se ha conseguido separar a bastantes individuos persuasibles (nodo a_1, con un 30% de la población) y "no molestar" (nodo a_3, con un 50% de la población) del resto (nodo a_2, con un 20% de la población). En el nodo a_1, la tasa de respuesta positiva en tratamiento y control es de 0.2 y 0.02 respectivamente, en el nodo a_2 del 0.04 y 0.04, y en el nodo a_3 de 0.02 y de 0.1. Las respuestas incrementales para estos nodos serían 0.18, 0 y -0.08 respectivamente. La divergencia KL en estos tres nodos sería:

$$D\big(p(R|T,a_1):p(R|C,a_1)\big) = \sum_{i=0,1} r_{t_i} \log \frac{r_{t_i}}{r_{c_i}} = 0.8 \log \frac{0.8}{0.98} + 0.2 \log \frac{0.2}{0.02} = 0.298164$$

$$D\big(p(R|T,a_2):p(R|C,a_2)\big) = \sum_{i=0,1} r_{t_i} \log \frac{r_{t_i}}{r_{c_i}} = 0.96 \log \frac{0.96}{0.96} + 0.04 \log \frac{0.04}{0.04} = 0$$

$$D\big(p(R|T,a_3):p(R|C,a_3)\big) = \sum_{i=0,1} r_{t_i} \log \frac{r_{t_i}}{r_{c_i}} = 0.98 \log \frac{0.98}{0.9} + 0.02 \log \frac{0.02}{0.1} = 0.051266$$

Por lo que divergencia de esta partición sería:

$$D\big(p(R|T,A):p(R|C,A)\big) = 0.3 \cdot 0.298164 + 0.5 \cdot 0 + 0.2 \cdot 0.051266 = 0.099702$$

Y así la ganancia en divergencia de la partición sería:

$$D_{ganancia}(A) = D\big(p(R|T,A):p(R|C,A)\big) - D\big(p(R|T):p(R|C)\big)$$

$$= 0.099702 - 0.033544 = 0.066158$$

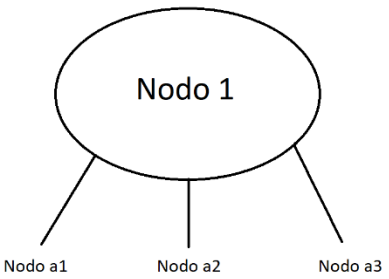


FIGURA 7

En Rzepakowski y Jaroszewicz (2012) utilizan la poda en lugar de un criterio de parada: dividen la

población total en una muestra de entrenamiento, donde construyen el árbol, y una de validación; desde los nodos terminales, hacia arriba, se va evaluando si sustituir el subárbol que cuelga de ese nodo por un único nodo (es decir, si podar esa rama) mejoraría el desempeño en la muestra de validación, evaluándolo con una métrica basada en las diferencias en las probabilidades de respuesta positiva en tratamiento y control para el nodo que se evalúa y sus hojas (si las diferencias entre el nodo y las hojas no son significativas, se poda la rama).

Este enfoque presenta como novedad admitir cortes multivía, pues no está restringido a cortes mediante reglas binarias. Además, al utilizar una media ponderada en el cálculo de la ganancia de divergencia, tiene en cuenta que los nodos hijos no están balanceados. Los autores señalan que el enfoque presenta un mejor comportamiento (en el capítulo siguiente se introducen las métricas de evaluación de estos modelos) que otros enfoques como el de dos modelos o el de Hansotia y Rukstales. Por otro lado, en Radcliffe y Surry (2011) ponen en cuestión diversas asunciones teóricas que hay detrás de la construcción del algoritmo.

2.4.1.3. Otros enfoques

En Radcliffe y Surry (2011), se intenta remediar o mitigar el problema de la diferencia de tamaño entre los nodos hijos que observan en el enfoque de Hansotia y Rukstales. Para ello, inicialmente proponen un enfoque que penaliza dicha diferencia, pero encuentran problemas para generalizar el factor de penalización y concluyen que dicha penalización (o, dicho de otro modo, el balance entre la maximización de la métrica objetivo y la minimización de la diferencia de tamaños en los nodos hijos) depende del conjunto de datos sobre el que se esté trabajando.

Finalmente proponen un enfoque similar a los algoritmos CART y C4.5 basado en un contraste de significación estadística para dividir cada nodo, que maximiza la diferencia entre la respuesta incremental de los nodos hijos e, indirectamente, minimiza la diferencia de tamaño entre nodos hijos. El contraste de significación estadística es un test-t que se realiza sobre el término de interacción (del efecto conjunto de pertenecer a la población de tratamiento y a uno de los dos nodos hijos) de una regresión auxiliar que se construye para dividir el nodo utilizando como variables explicativas únicamente la variable t, que indica si una observación pertenece al grupo de tratamiento o control, la variable h, que indica a qué nodo hijo pertenece cada observación, y la interacción entre ambas variables. En cada nodo, se busca la variable y el corte de ésta que haga más significativo el coeficiente de interacción. Además, se utiliza la varianza de la respuesta incremental entre subpoblaciones (obtenidas mediante remuestreos) para podar el árbol en lugar de utilizar un criterio de parada: es decir, se construye un árbol con una muestra, y se mide la varianza de la respuesta incremental de cada nodo en el resto de muestras, y se descarta la rama si supera cierto umbral.

2.4.2. Enfoques de Ensamblado

2.4.2.1. Enfoque *Bagging*

Los enfoques de ensamblado y, en particular, el enfoque *bagging* o de empaquetado, surgen ante un problema frecuente en los enfoques propuestos hasta el momento: reducir la inestabilidad de los métodos de modelización de la respuesta incremental.

La idea detrás de este enfoque está en combinar distintos modelos construidos utilizando el mismo enfoque o distintos enfoques sobre distintas muestras de construcción para así obtener una estimación conjunta (*aggregating*), que en general será la media de las estimaciones dadas por los distintos modelos, y así reducir la inestabilidad que estos modelos pudieran presentar por separado. Por otro lado, ante la dificultad de disponer de distintas muestras para desarrollar dichos modelos, lo habitual es generarlas a partir remuestreos de la muestra original (*bootstrap*).

Se trata de una idea similar al uso del *bagging* en modelos de predicción (James *et al* (2013)): reducir la alta varianza que presentan algunas técnicas de modelización, especialmente los árboles, donde es frecuente obtener modelos y predicciones distintas si se desarrollan sobre distintas muestras (especialmente si no se aplica un procedimiento de poda).

Algunas de las propuestas que destacan y que se pueden clasificar dentro de este enfoque son:

- En Radcliffe y Surry (2011) se propone un enfoque *bagging* sobre los árboles de significación estadística que introducen en ese mismo trabajo (expuesto en la sección 2.4.1.2) para reducir la inestabilidad que observan al desarrollar modelos con ellos sobre distintos conjuntos de datos: utilizan 10 o 20 modelos construidos sobre poblaciones obtenidas mediante remuestreo con repeticiones (cada modelo utiliza 8 muestras procedentes de un remuestreo distinto de la población original, una para construir y 7 para validar), y obtienen las predicciones medias de estos modelos.

- En Soltys *et al* (2014) se extiende el *bagging* a ensamblados entre cualquier enfoque de estimación directa; en particular, analizan en detalle el enfoque de árboles de divergencia euclídea (sección 2.4.1.2, propuesto por dos de estos tres autores). Los modelos que obtienen mejoran los enfoques de dos modelos y los de modelización directa basados en árboles (un solo árbol).

- En Rößler *el al* (2021) se construyen ensamblados ponderados que utilizan las mismas técnicas de agregación, pero ensamblando otros enfoques (dos modelos, transformación del *target* de Jaskowski) junto a los enfoques de estimación directa basados en árboles (árboles de divergencia euclídea). De hecho, también utilizan árboles como técnica de modelización en el enfoque de dos modelos y en el de trasformación del *target*. La ponderación de los distintos enfoques en la predicción se hace utilizando las métricas de evaluación que se introducen en el capítulo siguiente.

En cualquier caso, el uso de este enfoque no debería limitarse sólo a estos casos, sino que es razonable planteárselo cuando se observa inestabilidad o sobreajuste en un modelo

construido utilizando determinada técnica.

2.4.2.2. Enfoque *Random Forest*

Al igual que en el caso de las técnicas de *bagging*, se puede extender la noción de *random forest* de los problemas de predicción a los de respuesta incremental agregando árboles adaptados para estimar directamente el *uplift*. La novedad respecto a las técnicas de *bagging* es que limitan también el número de variables explicativas que se evalúan en los nodos de cada uno de los árboles que se construyen.

En Guelman *et al* (2012), los autores extienden en efecto los bosques aleatorios a árboles de respuesta incremental para reducir la varianza que observan en estos. En particular, utilizan árboles basados en el criterio de divergencia KL (introducido en la sección 2.4.2.2) para la división de cada nodo en dos nodos hijos, aunque para determinar ese corte eligen aleatoriamente un subconjunto de las variables explicativas disponibles: en general, la raíz del número de variables disponibles. Además, cada uno de esos árboles se construye sin limitación de profundidad (sin poda).

Obviamente, este enfoque no es válido únicamente para los árboles de divergencia KL: por ejemplo, en Soltys *et al* (2014) se construyen bosques aleatorios de respuesta incremental para árboles de divergencia euclídea. También utilizan la raíz de las variables disponibles como el número de variables candidatas para determinar el corte en cada nodo de los árboles del ensamblador, pero, en cambio, utilizan como criterios de parada un límite en la profundidad del árbol (limitada a 20) y un número mínimo de observaciones en cada nodo (3 por grupo de tratamiento y control), aunque apenas encuentran diferencias en aplicar estas restricciones.

2.4.2.3. Enfoque de *Causal Conditional Inference Forest*

Por último, en Guelman *et al* (2014) los autores refinan su enfoque previo mediante la introducción de árboles de inferencia condicional.

Este enfoque introduce un nuevo algoritmo para corregir aspectos negativos que observan en los *random forest*: el sobreajuste ocasionado por no utilizar un criterio de parada o poda, y la tendencia o sesgo de este tipo de algoritmos a elegir variables para las que es posible generar muchas divisiones (porque tienen muchas categorías y, por tanto, entran en muchos cortes de nodos). Para resolver estos problemas, separan la selección de variables y el proceso de división de nodos, añadiendo también un criterio de parada basado en la permutación de observaciones.

La base de este enfoque son los árboles de inferencia condicional (de Zeileis *et al* (2006)): en cada nodo terminal, se realiza un contraste de hipótesis nula global entre las variables explicativas y la variable de tratamiento (contrastes de independencia condicional); se para el proceso de división de un nodo si la hipótesis no puede ser rechazada, y en caso contrario se elige la variable con mayor interacción con la variable de tratamiento para generar la división.

El p-valor del contraste se obtiene además a partir de permutaciones en los datos de entrada.

CAPÍTULO 3

MÉTODOS DE EVALUACIÓN

De manera natural, surgen dos cuestiones cuando se construye un modelo estadístico: ¿cómo de "bueno" es? y ¿cómo de "bueno" es respecto a otros modelos que se podrían haber construido? Para responderlas, se hace necesario contar con métodos que permitan evaluar los modelos construidos y, además, compararlos con otros modelos.

En este sentido surgen los métodos de evaluación en los modelos predictivos, donde se miden los errores a nivel observación comparando el valor observado y el valor estimado por el modelo para la variable de interés, y se agregan estos errores en alguna métrica (por ejemplo, el Gini, AUROC o KS para los problemas de modelización de variable respuesta binaria). En general, el modelo construido se evalúa en una muestra distinta a la utilizada para estimar el modelo, para evitar que la estimación de estas métricas esté sesgada por hacerse sobre la muestra en la que se ha entrenado, por lo que se suele separar aleatoriamente[12] la muestra original en una muestra de entrenamiento con el 70% de las observaciones, que se utilizan para estimar el modelo, y una muestra de validación con el 30%[13] de las observaciones, donde se evalúa el modelo obtenido.

Por ejemplo, el coeficiente Gini mide la diferencia acumulada (a lo largo de los distintos valores que estima el modelo) de la distribución de la clase de respuesta positiva frente a la de la clase de respuesta negativa:

$$Gini = 1 - \sum_{i=1}^{n-1} (R_{i+1} - R_i) \cdot (N_{i+1} + N_i)$$

[12] Muestreo estratificado para mantener las tasas de respuesta y de tratamiento/control en ambas muestras.
[13] Algunos autores utilizan una división 80-20%.

donde R_i es el porcentaje acumulado de la distribución de respuesta positiva hasta la puntuación i, N_i es el porcentaje acumulado de la distribución de respuesta negativa hasta la puntuación i, y n es el número de posibles estimaciones diferentes del modelo (es decir, de la variable respuesta) cuando se puntúa una determinada población. Este coeficiente toma valores entre 0%[14] y 100% (o 0 y 1), y se puede interpretar cómo:

- Si vale 0%, el modelo no separa entre las dos clases, ya que las distribuciones de respuesta positiva y negativa se acumulan exactamente igual.
- Si vale 100%, el modelo presentaría una discriminación perfecta, ya que hasta cierta puntuación acumula el 100% de las respuestas positivas y el 0% de las negativas.
- Cuanto más cercano esté el coeficiente al 100%, mejor poder discriminante presenta el modelo.

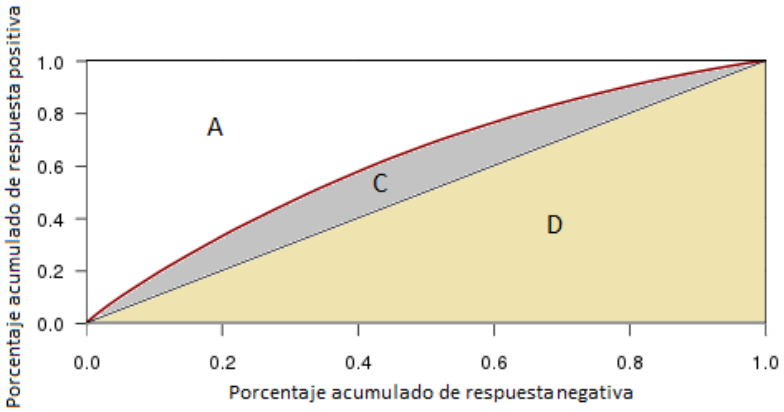


FIGURA 8

La curva ROC (Figura 8), que representa R_i en el eje vertical y N_i en el horizontal, facilita el entendimiento del coeficiente Gini: se puede calcular como el cociente entre las áreas C y $C + A$ (o como C/D), y se interpretaría como una comparativa entre la curva del modelo y la curva ideal (partiendo del origen, la que recorre el eje vertical y luego el horizontal), pues se obtiene como la proporción del área correspondiente a la curva del modelo, C, respecto del área $C + A$, asociada a la curva que discrimina perfectamente.

El problema en los modelos de respuesta incremental es que, por un lado, no se puede observar si un individuo es persuasible, una causa perdida, una causa segura, o un cliente al que no molestar, y, por otro, que no puede estar a la vez en el grupo de tratamiento y en el de control, por lo que no se puede saber cuál hubiera sido su desempeño en caso de estar en el otro grupo (problema de inferencia causal ya mencionado). Por tanto, el *uplift* no se puede medir u observar

[14] En realidad, es posible que el coeficiente Gini tome valores negativos, de hasta el -100%, si el modelo discrimina en sentido contrario al esperado.

a nivel individuo, y las métricas utilizadas en los modelos predictivos no son válidas para evaluar modelos de respuesta incremental.

La solución que proponen la mayoría de los métodos de evaluación de estas técnicas, que parten de la propuesta inicial de Lo (2002), es comparar la respuesta incremental en distintos segmentos, grupos o deciles en los que se divide la población de la muestra utilizada para evaluarlos. Es decir, para esas agrupaciones, que serán definidas utilizando el *uplift* estimado por el modelo, calcular la diferencia entre la tasa de respuesta positiva en el grupo de tratamiento y el grupo de control.

Los principales[15] métodos de evaluación de los modelos de respuesta incremental son:

- Enfoques visuales: gráfico de respuesta incremental, curva de respuesta incremental acumulada (o curva Qini).

- Métricas de evaluación: coeficientes Qini.

Por otro lado, están empezando a surgir enfoques con métricas de evaluación *ad hoc* al problema que se aborda, como es el caso de la métrica de rentabilidad (máxima respuesta incremental en el beneficio, MPU) que se introduce en Devriendt *et al* (2021) para el problema de fuga de clientes.

Para entender adecuadamente estos métodos de evaluación, es conveniente reincidir en la interpretación de la predicción de la respuesta incremental y en cómo ésta debería ser utilizada: una vez se estima el modelo, se puntuarían con él todas las observaciones de la muestra de evaluación, por lo que cada individuo i con características $\boldsymbol{x}^i$ tendrá una respuesta incremental estimada $\widehat{uplift}(\boldsymbol{x}^i)$. Si se ordenan todos los individuos de mayor a menor valor del *uplift* estimado, aquellos para los que la estimación sea positiva y muy grande se deberían incluir en la nueva campaña, y a medida que la estimación se aproxima a cero, y especialmente si se hace negativa, se deberían excluir.

3.1. Gráfico de respuesta incremental

En Lo (2002) se presenta un método para evaluar gráficamente los modelos de respuesta incremental que, aunque se introduce para el enfoque de *dummy* de tratamiento de Lo, puede ser aplicado para el resto de enfoques, pues no depende de la manera en la que se estima la respuesta incremental.

El procedimiento es el siguiente:

1. En primer lugar, se divide la muestra en un 70% que será utilizado para estimar el modelo, y un 30% que se utiliza para evaluarlo.

[15] Existen otras métricas menos utilizadas como el Gini ajustado (se puede consultar más detalle en Kane *et al* (2014)), el AUUC (ver Rzepakowski y Jaroszewicz (2012)), o una extensión del KS para modelos *uplift* (Surry y Radcliffe (2011)).

2. Una vez el modelo ha sido construido, se estima el *uplift* para los individuos de la muestra de evaluación y se ordenan las estimaciones de mayor a menor respuesta incremental.

3. Se agrupan las observaciones en deciles.

4. En cada decil, se observa la respuesta positiva y se calcula su media en el grupo de tratamiento y en el grupo de control, y se obtiene la respuesta incremental de ese decil como la diferencia de dichas tasas.

5. Se grafican estas respuestas incrementales por decil.

Idealmente, si el modelo funciona adecuadamente, en el primer decir habría sólo (o mayoritariamente) individuos persuasibles, mientras que el último estaría copado por individuos "no molestar", y en los deciles intermedios caerían las causas seguras y perdidas. Un buen modelo dejará en el primer decil a los que responden positivamente del grupo de tratamiento (TR), y en el último a los que responden positivamente del grupo de control y los que responden negativamente en el de tratamiento (CR y TN). Por lo tanto, se puede evaluar el modelo observando si la respuesta incremental decrece, y cuánto lo hace, a lo largo de los deciles, y también comparando la respuesta incremental de los primeros deciles con la que se tiene en la población total (que simplemente se debe a la campaña aleatoria previa).

En Lo (2002) se incorpora un ejemplo para ilustrar este método de evaluación, donde en primer lugar (Figura 9) observamos la tasa de respuesta en los grupos de tratamiento y control de cada decil, y en segundo lugar (Figura 10) la respuesta incremental observada y estimada en cada uno de los deciles.

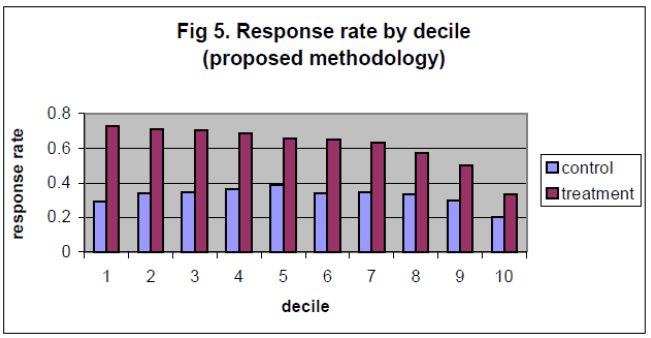


FIGURA 9

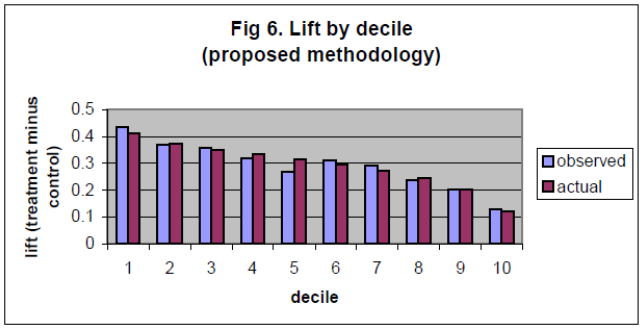


FIGURA 10

Como se ilustra en este ejemplo, es posible que la tendencia de la respuesta incremental no sea monótona a lo largo de los deciles de la muestra de validación, como sucede en los deciles 5 y 6. Nótese, además, que se pueden dar respuestas incrementales negativas en los últimos deciles, lo que significaría que en esos deciles la tasa de respuesta positiva es mayor en el grupo de control que en el grupo de tratamiento, como se puede observar en el siguiente ejemplo de Devriendt *et al* (2018), en el que también se pierde la monotonía en varios deciles al pasar de la muestra de entrenamiento (Figura 11) a la muestra de validación (Figura 12), y sirve para ilustrar la inestabilidad de la que pueden adolecer estos modelos.

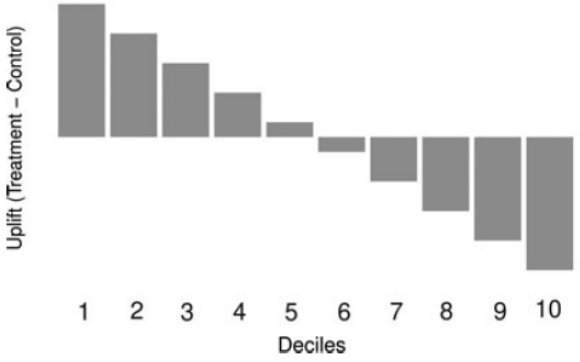


FIGURA 11

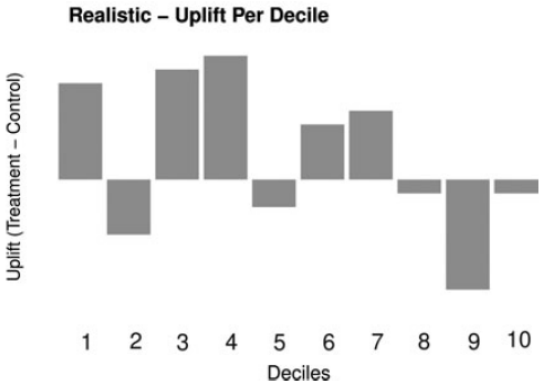


FIGURA 12

Este método de evaluación puede ser utilizado para comparar distintos modelos sobre el mismo conjunto de datos, pues se pueden comparar las respuestas incrementales observadas en los deciles generados por los distintos modelos, pero es deseable contar con una métrica que resuma la capacidad de estos modelos en una escala conocida y comparable.

3.2. Curve de respuesta incremental acumulada y Qini

En Radcliffe (2007) se construye una curva de respuesta incremental acumulada, la curva Qini, basada en el coeficiente Gini y su relación con la curva ROC.

Para dicha construcción:

- Se ordenan los individuos de la muestra de validación de mayor a menor respuesta incremental estimada.

- Se generan grupos de puntuaciones, y se van acumulando los individuos de dichos grupos (por ejemplo: decil 1, decil 1 y 2, decil 1 a 3…). Se pueden utilizar frecuencias relativas o absolutas (que requieren de un ajuste en el cálculo del *uplift* observado).

- En el eje horizontal se acumula la población de esos grupos, y el eje vertical corresponde a la respuesta incremental acumulada hasta cada uno de los grupos, midiendo la respuesta incremental de nuevo como la diferencia de la tasa de respuesta positiva entre el grupo de tratamiento y el de control.

- El último punto de este gráfico acumula a toda la población en el eje horizontal y representa la diferencia entre la tasa de respuesta positiva del grupo de tratamiento y del grupo de control (es decir, la respuesta incremental debida a la campaña aleatoria previa) en el eje vertical.

- Si se une el origen del gráfico con el último punto, se obtiene una recta que representa la respuesta incremental de una marcación de campaña aleatoria.

Este gráfico muestra cómo el modelo es capaz de acumular las respuestas incrementales de distintos grupos respecto a cómo lo hace la marcación aleatoria. Nótese que esta curva no será creciente si, por ejemplo, en los últimos deciles el grupo de control presenta una tasa de respuesta mayor a la del grupo de tratamiento (por lo que la diferencia entre las tasas de respuesta acumuladas hasta estos grupos se reduciría). El siguiente ejemplo de Radcliffe (2007) sirve para ilustrar este método de evaluación (la curva aparece por debajo de la recta aleatoria por tratarse de un problema de retención en el que el objetivo es reducir la tasa de fuga):

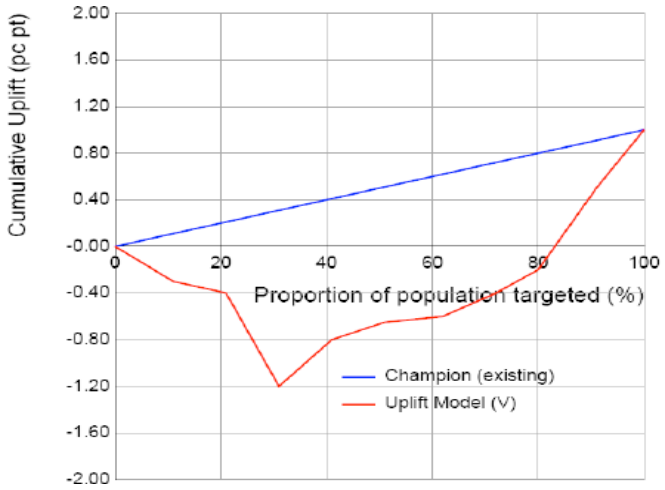


FIGURA 13

En este ejemplo se observa que, con la actual campaña aleatoria la tasa de fuga es aproximadamente 1 punto porcentual mayor en el grupo de tratamiento que en el de control, pero que, en el 30% de la población con mejor respuesta incremental estimada (es decir, más baja), el modelo es capaz de obtener una tasa de fuga 1.2 puntos porcentuales más baja en tratamiento respecto a control.

Junto con la curva de respuesta incremental acumulada, en Radcliffe (2007) se construye una métrica de evaluación similar al Gini, utilizando la idea de comparar la curva obtenida con la curva ideal, pero haciendo uso de la curva Qini: el coeficiente Qini. Se define el Qini, Q, como una ratio entre dos áreas: la que hay entre la curva Qini y la diagonal aleatoria (escenario base o *baseline*, sin modelo), y la que hay entre la curva óptima y la diagonal aleatoria, donde la curva del modelo óptimo es la que ordena en primer lugar a los individuos que pertenecen a TR, y en último lugar a los que pertenecen a CR. Esta métrica es fácilmente interpretable, pues presenta la misma escala que el coeficiente Gini: su valor máximo es 1 (o 100%), cuando coincide con el modelo óptimo, y el mínimo es -1 (o -100%), cuando el modelo ordena el primer lugar a los individuos de CR y en último lugar a los de TR.

Debido a las dudas[16] que hay sobre si efectivamente existe la curva teórica óptima (de acuerdo con Radcliffe (2007)), se introduce otra métrica que remplaza la curva óptima por la curva óptima *zero-downlift*, que es la línea horizontal que pasa por el último punto de las curvas (el que representa la respuesta incremental en la muestra total). Se calcula esta nueva métrica qini, q_0, como la ratio entre el área que hay entre la curva Qini y la diagonal aleatoria (como en el cálculo de Q), y la que hay entre la curva óptima *zero-downlift* y la diagonal aleatoria (por lo que es necesario que la respuesta incremental en la población total sea distinta de cero para no tener un cero en el denominador).

[16] Por ejemplo, no se especifica cómo se ordenarían las observaciones que hay en CN o en TN, y no es posible conocer cuál sería su respuesta si estuvieran en el otro grupo (problema de inferencia causal).

El siguiente ejemplo, extraído de Radcliffe (2007) y que utiliza frecuencias absolutas en lugar de relativas, ilustra el cálculo de la métrica Qini, pues incorpora la curva de respuesta incremental óptima (curva roja), donde el modelo ordena en primer lugar a los individuos del grupo de tratamiento que responden positivamente (primeros 30,000) y en último lugar a los que responden positivamente en el grupo de control (últimos 10,000).

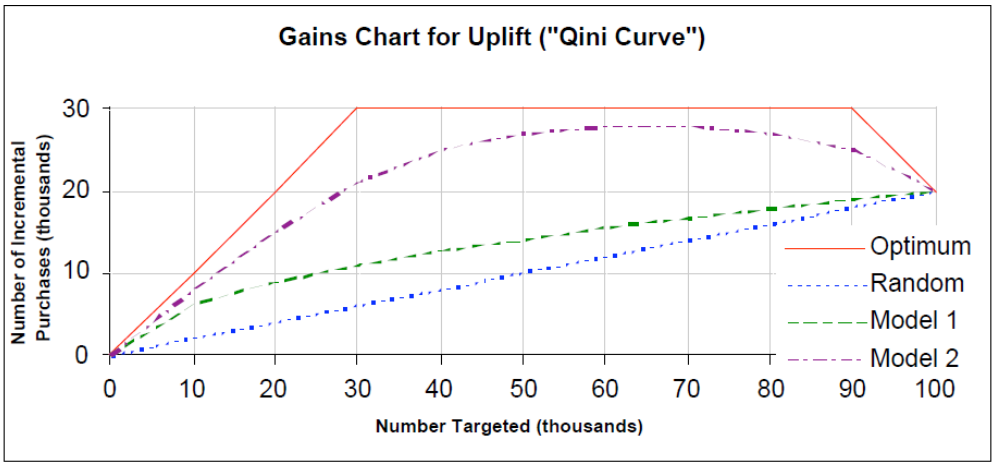


FIGURA 14

En este ejemplo, donde la curva óptima *zero-downlift* sería una línea horizontal que cruza el eje vertical por el valor 20, Radcliffe obtiene que para el primer modelo $Q = 21\%$ y $q_0 = 32\%$, mientras que para el segundo $Q = 76\%$ y $q_0 = 114\%$, valores alineados con lo que se observa en las curvas Qini.

Por último, como se señala en Raddcliffe y Surry (2011), en la práctica[17] se utiliza una métrica Qini más generalizada, a la que aquí llamaremos $\widetilde{Q}$ para distinguirla de Q, que se obtiene simplemente como el área entre la curva Qini y la diagonal aleatoria, lo que facilita su cálculo (áreas de los trapezoides formados por los puntos que definen la curva Qini menos área del triángulo bajo la curva aleatoria) y evita el problema de denominador igual a cero de q_0. De hecho, $\widetilde{Q}$ es el numerador en las dos métricas Qini definidas anteriormente (Q y q_0). Por ejemplo, si se construye la curva Qini utilizando deciles, se puede calcular esta métrica como

$$\widetilde{Q} = \sum_{i=1}^{10} \left(\min\left(uplift_{10\cdot i}, uplift_{10\cdot(i-1)}\right) \cdot 0.1 + \text{abs}\left(uplift_{10\cdot i} - uplift_{10\cdot(i-1)}\right) \cdot 0.1/2\right)$$

donde $uplift_{10}, \ldots, uplift_{90}$ denotan las respuestas incrementales acumuladas en cada uno de los grupos (decil 1, decil 1 y 2, decil 1 a 3…), $uplift_0 = 0$, y $uplift_{100}$ es la respuesta incremental en esa población (diferencia de la tasa de respuesta positiva entre el grupo de tratamiento y el de control).

[17] Por ejemplo, en Devriendt *et al* (2021) se define únicamente de esta manera.

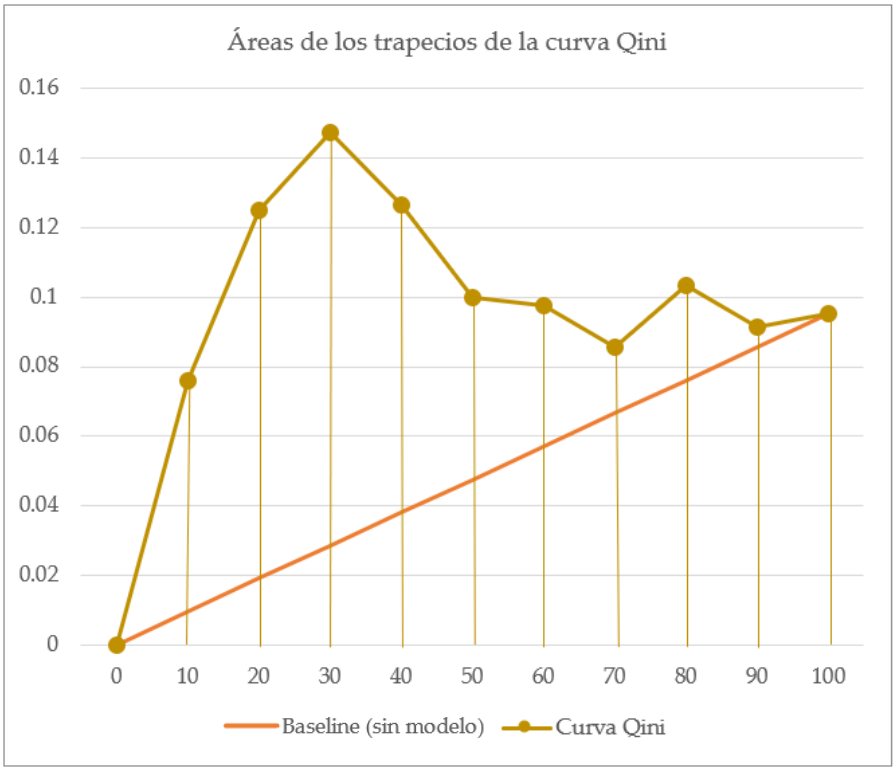

Áreas de los trapecios de la curva Qini
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
40
50
60
70
80
90
100
Baseline (sin modelo)
Curva Qini

FIGURA 15

CAPÍTULO 4

APLICACIÓN A UN CASO DE USO: COMERCIALIZACIÓN DE DEPÓSITOS BANCARIOS

En este capítulo se van a utilizar algunos de los modelos presentados en este trabajo para abordar un problema real. El problema en cuestión es una campaña de marketing de un banco minorista portugués que quiere comercializar depósitos a largo plazo. El conjunto de datos está disponible en el repositorio de *Machine Learning* de la Universidad de California en Irvine[18] a raíz del trabajo que aparece en Moro *et al* (2014), donde se utilizan únicamente técnicas predictivas.

El banco está interesado en realizar una nueva campaña para comercializar depósitos y quiere aprovechar los datos que tiene de campañas anteriores realizadas en el período 2008-2013 sobre una preselección de clientes a los que se contactó de manera aleatoria por teléfono fijo o móvil. Se quiere realizar una nueva campaña sobre una preselección de clientes similar, y se espera definir el grupo de clientes a contactar por cada canal con el apoyo de un modelo estadístico.

Este conjunto de datos contiene 41,188 clientes y 21 variables, incluyendo la variable de respuesta, que indica si finalmente el cliente contrató un depósito ($y = 'yes'$) o no ($y = 'no'$), y la variable de tratamiento $contact$, que indica si el cliente fue contactado por teléfono fijo ($contact = 'telephone'$) o móvil ($contact = 'cellular'$). En el grupo contactado por teléfono móvil, que contiene el 63.47% de esta población, la tasa de respuesta positiva es del 0.147, mientras que en el grupo de teléfono fijo, con el 36.53% de la muestra, la respuesta positiva es del 0.052. La tasa de respuesta positiva en la población total es del 0.113. Para estimar los modelos es necesario utilizar variables binarias, y como se observa una mayor tasa de respuesta en el grupo contactado por teléfono móvil, se va a definir la variable de respuesta y tratamiento para un individuo i como

[18] Se puede consultar y descargar en https://archive.ics.uci.edu/ml/datasets/bank+marketing

$$r^i = \begin{cases} 1 & y^i = 'yes' \\ 0 & y^i = 'no' \end{cases}$$

$$t^i = \begin{cases} 1 & contact^i = 'cellular' \\ 0 & contact^i = 'telephone' \end{cases}$$

Con esta definición, la respuesta incremental debida al tratamiento fue de $0.147 - 0.052 = 0.095$.

Se puede ver una descripción de las variables restantes en el Apéndice A.

4.1. Modelos construidos

Antes de estimar los modelos, se ha separado la muestra en una de entrenamiento, con el 70% de las observaciones, que será utilizada para estimar los modelos, y una muestra de validación, con el 30% de las observaciones, que se utiliza para evaluar los modelos. Esta partición se ha realizado estratificando por la variable de respuesta y de tratamiento. Además, se han excluido las variables macroeconómicas ($cons.conf.idx, cons.price.idx, euribor3m, nr.employed, empvarrate$) debido a que son variables poco accionables desde un punto de vista comercial. También se excluyen la variable $duration$, pues no se conoce la duración de la llamada antes de contactar con el cliente, y la variable $campaign$, que indica el número de veces que se ha contactado con el cliente en esta campaña (pues inicialmente será cero para todos).

Para poder comparar diferentes estimaciones y que cubran la mayoría de los casos que se han estudiado en este trabajo, se van a construir[19] modelos utilizando los siguientes enfoques:

- Tradicional estimado con un árbol (modelo *ctree* del paquete *partykit* de R).
- Dos modelos estimados con un árbol cada uno (modelo *ctree* del paquete *partykit* de R).
- *Dummy* de tratamiento estimado con un árbol (modelo *ctree* del paquete *partykit* de R).
- Transformación del *target* con el enfoque de Lai estimado con un árbol (modelo *ctree* del paquete *partykit* de R).
- Basado en árboles utilizando la divergencia KL (modelo *upliftRF* del paquete *uplift* de R, parametrizado[20] con *mtry = 13, ntree = 1, minsplit=300, minbucket_ct0=1, minbucket_ct1=1, split_method = "KL", bag.fraction = 1*).

[19] Los modelos han sido estimados utilizando el software estadístico R. El proyecto de RStudio con estos desarrollos se adjunta en el Apéndice B junto con un Excel en el que se generan los gráficos de respuesta incremental y curvas y métricas Qini, y las librerías utilizadas aparecen en el Apéndice C.

[20] En los modelos en los que es posible llevar a cabo un ejercicio de *parameter tuning*, se han entrenado algunos modelos con distintos valores para los parámetros y se ha elegido el mejor de los modelos atendiendo a la métrica de evaluación $\widetilde{Q}$, pero no se han explorado en detalle las distintas combinaciones de parámetros porque su optimización queda fuera del alcance de este trabajo.

- Basado en *random forest* utilizando la distancia euclídea (modelo *upliftRF* del paquete *uplift* de R, parametrizado con *mtry = 4, ntree = 100, minsplit=100, minbucket_ct0=3, minbucket_ct1=3, split_method = "ED", bag.fraction = 1*).

El primero de los enfoques de respuesta incremental que se ha trabajado ha sido el de dos modelos. Debido a que los dos modelos se pueden estimar con cualquiera de las técnicas habituales, se han construido tanto utilizando árboles como con regresiones logísticas (modelo *DualUplift* del paquete *tools4uplift* de R). La métrica Qini $\tilde{Q}$ toma el valor 0.132 para el enfoque de doble modelización con árboles y 0.127 para dicho enfoque cuando se emplea la regresión logística, por lo que se considera más adecuado el que utiliza árboles. Además, como se puede observar en la Figura 17, la curva Qini de los dos modelos construidos con árboles alcanza un *uplift* observado superior en los primeros deciles acumulados. Por este motivo se han utilizado árboles cuando en algunos de los enfoques era posible estimar el modelo con cualquier técnica conocida (regresión, árboles…), para estar alineado con este primer enfoque y evitar tener que valorar en cada uno de ellos qué técnica de estimación utilizar (regresión o árbol, pero también red neuronal, SVM…). Estas dos estimaciones para el enfoque de dos modelos han sido comparadas también con un modelo estimado con un árbol siguiendo el enfoque tradicional, con un Qini de 0.048. En la Figura 16 se pueden comparar los gráficos de respuesta incremental para estas tres estimaciones, y en la Figura 17 se pueden comparar sus curvas Qini.

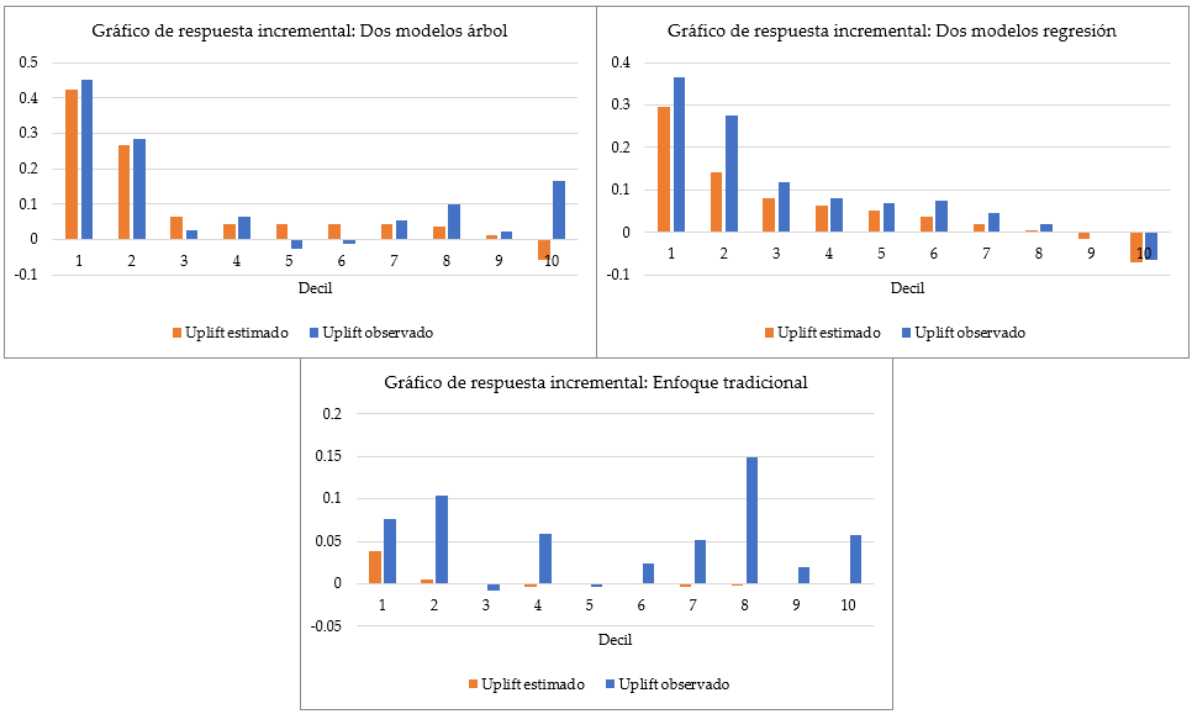


FIGURA 16

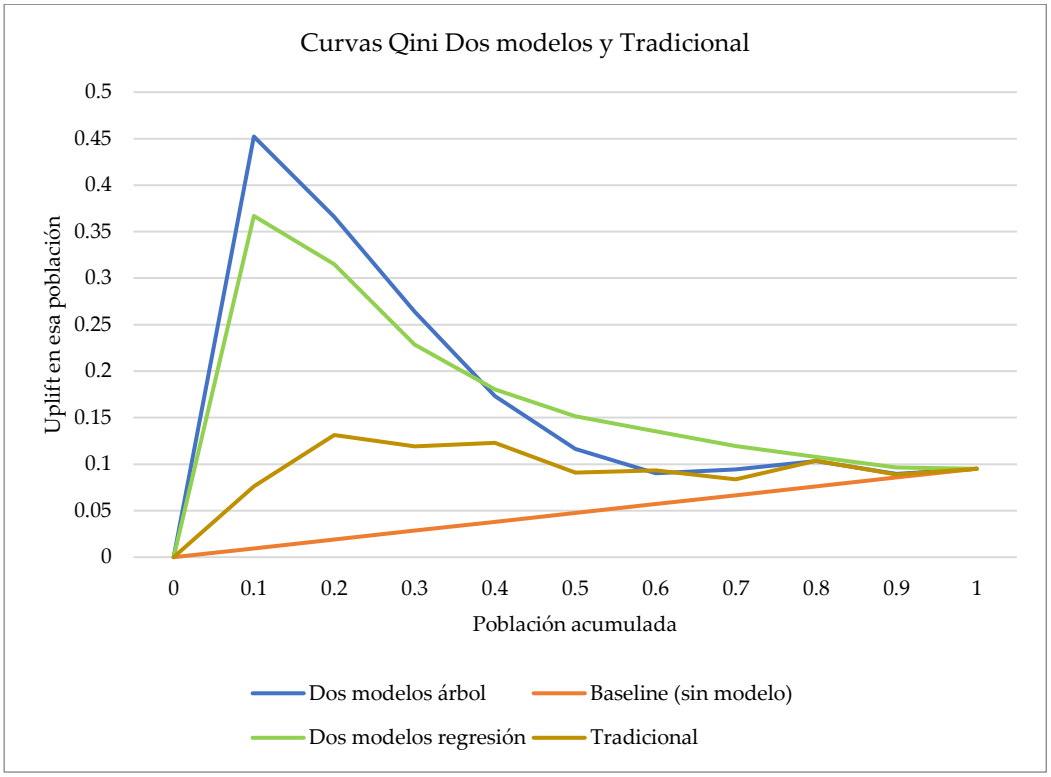
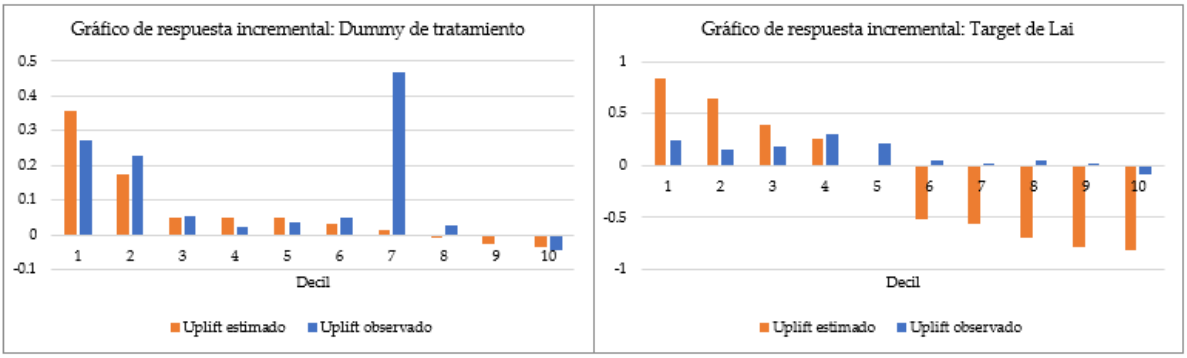


FIGURA 17

A continuación, se han obtenido modelos siguiendo el enfoque de *dummy* de tratamiento y el de transformación del *target* de Lai utilizando un árbol para estimarlos, resultando en coeficientes Qini de 0.091 y 0.127 respectivamente, y dando lugar a los gráficos de respuesta incremental de la Figura 18. Para los dos enfoques restantes, basados en un árbol y en un *random forest* con métricas de divergencia, sí que cabe destacar que se han estimado modelos con las distintas métricas propuestas en el enfoque de Rzepakowski y Jaroszewicz (divergencia KL, distancia euclídea y divergencia chi-cuadrado), pero los que aquí se presentan son los que obtienen un mayor valor en el coeficiente Qini, de 0.17 para el árbol y 0.165 para el *random forest*. En la Figura 18 se muestras los gráficos de respuesta incremental de estos dos enfoques, y en la Figura 19 se presentan todas las curvas Qini.



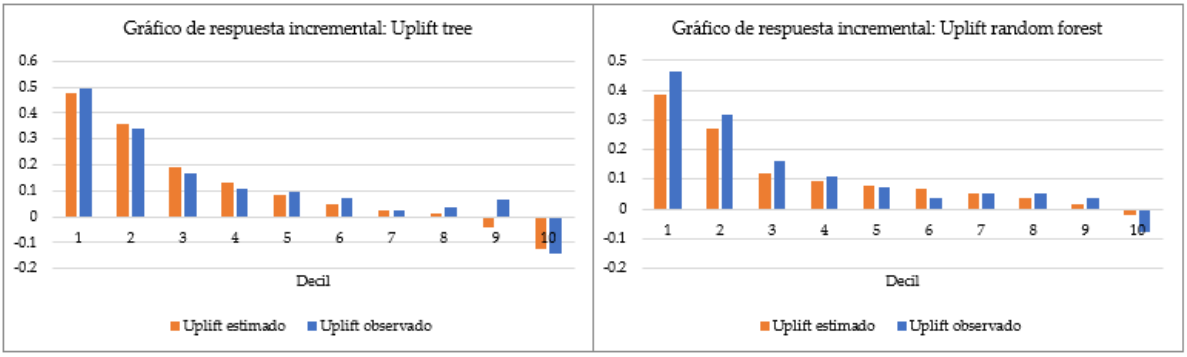


FIGURA 18

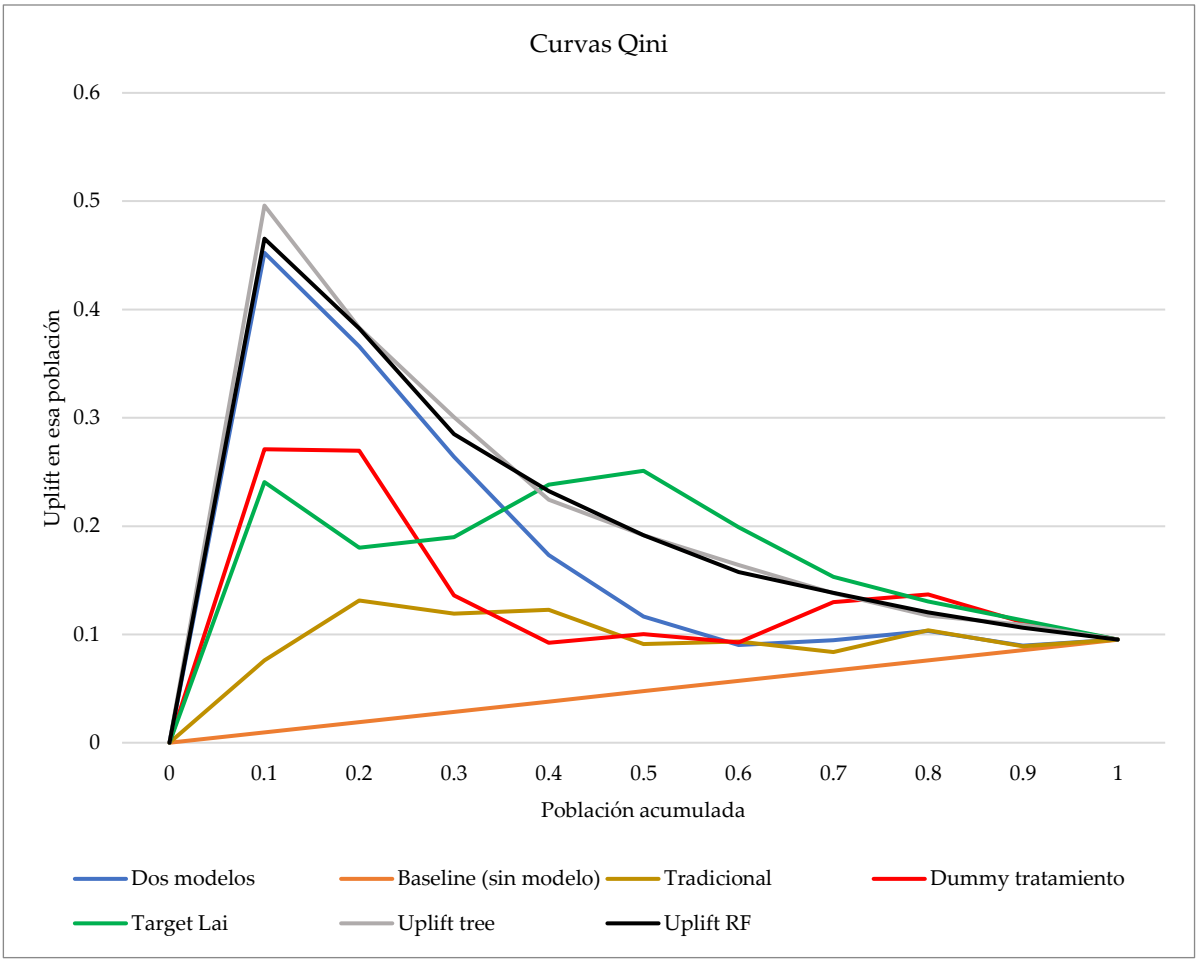


FIGURA 19

En los gráficos de respuesta incremental, se observa que algunos de estos enfoques son bastante inestables (dos modelos, *dummy* de tratamiento, *target* de Lai) o no están diseñados para modelizar el *uplift* (tradicional). En la curva Qini, destacan dos modelos sobre el resto: árbol y *random forest* de respuesta incremental. Aparte de ser los que tienen un mayor *uplift* en los primeros deciles, presentan el comportamiento deseable para la curva Qini, pues a medida que se incrementa la población

acumulada, desciende la respuesta incremental observada en esa población. En particular, el árbol de respuesta incremental presenta una curva incluso un poco mejor a la del *random forest* en los primeros deciles acumulados (respuesta incremental de 0.496 frente a 0.465 en el primer decil, y de 0.301 frente a 0.285 en los primeros tres deciles). Si se comparan las métricas Qini $\tilde{Q}$ de estos modelos, el árbol de respuesta incremental también resulta ser el mejor de los modelos construidos:

Enfoque	Qini
Tradicional	0.048
Dos modelos	0.132
Dummy	0.091
Lai	0.127
Uplift tree	0.170
Uplift RF	0.165

TABLA 1

Nótese que, como se destaca en Devriendt *et al* (2018) o en Rößler *et al* (2021), el hecho de que el árbol de respuesta incremental resulte ser el mejor modelo entre los estimados para este conjunto de datos no quiere decir que sea el mejor para todos los conjuntos de datos, y al abordar otros problemas deberían barajarse todos los enfoques posibles.

4.2. Detalles del modelo final

Aunque las curvas y métricas Qini hacen evidente que aplicar cualquiera de los enfoques, incluso el tradicional, supondría una mejora respecto a la marcación de campaña aleatoria, el modelo propuesto al banco portugués sería el árbol de respuesta incremental.

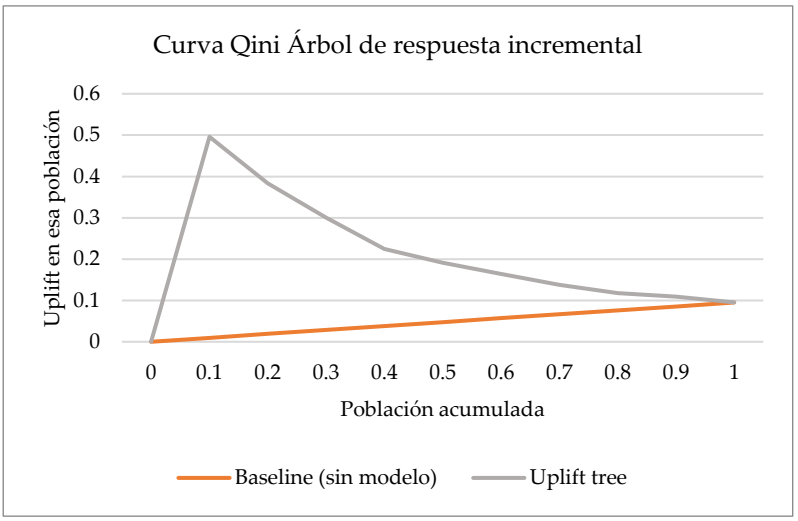


FIGURA 20

El *uplift* que hay en los tres primeros deciles es de 0.496, 0.337 y 0.165 respectivamente. Por otro

lado, el *uplift* en el último decil es de -0.140, por lo que se trata del menor *uplift* por decil entre todos los modelos estimados, lo que indica que además es capaz de identificar a clientes a los que nunca habría que incluir en el grupo de tratamiento de la nueva campaña, por observarse mayores tasas de respuesta positiva en el grupo de control.

	Tasa respuesta positiva		
Decil	**Tratamiento**	**Control**	**Uplift observado**
1	0.555	0.059	0.496
2	0.388	0.051	0.337
3	0.211	0.046	0.165
4	0.157	0.052	0.105
5	0.136	0.042	0.094
6	0.115	0.041	0.074
7	0.106	0.084	0.023
8	0.060	0.024	0.036
9	0.106	0.042	0.064
10	0.086	0.226	-0.140
Total	0.147	0.052	0.095

TABLA 2

En términos acumulados, por ejemplo, en los tres primeros deciles hay un *uplift* de 0.301, mientras que en los seis primeros es de 0.164.

	Tasa respuesta positiva		
Población	**Tratamiento**	**Control**	**Uplift observado**
Decil 1	0.555	0.059	0.496
Deciles 1 - 2	0.439	0.056	0.383
Deciles 1 - 3	0.353	0.053	0.301
Deciles 1 - 4	0.277	0.052	0.224
Deciles 1 - 5	0.242	0.051	0.191
Deciles 1 - 6	0.214	0.050	0.164
Deciles 1 - 7	0.189	0.051	0.138
Deciles 1 - 8	0.167	0.049	0.117
Deciles 1 - 9	0.158	0.049	0.109
Deciles 1 - 10	0.147	0.052	0.095

TABLA 3

Si efectivamente la marcación de tratamiento en la campaña previa fue aleatoria, y se va a incluir un determinado decil en una nueva campaña, se podría utilizar como estimador de la tasa de respuesta positiva en ese decil la tasa de respuesta positiva que se observa en el grupo de tratamiento de ese decil; por el contrario, si no se va a incluir un decil en la campaña, se puede utilizar como estimador de la tasa de respuesta positiva para ese decil la que se observa en su grupo de control. Por tanto, si se asume que la marcación en la campaña previa fue aleatoria, es posible construir una sencilla métrica para proporcionarle al banco portugués un punto de corte

determinando qué deciles deberían ser incluidos en una nueva campaña[21], y cuáles no.

4.3. Estudio de rentabilidad

Dado que para este problema no se han publicado cifras reales sobre indicadores de rentabilidad, se van a asumir como se indica a continuación:

- Incluir a un cliente en la campaña conlleva un coste de 65€.
- La ganancia asociada a un cliente con respuesta positiva es de 1,000€.

Así, por cada respuesta positiva en el grupo de tratamiento y control, el beneficio sería de 935€ y 1,000€ respectivamente, mientras que por cada respuesta negativa se tendría un beneficio de -65€ o 0€ en cada caso. Para cada decil, se puede estimar el beneficio unitario esperado, tanto si fuera incluido en el grupo de tratamiento como si lo fuera en el de control, al multiplicar la tasa de respuesta positiva esperada correspondiente por el beneficio unitario y restar el coste unitario.

En la siguiente tabla se puede observar el beneficio unitario esperado de cada decil si fuera incluido en el grupo de tratamiento o en el grupo de control, así como su diferencia.

	Tasa respuesta positiva		Beneficio unitario esperado		
Decil	Tratamiento	Control	Tratamiento	Control	Diferencia
1	0.555	0.059	489.66	58.70	430.95
2	0.388	0.051	323.40	51.05	272.35
3	0.211	0.046	145.63	45.58	100.06
4	0.157	0.052	92.07	52.37	39.71
5	0.136	0.042	70.84	42.28	28.56
6	0.115	0.041	49.77	40.77	9.01
7	0.106	0.084	41.38	83.87	-42.49
8	0.060	0.024	-5.22	24.10	-29.32
9	0.106	0.042	40.86	41.83	-0.97
10	0.086	0.226	20.94	226.19	-205.25

TABLA 4

Como la diferencia en el beneficio unitario esperado al incluir el decil en tratamiento o control es positiva en los seis primeros deciles, se recomendaría al banco incluirlos en el grupo de tratamiento.

El beneficio de utilizar este enfoque frente a la definición aleatoria del grupo de tratamiento es evidente: con la campaña aleatoria se observaba una tasa de respuesta de 0.147 en el grupo de tratamiento, que generaría un beneficio unitario de 82€, mientras que la tasa de respuesta positiva en el grupo de control era de 0.052, con un beneficio unitario asociado de 52€.

[21] Nótese que, para ello, se debería puntuar la nueva muestra con el modelo y elegir los deciles correspondientes.

CONCLUSIONES

En este trabajo se ha realizado una revisión de los principales enfoques para modelizar la respuesta incremental, ilustrando sus diferencias y destacando sus beneficios frente a la metodología tradicional de predicción de la respuesta positiva. Además, se complementa el trabajo con un caso de uso en el que se han estimado y comparado algunos modelos para un problema real de comercialización de depósitos bancarios, sirviendo de ejemplo de cómo este tipo de modelos pueden aportar valor al negocio más allá su interés teórico.

Como principales conclusiones de este trabajo se puede destacar que:

1. Los modelos de respuesta incremental suponen una mejora respecto a los modelos de predicción tradicionales en los problemas de definición de una campaña (de su grupo de tratamiento), pues, al incorporar la separación de la población en individuos persuasibles, causas seguras, causas perdidas y "no molestar", hacen que la modelización esté más alineada con el objetivo de la definición de una nueva campaña, que sería detectar a los individuos que responden positivamente debido a la campaña, pero que no lo harían en ausencia de ella.

2. Aunque las técnicas de ensamblado mitiguen los problemas de inestabilidad de otros enfoques, no existe un modelo que sea el mejor para cualquier conjunto de datos, por lo que es necesario construir y comparar modelos utilizando distintos enfoques y parametrizaciones cuando se trabaje con problemas de respuesta incremental.

Por otro lado, el trabajo de esta memoria ha abierto posibles vías de ampliación de lo estudiado:

1. Extensión al problema multitratamiento y de respuesta continua.

2. Preselección de variables: se trata de un problema en el que de hecho se ha trabajado en el desarrollo de esta memoria, definiendo un *information value* basado en la definición del *target* de Lai, que bien podríamos llamar IV_{Lai}, y que en los primeros resultados parecía estar alineado con la métrica *NIV* que se utiliza por ejemplo en Devriendt *et al* (2018). Una preselección de variables sería de especial interés cuando se quieren optimizar las parametrizaciones de modelos como el árbol y bosque de respuesta incremental, pues

reducen el tiempo de computación. Otro enfoque razonable podría ser estimar un árbol de respuesta incremental para cada variable y medir su Qini para realizar la preselección con este Qini a nivel variable.

3. Debido a su potencial, sería deseable disponer de casos reales en los que estos modelos se han aplicado, para poder obtener cifras de efectividad de esas campañas y evaluar así los valores estimados por los modelos.

4. Analizar qué pasa si la campaña no ha sido diseñada de manera aleatoria y por tanto existe un sesgo en la marca de tratamiento y control. Por ejemplo, cómo se definiría una nueva campaña si en la campaña previa ya se utilizó algún enfoque de respuesta incremental.

5. Algunos enfoques sufren de inestabilidad, y sería interesante estudiar si añadiendo alguna restricción en la construcción sería posible mitigarla, tomando como ejemplo la imposición de la monotonía en algunos problemas de respuesta binaria con regresión logística.

6. Estudiar transformaciones de variables que maximicen la ganancia en Qini al incorporarlas al modelo y que también sirvan para mejorar su interpretabilidad. Puede servir de ejemplo el WOE de los modelos de predicción y la extensión al $NWOE$ de los modelos de respuesta incremental que aparece en Devriendt *et al* (2018), pero también se podría estudiar alguna otra alternativa que utilizara ideas similares a los *multivariate adaptive regression splines* (MARS).

7. Estudiar la posibilidad de adaptar otras técnicas de predicción existentes (como redes neuronales) a modelos de respuesta incremental.

8. Sustituir las métricas de evaluación existentes por métricas basadas en la rentabilidad y el beneficio. Las métricas de evaluación podrían no estar totalmente alineadas con los objetivos de negocio, pues se limitan a comparar las áreas de las curvas Qini y no tienen en cuenta el umbral a fijar para el lanzamiento de una nueva campaña, por lo que podría haber un modelo con peor Qini pero mejor rentabilidad (si por ejemplo tiene buen comportamiento en el primer decil, pero mal en el resto). Existen trabajos en este sentido, como Devriendt *et al* (2021), pero no hay una generalización de estas métricas, sino que las nuevas propuestas surgen asociadas a problemas concretos (retención de clientes en ese caso).

9. Estudiar la posibilidad de generar enfoques de modelización basados en rentabilidad o beneficio, por ejemplo incorporando estas métricas en la división de los nodos en los árboles y bosques de respuesta incremental.

APÉNDICE A

DICCIONARIO DE DATOS

A continuación se presenta una explicación de las variables incluidas en el conjunto de datos utilizado en el Capítulo 4.

- age: edad del cliente.
- job: tipo de trabajo (categorías: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown').
- $marital$: estado civil (categorías: 'divorced', 'married', 'single', 'unknown', aunque 'divorced' incluye viudos).
- $education$: nivel educativo (categorías: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown').
- $default$: si el cliente tiene un crédito en default (categorías: 'no', 'yes', 'unknown').
- $housing$: si el cliente tiene una hipoteca (categorías: 'no', 'yes', 'unknown').
- $loan$: si el cliente tiene un préstamo personal (categorías: 'no', 'yes', 'unknown').
- $contact$: canal con el que se contacta al cliente (categorías: 'cellular','telephone').
- $month$: mes del último contacto con el cliente (categorías: 'jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'oct', 'nov', 'dec').
- day_of_week: día de la semana del último contacto con el cliente (categorías: 'mon', 'tue', 'wed', 'thu', 'fri').
- $duration$: duración de la última llamada con el cliente.
- $campaign$: número de contactos realizados durante la campaña.
- $pdays$: días desde que al cliente se le contactó en una campaña previa.
- $previous$: número de contactos realizados durante la campaña anterior.
- $poutcome$: resultado de la campaña previa ('failure', 'nonexistent', 'success').
- $emp.var.rate$: tasa de variación trimestral del empleo.
- $cons.price.idx$: índice de precios de consumo mensual.

- $cons.conf.idx$: índice de confianza del consumidor mensual.
- $euribor3m$: euríbor a tres meses diario.
- $nr.employed$: número de empleados del trimestre.
- y: indicador de si el cliente ha contratado el depósito.

APÉNDICE B

PROYECTO RSTUDIO

Se adjuntan[22] los códigos en R del proyecto generado en el último capítulo, así como un Excel que contiene la muestra de validación con la respuesta incremental estimada por cada uno de los enfoques y la construcción de las curvas Qini asociadas.


Proyecto uplift.rar


Salidas_modelos_y_curvas.xlsx

[22] Aunque en la versión entregada para finalizar el Máster Universitario en Matemáticas Avanzadas de la UNED se adjuntaba este material, se ha excluido de la versión publicada. En cualquier caso, se pondrá a disposición del lector si así lo requiere contactando con el autor (javicarrascos92@gmail.com).

APÉNDICE C

DOCUMENTACIÓN LIBRERÍAS R

El proyecto del apéndice anterior ha sido desarrollado en la versión 4.0.3 de R; nótese que alguna de las librerías utilizadas, que se listan a continuación, podría no estar disponible en versiones anteriores de este software. Además, las librerías están en una continua revisión y mejora, por lo que es posible que alguno de los códigos no funcione en el futuro si una de las librerías es actualizada y, por ejemplo, alguna función requiere de nuevos parámetros de entrada. Mención especial requiere la librería *tools4uplift*, pues ha sido eliminada del repositorio de CRAN y se tiene que instalar manualmente utilizando versiones previas que aparecen en el archivo, aunque es necesario instalar antes las librerías dependientes (*RItools, coin, tables, penalized*).

- https://cran.r-project.org/web/packages/caTools/caTools.pdf
- https://cran.r-project.org/web/packages/RItools/RItools.pdf
- https://cran.r-project.org/web/packages/coin/coin.pdf
- https://cran.r-project.org/web/packages/tables/tables.pdf
- https://cran.r-project.org/web/packages/penalized/penalized.pdf
- https://cran.microsoft.com/snapshot/2019-12-24/web/packages/uplift/uplift.pdf
- https://cran.r-project.org/web/packages/tools4uplift/tools4uplift.pdf
- https://cran.r-project.org/web/packages/Information/Information.pdf
- https://cran.r-project.org/web/packages/partykit/partykit.pdf
- https://cran.r-project.org/web/packages/gmodels/gmodels.pdf
- https://cran.r-project.org/web/packages/sqldf/sqldf.pdf
- https://cran.r-project.org/web/packages/dplyr/dplyr.pdf
- https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
- https://cran.r-project.org/web/packages/smbinning/smbinning.pdf
- https://cran.r-project.org/web/packages/ROCR/ROCR.pdf
- https://cran.r-project.org/web/packages/RSQLite/RSQLite.pdf
- https://cran.r-project.org/web/packages/proto/proto.pdf

- https://cran.r-project.org/web/packages/gsubfn/gsubfn.pdf

APÉNDICE D

BIBLIOGRAFÍA

[1] Samuelson, D.A. (2013). Analytics: Key to Obama's victory. OR/MS Today February: 20–24, https://www.informs.org/ORMS-Today/Public-Articles/February-Volume-40-Number-1/Analytics-key-to-Obama-s-victory (último acceso el 30 de enero de 2022).

[2] Siegel, E. (2013). *The real story behind Obama's election victory*. The Fiscal Times 21 January, http://www.thefiscaltimes.com/Articles/2013/01/21/The-Real-Story-Behind-Obamas-Election-Victory.aspx#page1 (último acceso el 30 de enero de 2022).

[3] Scherer, M. (2012). *How Obama's data crunchers helped him win*. CNN News, http://www.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team/index.html?hpt=hp_bn5 (último acceso el 30 de enero de 2022).

[4] Issenberg S. (2013). *How President Obama's campaign used big data to rally individual voters*. MIT Technology Review, 2013;116:38–49, https://www.technologyreview.com/2012/12/19/114510/how-obamas-team-used-big-data-to-rally-voters/ (último acceso el 30 de enero de 2022).

[5] Porter, D. (2013). *Pinpointing the persuadables: Convincing the right voters to support Barack Obama*. Presentado en Predictive Analytics World, Oct, Boston, MA, https://www.predictiveanalyticsworld.com/machinelearningtimes/pinpointing-the-persuadables-convincing-the-right-voters-to-support-barack-obama/ (último acceso el 30 de enero de 2022).

[6] Rudas, K., Jaroszewicz, A. (2018). *Linear regression for uplift modelling*. Data Mining and Knowledge Discovery (2018) 32:1275–1305. https://doi.org/10.1007/s10618-018-0576-8

[7] Gubela, R. M., Lessmann, S., Jaroszewicz, S. (2020). *Response transformation and profit decomposition for revenue uplift modelling.* European Journal of Operational Research 283 (2020) 647–661.

[8] Baier, D., Stöcker, B. (2021). *Profit uplift modeling for direct marketing campaigns: approaches and applications for online shops*. Journal of Business Economics (2021). https://doi.org/10.1007/s11573-021-01068-3

[9] Devriendt, F., Moldovan, D., Verbeke, W. (2018). *A Literature Survey and Experimental Evaluation of the State-of-the-Art in Uplift Modeling: A Stepping Stone Toward the Development of Prescriptive Analytics*. Big Data. Mar;6(1):13-41. doi: 10.1089/big.2017.0104. PMID: 29570415.

[10] Kane, K., Lo, V., Zheng, J. (2014). *Mining for the truly responsive customers and prospects using true-lift modeling: Comparison of new and existing methods*. J Market Anal 2, 218–238. https://doi.org/10.1057/jma.2014.18

[11] Rößler, J., Tilly, R., Schoder, D. (2021). *To Treat, or Not to Treat: Reducing Volatility in Uplift Modeling Through Weighted Ensembles*. HICSS: 1-10

[12] Devriendt, F., Berrevoets, J., Verbeke, W. (2021). *Why you should stop predicting customer churn and start using uplift models*. Information Sciences, Volume 548, Pages 497-515, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.12.075.

[13] Gutiérrez, P., Gérardy, J. (2017). *Causal Inference and Uplift Modelling: A Review of the Literature*. Proceedings of The 3rd International Conference on Predictive Applications and APIs, in Proceedings of Machine Learning Research 67:1-13, https://proceedings.mlr.press/v67/gutierrez17a.html

[14] Olaya, D., Coussement, K., Verbeke, W. (2020). *A survey and benchmarking study of multitreatment uplift modeling*. Data Min Knowl Disc 34, 273–308. https://doi.org/10.1007/s10618-019-00670-y

[15] Radcliffe N. J., Surry P. D. (2011). *Real-world uplift modelling with significance based uplift trees*. White Paper TR-2011-1, Edinburgh, UK: Stochastic Solutions.

[16] Lo, V. S. Y. (2002). *The true lift model: A novel data mining approach to response modeling in database marketing*. SIGKDD Explor Newsl. 2002; 4:78–86.

[17] Lai, L. Y. T. (2006). *Influential marketing: A new direct marketing strategy addressing the existence of voluntary buyers*. Simon Fraser University (Canada), M.Sc. Thesis, 2006. https://summit.sfu.ca/item/6629

[18] Shaar, A., Abdessalem, T., Segard, O. (2016). *Pessimistic uplift modeling*. KDD 2016 : 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2016, San Francisco, California, United States.

[19] Jaskowski, M., Jaroszewicz, S. (2012). *Uplift modeling for clinical trial data*. ICML Workshop on Clinical Data Analysis. Edinburgh, UK: ICML Workshop on Machine Learning for Clinical Data Analysis, 2012.

[20] Athey, S., Imbens, G. W. (2015). *Machine learning methods for estimating heterogeneous causal effects*. stat, 1050:5, 2015.

[21] Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series

in Statistics Springer New York Inc., New York, NY, USA, 2001.

[22] Kass, G. V. (1980). *An Exploratory Technique for Investigating Large Quantities of Categorical Data*. Applied Statistics, Vol. 29, No. 2, 1980, pp. 119–127.

[23] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). A*n introduction to statistical learning with applications in R*. New York: Springer, 2013.

[24] Hansotia, B., Rukstales, B. (2002). *Incremental value modelling*. Journal of Interactive Marketing; Philadelphia 16(3), 2002, pp. 35–46.

[25] Rzepakowski, P., Jaroszewicz, S. (2012). *Decision trees for uplift modeling with single and multiple treatments*. Knowl. Inf. Syst. 32 (2) (2012) 303–327.

[26] Soltys, M., Jaroszewicz, S., Rzepakowski, P. (2014). *Ensemble methods for uplift modeling*. Data Min Knowl Discov. 2014;29:1–29.

[27] Guelman, L., Guillén, M., Pérez-Marín, A.M. (2012). *Random forests for uplift modeling: an insurance customer retention case*. Lecture Notes in Business Information Processing 115: 123-133.

[28] Guelman L., Guillén M., Pérez-Marín, A.M. (2014). *Optimal personalized treatment rules for marketing interventions: A review of methods, a new proposal, and an insurance case study*. Working Papers 2014-06. Universitat de Barcelona, UB Riskcenter, 2014. http://ideas.repec .org/p/bak/wpaper/201406.html

[29] Zeileis, A., Hothorn, T., Hornik, K. (2008). *Model-based recursive partitioning*. J. Comput. Graph. Statist. 17 (2), 2008, 492–514.

[30] Radcliffe N. J. (2007). *Using control groups to target on predicted lift: Building and assessing uplift models*. Direct Market J Direct Market Assoc Anal Council. 2007;1:14–21.

[31] Surry P. D., Radcliffe N. J. (2011). *Quality measures for uplift models*. Submitted to KDD 2011.

[32] Moro S., Cortez P., Rita P. (2014). *A data-driven approach to predict the success of bank telemarketing*. Decision Support Systems, Elsevier, 62:22-31, 2014.