THE
BEHAVIOURAL
INSIGHTS TEAM

El uso de la ciencia de datos en políticas públicas

Un informe del Behavioural Insights Team

THE BEHAVIOURAL INSIGHTS TEAM

Contenidos

Prefacio

John Manzoni
Jefe Ejecutivo del Servicio Civil

El enorme aumento en los datos, y de herramientas para analizarlos y utilizarlos, ha transformado al mundo continuamente hasta el presente. Ya sea segmentándonos con anuncios en línea, o utilizando los datos de los motores de búsqueda para predecir dónde ocurrirán los nuevos brotes de gripe y así garantizar que haya medicamentos disponibles, todos estos cambios nos afectan a todos.

Existe un gran potencial para que los gobiernos mejoren el rendimiento y la productividad de los servicios a través del uso inteligente de datos. Estos datos incluyen resultados, patrones de uso, costos, y experiencias de los ciudadanos. Con esta gran cantidad de datos, tenemos la obligación de hacer que los servicios gubernamentales se efectúen de la mejor manera posible. Esto implica aprender de los contextos donde los servicios funcionan bien, y mejorar aquellos donde no lo hacen. Significa personalizar y enfocar los servicios públicos en torno a las necesidades y deseos de las personas y las empresas, e implica el uso de métodos experimentales, y una comprensión desde dentro de la variación del servicio, para determinar rápidamente cómo los servicios y sistemas pueden ser mejorados.

La plena realización de este potencial no ocurrirá de la noche a la mañana. Como cualquier nuevo esfuerzo, aparecerán desafíos a lo largo del camino. Necesitamos estar preparados para innovar, ampliar los enfoques que funcionen, y aprender de los aspectos que no funcionen y adaptarlos. Este informe refleja esta filosofía, y destaca dónde los enfoques han de seguir siendo perfeccionados. Pero también deja en claro por qué esta perseverancia es justificable: el trabajo contenido en este informe representa solamente el comienzo de lo que podemos hacer y, sin embargo, los resultados ya están demostrando el gran impacto que se logrará en servicios básicos del gobierno tales como las escuelas, la salud y la atención social.

Dentro del gobierno, y de manera transversal, ya estamos transformando la forma en que participan los ciudadanos con el Estado, mediante la expansión de la tecnología digital. El siguiente paso es buscar garantizar que los datos obtenidos conduzcan a una mejora constante. Para eso, la aplicación de la ciencia de datos será clave.

John Manzoni
Jefe Ejecutivo del Servicio Civil

Prefacio

David Halpern
Director Ejecutivo del Behavioural Insights Team (BIT)

Desde la creación del equipo de Aprendizajes Conductuales en 2010, hemos realizado más de 400 ensayos controlados aleatorizados, y nos hemos hecho conocidos por nuestro compromiso con los métodos empíricos y una evaluación rigurosa. Los ensayos son una poderosa herramienta para establecer "qué funciona", y son esenciales para incrementar la innovación y el rendimiento del gobierno. Estos proporcionan un criterio contra el cual se pueden medir tanto las innovaciones "mínimas" como las "disruptivas".

Si los ensayos son un motor clave de la innovación, entonces el combustible que los impulsa son los datos.

Sin embargo, por muy poderosos que sean los ensayos, no son la única herramienta a la que recurren los gobiernos empíricos e innovadores. El análisis predictivo y el aprendizaje automático están abriendo nuevos caminos para mejorar los servicios públicos a través del estudio sistemático de variaciones sutiles y relaciones complejas en las experiencias y resultados del servicio público.

Como muestran los resultados de este informe, existen muchas maneras en que las técnicas de la ciencia de datos se pueden utilizar para mejorar la política y la práctica del gobierno. Los reguladores pueden utilizar estas técnicas para mejorar en gran medida la detección y focalización de servicios con bajo rendimiento, ya que hemos demostrado que un modelo de aprendizaje automático podía conducir a una tasa de inspección del 20% para identificar el 95% de las prácticas inadecuadas de consultorios de medicina general. A los profesionales se les puede ayudar a tomar decisiones más informadas, ya que hemos demostrado que el aprendizaje automático se puede utilizar para ayudar a identificar a aquellos niños con alto riesgo de reincidir en el sistema de servicio social, a pesar de que sus casos estuviesen cerrados. El análisis predictivo también ofrece la posibilidad de desarrollar intervenciones mucho más focalizadas; en esencia, ayudando a identificar no solo "qué funciona" en promedio, sino también qué intervención es la que mejor funciona para cada caso, como en nuestro estudio de apoyo estudiantil en el King's College de Londres.

Ha pasado una década desde que Bloomberg demostró que el uso más inteligente de los datos podría mejorar los servicios públicos de la ciudad de Nueva York a partir de la toma de mejores decisiones acerca de dónde enviar inspectores para prevenir el crimen. Desde aquellos años, el volumen y la disponibilidad de datos ha crecido en todo el mundo, al igual que nuestra comprensión de la toma de decisiones y las ciencias del comportamiento. En este informe, buscamos mostrar el potencial que esto ha comenzado a desencadenar.

David Halpern
Director Ejecutivo del equipo de Aprendizajes Conductuales

Resumen ejecutivo

La gama de técnicas que conforman la ciencia de datos (nuevas herramientas para analizar datos, nuevos conjuntos de datos y nuevas formas de datos) tiene un gran potencial de utilización en políticas públicas. Sin embargo, a la fecha, estas herramientas han estado principalmente bajo el dominio de los académicos y, donde se ha utilizado, el sector privado ha estado en la vanguardia.

Al mismo tiempo, muchas de las aplicaciones del "machine learning" (aprendizaje automático de inteligencia artificial, en adelante "aprendizaje automático") han sido de un interés bastante abstracto para el gobierno. Por ejemplo, identificar tendencias en Twitter es útil, pero no intrínsecamente valioso. Los proyectos que muestran el poder de los nuevos datos y herramientas, como la utilización de algoritmos de aprendizaje automático para derrotar a expertos humanos en el juego de Go, o identificar la prevalencia de videos de gatos que apoyan a un candidato político han estado algo ajenos a una aplicación con fines gubernamentales. Incluso cuando han sido aplicables, a menudo no se han probado adecuadamente en terreno, y las herramientas desarrolladas a partir de ellos no se han basado en la comprensión de las necesidades de los usuarios finales.

Por lo tanto, junto con muchas otras partes, en el último año hemos estado trabajando en proyectos modelo basados en el uso de la ciencia de datos, de manera que produzcan inteligencia o un entendimiento aplicable que pueda ser utilizado no solo como herramienta para comprender el mundo, o para monitorear el desempeño, sino también para sugerir intervenciones prácticas que puedan ser implementadas por los gobiernos.

Hemos realizado ocho de estos proyectos modelo enfocados en cuatro áreas: focalización de inspecciones, mejoramiento de la calidad de los ensayos controlados aleatorizados (ECA), asistencia a los profesionales para tomar mejores decisiones, y predicción de accidentes de tráfico que probablemente conduzcan a que alguien fallezca o resulte gravemente herido. Este informe cubre seis de estos ocho proyectos.

Focalización de inspecciones

- Encontramos que el **65 por ciento de las escuelas que "requieren mejoras" y las "inadecuadas" estaban dentro del 10 por ciento de las escuelas identificadas como las que se encuentran en mayor riesgo según nuestro modelo**. Aumentando esto al 20 por ciento más riesgoso, nuestro modelo captó el 87 por ciento de estas escuelas.

- Utilizando los datos disponibles en la Care Quality Commission (CQC) y otras fuentes, se puede identificar el **95% de las prácticas inadecuadas en consultorios de medicina general al inspeccionar solo una de cada cinco consultas.**

- Utilizando solamente los datos públicos del sistema de Monitoreo Inteligente de CQC, que se basa en varios indicadores clínicos, un modelo similar captaría únicamente el 30% de las prácticas inadecuadas para el mismo esfuerzo de inspección.

- También hemos creado un modelo para predecir los resultados de inspección de residencias geriátricas, pero este modelo tiene mucho menos éxito, lo que sugiere que se necesitan más datos o que las técnicas de aprendizaje automático podrían tener un uso limitado en esta área.

Mejoramiento de los ensayos controlados aleatorizados

- Anteriormente, hemos utilizado datos ECA para estudiar cómo varía la efectividad de las intervenciones para subgrupos específicos, lo que permite que las intervenciones estén mejor focalizadas.
- Estos subgrupos tendieron a definirse ampliamente por una o dos características predeterminadas, y las combinaciones de características fueron en gran medida ignoradas.
- Al aplicar algoritmos de aprendizaje automático causal a los datos ECA, podemos identificar transversalmente los impactos diferenciales de una intervención en todas las características observables, buscando garantizar que las personas reciban la mejor intervención para su caso y ayudando a prevenir fracasos.
- Reproducimos un experimento realizado en 2016 en el King's College de Londres, en el que se animó a los estudiantes a asistir a una "feria de bienvenida" a través de mensajes de texto que enfatizaban la empleabilidad o la pertenencia social, siendo la condición de pertenencia la que tuvo un mejor desempeño.
- En nuestro estudio de replicación, se asignaron participantes al azar para que recibieran uno de los brazos de mensajería asignados de forma aleatoria, o el mensaje que el algoritmo de aprendizaje automático predijo que les daría el mejor resultado en función de sus características observables.
- En nuestro primer estudio basado en estas técnicas, **encontramos que la asignación de mensajes mediante algoritmos tuvo un pequeño efecto positivo, aunque no estadísticamente significativo**, lo cual creemos se debe a una mala regulación en la complejidad del modelo. Estamos mejorando el diseño de nuestra focalización mediante la utilización de un consenso de modelos en lugar de uno solo.

Ayudar a los profesionales a tomar mejores decisiones

- Los trabajadores sociales necesitan tomar una gran cantidad de decisiones, muy a menudo con escaso tiempo e información incompleta.
- Nuestro trabajo anterior en esta área ha demostrado que el alto volumen de trabajo para los trabajadores sociales de evaluación puede influir en las decisiones que tomen.
- Al trabajar con una autoridad local, utilizamos el procesamiento de lenguajes naturales para predecir qué casos que habían sido marcados para que no se tomaran medidas adicionales, volverían a aparecer en un plazo de tres meses y devenir en un plan de protección infantil o en el ingreso de un niño a un centro de acogida.
- El análisis del texto y los datos estructurados nos permitió predecir, **8,3 veces mejor que el azar**, qué casos se remitirían al sistema.
- Utilizando solo el análisis del texto, **podemos detectar el 45,6% de los casos que retornarán con base en menos del 6% de todos los casos**, lo que permite que las intervenciones se centren con precisión para ayudar a las familias más necesitadas.
- Estamos trabajando con trabajadores sociales para desarrollar una herramienta digital que se pueda utilizar para ayudarlos a tomar decisiones más informadas.

Predicción de accidentes de tráfico serios

- Los accidentes de tráfico en East Sussex, Inglaterra, han esquivado la tendencia nacional de presentar menos incidentes de víctimas muertas y heridos graves (KSI, por su sigla en inglés).
- Podemos predecir qué accidentes resultarán en que alguien se convierta en KSI, **siendo los factores de comportamiento de los conductores, y no las condiciones del camino**, lo que más contribuye a la explicación.
- Hemos logrado superar algunos mitos, por ejemplo, con respecto a los conductores de edad avanzada y los vehículos de mercancías.
- Los motociclistas, los jóvenes y las personas de mediana edad son desproporcionadamente más propensas a estar involucradas en incidentes KSI dentro de East Sussex.

Si desea mantenerse actualizado acerca de nuestro trabajo, hallazgos y publicaciones, por favor suscríbase en:
http://www.behaviouralinsights.co.uk/subscribe.

Manténgase en contacto: escríbanos a info@bi.team.

Introducción

El Behavioural Insights Team (BIT) siempre ha formado parte del movimiento de "qué funciona", respaldando la formulación de políticas basadas en evidencia. Hasta ahora esto ha implicado en gran medida realizar ensayos controlados aleatorizados (ECA) para tratar de proporcionar evidencia de altísimo estándar acerca de si las intervenciones de política, extraídas tanto de las ciencias del comportamiento como de otras áreas, son efectivas o no. En los últimos siete años, hemos realizado más de 450 de estos ECA a través de todo el espectro de políticas, desde los impuestos, educación, desarrollo internacional hasta la atención médica, en países desde el Reino Unido a Australia, y desde EE. UU. a Siria.

Sin embargo, existen ocasiones en que un ECA no es adecuado, y hay mucho que se puede aprender acerca de cómo mejorar una política sin ejecutarla. El uso inteligente de los datos y su aprovechamiento para enfrentar problemas graves puede ayudar a mejorar la eficacia y la eficiencia de los programas existentes. Fue por esta razón que conformamos nuestro equipo de Ciencia de Datos hace poco menos de un año. En el primer año tuvimos una misión clara: salir y trabajar con los departamentos gubernamentales, con los gobiernos locales y con otras organizaciones para demostrar rápidamente que la ciencia de datos puede tener un impacto positivo en los servicios públicos, en una amplia gama de aspectos.

¿A qué nos referimos con "ciencia de datos (data science)"?

La "ciencia de datos" se ha convertido en una especie de término de moda para referirse a "cualquier cosa innovadora que utilice datos". Esto puede incluir la visualización de datos, su síntesis de variadas formas, o su utilización para predecir eventos o resultados. Nuestro enfoque se centra principalmente en el último de estos tres aspectos, que generalmente se denomina "modelación predictiva". Este generalmente consiste en recopilar un conjunto grande de datos históricos, utilizando algoritmos informáticos para encontrar patrones en los datos que serían poco prácticos o imposibles de encontrar para un humano, y luego usar estos patrones para comprender el proceso en cuestión o para predecir dónde y cuándo es probable que sucedan eventos específicos, para así poder planificar y responder a esos eventos.

Conforme a la tradición BIT, el equipo de Ciencia de Datos se constituyó con una cláusula de suspensión: teníamos 12 meses a partir del 16 de enero de 2017 para cumplir con ciertos objetivos, o el equipo se disolvería. Estos objetivos eran:

1. Producir al menos tres ejemplares a través de al menos dos áreas de política;
2. Utilizar datos públicos disponibles, datos extraídos de la web y datos textuales, para producir mejores modelos predictivos con el fin de ayudar al gobierno;
3. Probar las implicaciones de estos modelos utilizando ECA;
4. Comenzar a desarrollar herramientas que nos permitan acercar las consecuencias de nuestros datos a los formuladores de políticas y profesionales.

Entre enero y noviembre de 2017, completamos **ocho** proyectos ejemplares en salud, educación, servicio social, y seguridad vial. Trabajamos con entidades gubernamentales, locales y nacionales, y con inspectorados.

Trabajamos para determinar cómo las inspecciones de escuelas, de consultorios de medicina general y de residencias geriátricas, podrían ser más efectivas, permitiendo que los reguladores y los inspectorados utilicen el conjunto de datos, disponibles públicamente, para predecir qué instituciones tienen más probabilidades de fracasar y poder focalizar sus inspecciones en consecuencia. Mostramos que estos datos, combinados con técnicas de aprendizaje automático, tales como los árboles de decisión con potenciación de gradiente, pueden superar significativamente tanto una focalización de inspección aleatoria como sistemática. No obstante, en el caso de las residencias geriátricas, el impacto no es grande, lo que nos sirvió para articular los límites del aprendizaje automático, o al menos de los datos actuales. Estamos muy entusiasmados de estar trabajando con Ofsted para poner los aprendizajes de este trabajo en acción, y para mejorar el trabajo que ya ha sido realizado en este ámbito.

Trabajando con el King's College de Londres (KCL) como parte de nuestro proyecto KCLXBIT durante el año académico 2016-2017, hemos utilizado técnicas de aprendizaje automático causal para identificar varios grupos dentro de nuestra muestra que obtuvieron un beneficio particularmente alto, o particularmente bajo, a partir de los mensajes que enviamos a los estudiantes para que se inscribieran en sociedades, o para que utilizaran entornos de aprendizaje en línea. Esto nos permitió elaborar la mejor intervención posible a cualquier individuo, con base en sus características y su respuesta a la prueba.

Sin embargo, la calidad de una predicción depende de las decisiones que se puedan tomar con base en ella. Al asociarnos de nuevo con KCL, pudimos probar las predicciones de nuestro modelo en el mundo real. Descubrimos que, si bien el algoritmo mejoró la orientación de los mensajes en una pequeña proporción, el efecto no fue significativo. Estamos iterando el proceso que utilizamos para focalizar las intervenciones y continuaremos probando y adaptando nuestros algoritmos como lo hemos hecho con intervenciones desde nuestros inicios.

Además de desarrollar intervenciones focalizadas y puntuales, es importante que la ciencia de los datos se vuelva útil para los profesionales y especialistas en su día a día. En el proyecto en el cual trabajamos con una autoridad local para predecir futuras intensificaciones de casos concernientes al servicio social para niños, llevamos a cabo un trabajo cualitativo basado en el aprendizaje automático, para ayudar a encontrar la mejor manera de presentar los hallazgos a una audiencia de trabajadores sociales (a menudo justificadamente escépticos), de modo que pudieran ser útiles para ellos. Al unir el elemento humano de la práctica, con el aprendizaje automático, estamos ahora desarrollando una herramienta digital para ayudar a los trabajadores sociales a utilizar estos conocimientos en tiempo real.

Una lección importante extraída de esta experiencia, es que los proyectos tienden a tener más éxito cuando hay un problema predictivo claro; datos de alta calidad, a gran escala y preferentemente a nivel individual; certeza suficiente como para implementar los hallazgos en la práctica; y autorización ética y legal clara. Recomendamos realizar el análisis inicial de proyectos utilizando el Data Maturity Framework[2] del Centro de Ciencia de Datos y Políticas Públicas de la Universidad de Chicago.

El uso de la ciencia de datos en la política es algo nuevo, excitante, y con un gran potencial que aún debe ser concretado. Queda mucho trabajo por hacer para poder desarrollar conocimientos procesables a partir de estos datos de forma que, posteriormente, puedan ser utilizados en la práctica. Creemos, sin embargo, que los proyectos aquí descritos representan un paso en la dirección correcta.

Identificación del bajo desempeño

Inspecciones CQC de consultorios de medicina general

Los consultorios de medicina general tienen un profundo impacto en nuestra salud: son el primer punto de contacto para muchas asistencias médicas no urgentes, y un canal vital para las campañas de salud pública. Sin embargo, debido a los altos números de casos, la gran variedad de condiciones atendidas, y los distintos niveles de destreza profesional, algunos consultorios en el Reino Unido no cumplen con el estándar requerido.

El Comité de Calidad de la Atención, o Care Quality Commission (CQC) tiene el deber de inspeccionar los consultorios de medicina general, pero cuenta con recursos limitados en términos de personal y tiempo, las inspecciones, a su vez, representan una carga para los consultorios. CQC califica las prácticas de los consultorios en una escala de cuatro puntos, desde excepcional hasta inadecuado. Estas inspecciones se centran no solo en los resultados clínicos (es decir, la salud del paciente) sino también en cinco dominios amplios: si las prácticas son seguras, (clínicamente) efectivas, cuidadosas, receptivas, y bien dirigidas.

Nosotros investigamos cómo podríamos mejorar los sistemas de focalización existentes intentando predecir las clasificaciones de inspecciones pasadas utilizando datos anteriores a estas inspecciones. En particular, analizamos si es que las prácticas obtuvieron un resultado positivo (una calificación "sobresaliente" o "bueno") o uno menos deseable ("debe mejorar " o "inadecuado"). Encontramos que el 11 por ciento de las prácticas recibieron calificaciones bajas, incluido el 2 por ciento que recibió clasificaciones inadecuadas, lo que provoca una re-inspección automática dentro de seis meses y debería derivar en medidas correctivas por parte de los consultorios.

Utilizamos una amplia variedad de datos públicos disponibles para este proyecto. Además de indicadores clínicos publicados por CQC, utilizamos datos de la Oficina Nacional de Estadística del Reino Unido, y datos acerca del tipo y número de medicamentos recetados en los consultorios. También extrajimos el texto de reseñas dejadas por pacientes en el sitio web de NHS Choices (ver recuadro de abajo).

"Raspado" o extracción de información de sitios web

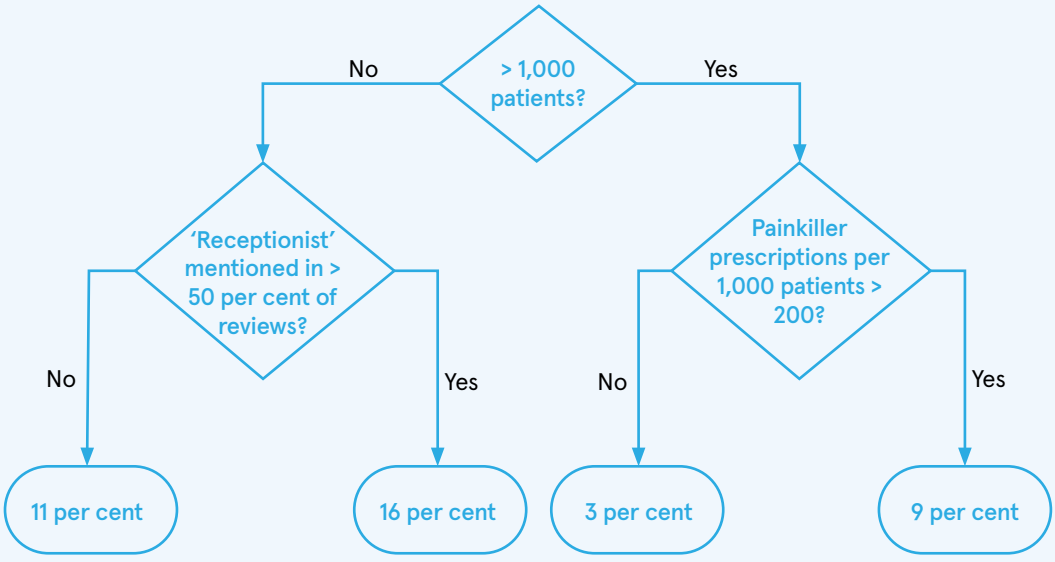
El "raspado" ("scraping" por su nombre en inglés) o extracción es el proceso de recopilación y conversión de datos de sitios web que pueden ser utilizados en el análisis. En nuestro trabajo acerca de consultorios de medicina general, utilizamos la extracción para obtener las clasificaciones por estrellas y las reseñas para todas las prácticas que tenían calificaciones dentro del sitio web de NHS Choices. Esto nos permitió incluir algunos aspectos de la opinión pública acerca de cada consultorio en nuestro análisis, los que supusimos que podrían ayudarnos a identificar el bajo desempeño. Las reseñas de pacientes (o usuarios) a menudo son valiosas fuentes de información y se encuentran disponibles de forma pública, pero normalmente se hallan solo en sitios web y no en un conjunto de datos preparado para un análisis. La extracción nos permite acceder al valor oculto en estos sitios web.

La imagen de arriba a la izquierda muestra la forma en que la página web normalmente aparece cuando alguien está navegando en https://www.nhs.uk en busca de reseñas de consultorios. A la derecha está el código HTML subyacente que genera la vista a la izquierda. Nosotros escribimos programas que descifran este código HTML y extraen las partes que nos interesan. Por ejemplo, para cada reseña, los programas debían extraer la porción escrita en formato libre, resaltada arriba en amarillo. También descifrarían cuántas estrellas le habría otorgado al consultorio cada usuario. El código HTML y los sitios web, en general, no suelen escribirse para facilitar la extracción de información, por lo que fue necesario escribir programas personalizados para interpretar cada elemento individual en cada página del sitio web de NHS Choices que nos interesaba.

La utilización de un gran conjunto de datos de recetas médicas junto con el texto de 99.644 reseñas en línea de NHS Choices nos permitió entender los matices de la conducta de los médicos que, de otra manera, serían indetectables. Para captar este comportamiento contingente, analizamos los datos usando árboles de decisión con potenciación de gradiente (ver recuadro a continuación).

Utilización de árboles de decisión como modelos estadísticos

Un árbol de decisión es un objeto flexible que puede describir muchos procesos del mundo real y que permite utilizar diferentes datos como predictores según el contexto. En este ejemplo, el árbol de decisión predice la posibilidad de un resultado de inspección deficiente:[3]

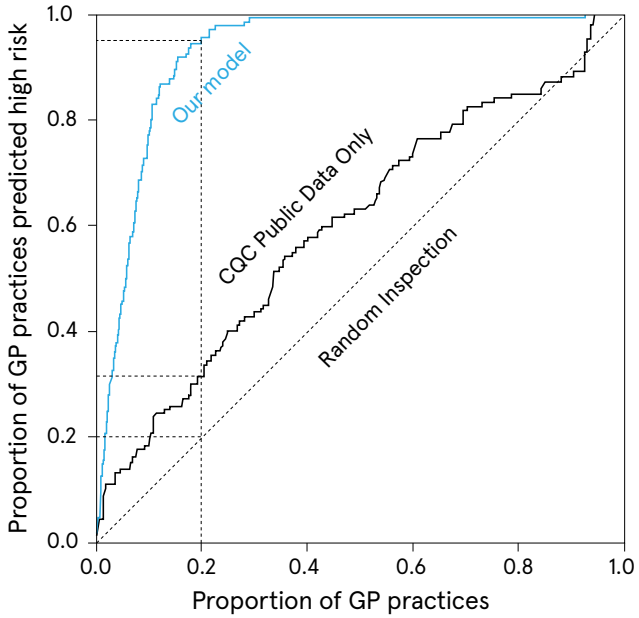


Nótese cómo los datos de texto se utilizan para consultorios más pequeños, y los datos de recetas médicas para consultorios más grandes. Este conocimiento del contexto no se encuentra en muchas otras técnicas, y nos permite alcanzar un alto poder de predicción en problemas grandes y complejos, como en este caso.

Cualquier modelo estadístico se puede mejorar mediante un proceso de refocalización iterativa en los casos donde sus predicciones fueron menos precisas. A modo de analogía, imagínese a un maestro que le asigna tareas a un niño: si el maestro establece un trabajo de seguimiento, es mejor que cubra aquel material que el alumno no pudo abordar correctamente que aquellos tópicos con los que el alumno no tuvo problemas. Este proceso se denomina "potenciar", y los "árboles de decisión con potenciación de gradiente" son el resultado de aplicar este proceso a un modelo de árbol de decisión. El resultado de esto es una gran colección de diferentes árboles de decisión, que luego se promedian para producir predicciones.

Los resultados del modelo fueron sorprendentes (ver Figura 1). Pudimos identificar casi todas (95%) las clínicas inadecuadas (aquellas que recibieron la calificación de inspección más baja) al inspeccionar solo el 20 por ciento de clínicas que nuestro modelo identificó como las más riesgosas. Por el contrario, solo pudimos identificar el 30 por ciento de las prácticas inadecuadas al restringir nuestros datos de entrada a las fuentes publicadas por CQC, lo que demuestra el gran poder predictivo de estas fuentes de datos adicionales.

Figura 1: La curva de ganancia para el modelo de consultorios de medicina general de CQC.



Fue más difícil detectar los consultorios de medicina general calificados como "requiere mejoras": pudimos detectar alrededor del 55 por ciento de los consultorios "inadecuados" y "requiere mejoras" al inspeccionar el 30 por ciento de estos. Utilizando sólo datos públicos CQC, pudimos detectar solamente el 40 por ciento.

Dentro de nuestro modelo, las decisiones se basan más en los datos de texto que en cualquier otra fuente de datos. El modelo parece capaz de extraer palabras y frases particulares que aparecen en las reseñas para indicar una práctica buena o mala.

Ejemplo de un texto asociado a un buen consultorio de medicina general

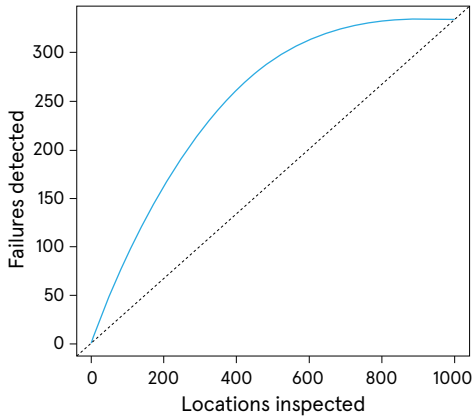
> *"Como resultado de mi asistencia y la excelente atención de enfermería, mi condición ha mejorado notablemente, por lo que estoy muy agradecido. Cuando he acudido a un Médico General también he recibido una atención excelente y amable."*

¿Cómo medimos el rendimiento de un modelo predictivo?

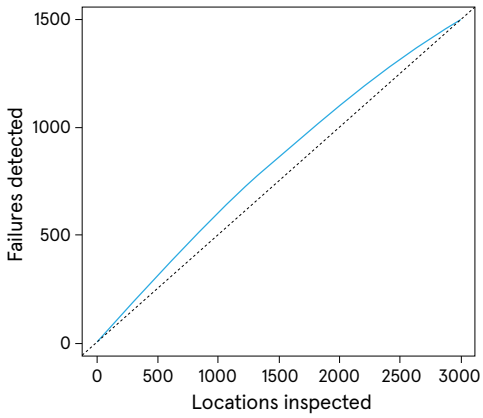
Podemos utilizar datos históricos para comparar el resultado modelado con el resultado real y producir tasas de falsos positivos (donde el modelo predice un resultado que no ocurre) y falsos negativos (donde el modelo no predice ningún resultado cuando este sí ocurre). Desafortunadamente, estas dependen de dónde establecemos la sensibilidad del modelo, y pueden ser intercambiables entre sí: un modelo con alta sensibilidad a menudo predecirá un resultado y tendrá una tasa alta de falsos positivos y una tasa baja de falsos negativos. Lo contrario es cierto para un modelo con baja sensibilidad.

Para evaluar el rendimiento de un modelo predictivo independientemente de esta decisión de dónde establecer la sensibilidad, observamos el número de casos positivos correctamente identificados por sobre todos los valores posibles de este umbral, y lo trazamos contra el número de casos donde el modelo predice un resultado positivo (independientemente de los hechos). Para un modelo fuertemente predictivo, esta curva se arqueará fuertemente, como en el ejemplo de abajo a la izquierda. Para un modelo débilmente predictivo, esta curva estará ligeramente por encima de la línea diagonal que representa el comportamiento promedio de seleccionar casos al azar. Esta curva se conoce generalmente como "curva de ganancia".

Fuerte



Débil



Los indicadores clínicos fueron menos predictivos de lo que podríamos haber esperado, en parte porque omiten algo que el texto y los indicadores demográficos más detallados sí recogen: el estándar de atención.

En algunos casos, las asociaciones fueron sorprendentes, por ejemplo, cuando los consultorios recetaban menos dosis de ciertos medicamentos, era probable que se necesitara mejorar. Esto muestra cómo se puede utilizar un proceso riguroso basado en datos para evitar el pensamiento motivado, y resalta el potencial de la focalización de inspecciones para reducir la carga cognitiva mediante el filtrado automático de muchos factores distintos.

¿Qué podemos aprender de esto? Descubrimos que es factible aprender de los resultados de inspecciones anteriores para dirigir futuras inspecciones de manera más eficiente. También es posible encontrar más consultorios de medicina general que son inadecuados o que requieren mejoras con el mismo número de inspecciones.

Residencias geriátricas

También nos volcamos a la predicción de las calificaciones de las residencias geriátricas, utilizando un enfoque similar al que utilizamos para los consultorios de medicina general. Desafortunadamente, existen significativamente menos datos públicos acerca de residencias geriátricas: los indicadores CQC no son públicos, y los datos de la fuerza de trabajo son de recopilación opcional y solamente se pueden obtener a nivel regional, y no a nivel de centro asistencial mediante Skills for Care[4].

Otra dificultad con las residencias geriátricas es que las reseñas son abrumadoramente positivas, con más del 99 por ciento galardonadas con tres estrellas o calificaciones en promedio más altas. Esto ocurre tal vez porque las quejas serias contra las residencias geriátricas tienden a ser hechas a los reguladores; mientras que, con los consultorios de medicina general, una gran cantidad de quejas se centraron en problemas de menor nivel.

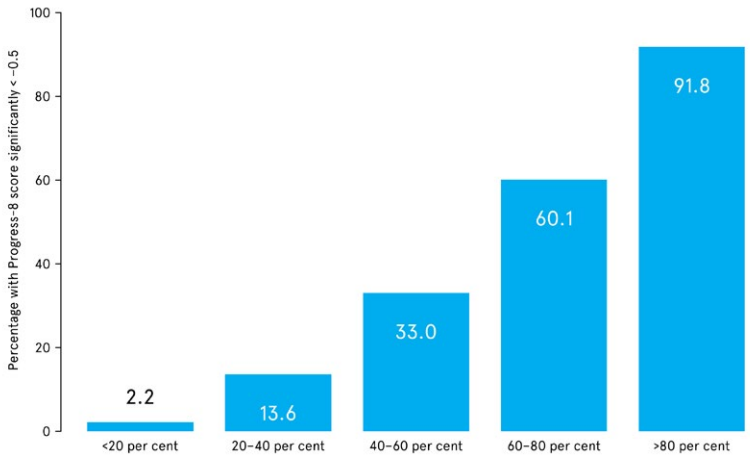
A pesar de esto, aún podemos detectar aproximadamente el 45 por ciento de las residencias que "requieren mejoras" o que son "inadecuadas" si inspeccionamos el 30 por ciento de las residencias, lo que demuestra que podemos alcanzar un poder predictivo razonable incluso con menos información disponible o datos relevantes. La tasa de detección equivalente para los consultorios de medicina general fue del 55 por ciento.

Progress-8

Progress-8 es un indicador clave de rendimiento escolar que mide el progreso que realiza un alumno entre el final de la escuela primaria y el final de Key Stage 4 (GCSE; sistema de certificados en Gran Bretaña) en comparación con otros alumnos con un rendimiento escolar de primaria similar. Se usa como un estándar de logros y progresos mínimos5, lo que significa que identifica las escuelas donde los estudiantes están teniendo peores resultados de los que "deberían obtener" con respecto a su desempeño anterior en su vida escolar.

Progress-8 es de interés para Ofsted y sirve como un punto de comparación útil para las inspecciones mismas. Encontramos que Progress-8 es altamente predecible (con una relación casi perfecta) con una sola variable: el porcentaje de alumnos cursando Key Stage 4 elegibles para la bonificación al alumnado (ver Figura 2). La misma relación no es cierta para los resultados de inspección Ofsted.

Figura 2: Porcentaje de escuelas con puntajes de Progress-8 que no cumplen con el estándar mínimo por porcentaje de alumnos cursando Key Stage 4 elegibles para la bonificación al alumnado.



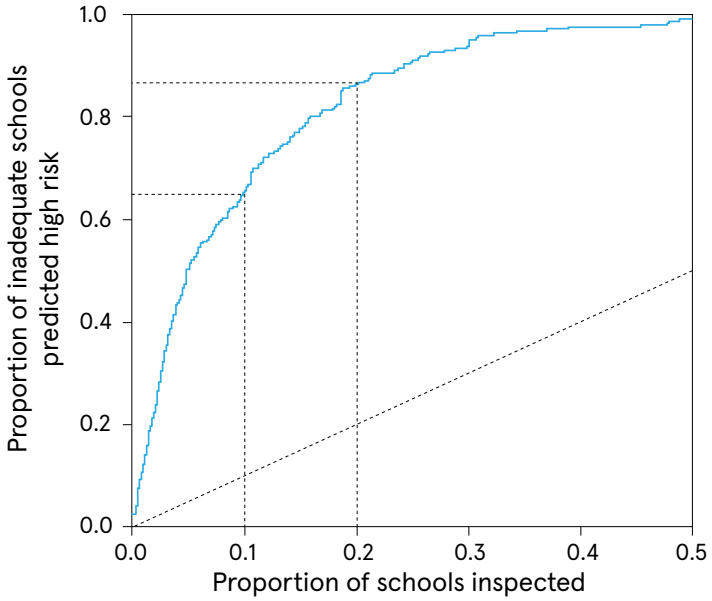
Inspecciones escolares Ofsted

Las inspecciones escolares son vitales para garantizar que se mantengan los estándares en todo el sistema escolar estatal. Además, sabemos por investigaciones previas que las intervenciones en escuelas con calificaciones bajas, mejoran los resultados.[6] Ofsted ya es líder en la configuración, validación, uso de escalas de calificación, y de modelos estadísticos para focalizar inspecciones. Intentamos predecir los resultados de las re-inspecciones de la Sección 5 de escuelas previamente calificadas como "buenas" mediante el aprendizaje automático, debido a que Ofsted tiene cierto grado de control sobre el calendario de estas re-inspecciones.[7]

Utilizamos datos disponibles públicamente desde el año anterior a la realización de una inspección, incluidos los datos de la fuerza laboral, datos del censo del Reino Unido y de privaciones del área local, tipo de escuela, datos financieros (fuentes de financiación y gasto), datos de rendimiento (Key Stages ("Etapas Clave") 2 para escuelas primarias y Key Stages 4 y 5 para escuelas secundarias) y respuestas de Ofsted Parent View para preguntas de la encuesta. Descubrimos que el 65 por ciento de las escuelas clasificadas como "requiere mejoras" e "inadecuadas" estaban dentro del 10 por ciento de las escuelas identificadas como en mayor riesgo según nuestro modelo. Aumentando esto al 20 por ciento más riesgoso, nuestro modelo capturó el 87 por ciento de estas escuelas (ver Figura 3).

Nótese que el modelo que presentamos en este trabajo se basa en datos anteriores a 2015. Este no es el modelo que utiliza actualmente Ofsted.

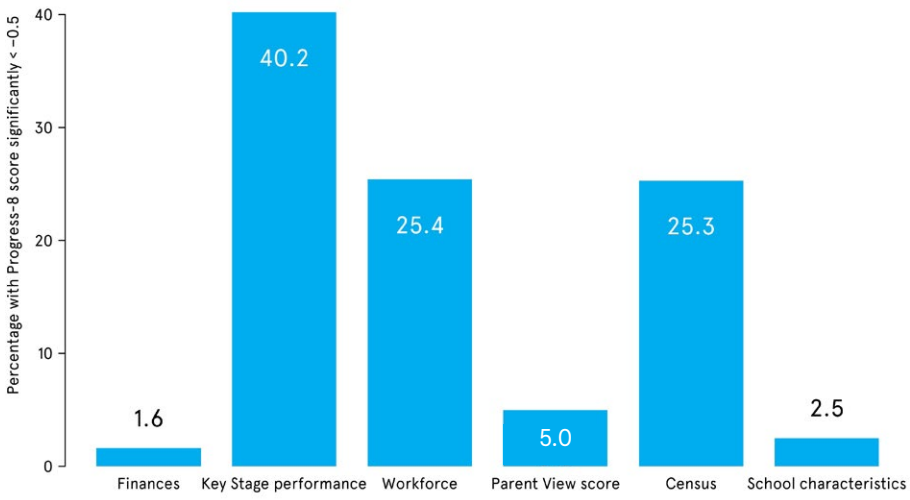
Figura 3: La curva de ganancia para nuestro modelo.



Las diferencias generales entre los dos grupos de escuelas son bastante sutiles: el poder predictivo del modelo proviene de las interacciones entre las variables, en lugar del poder de cualquier variable por sí misma. Por ejemplo, tener buenas reseñas y una baja rotación del personal puede ser mucho más predictivo que tener cualquier indicador por sí solo. Esta perspectiva de matices, en comparación con los resultados que vimos al predecir los puntajes de Progress-8, sugiere que las inspecciones Ofsted están captando algo más profundo acerca de las escuelas que solo el rendimiento educativo. La Figura 4 muestra las categorías más importantes para predecir el rendimiento Ofsted.

Algunos de los factores más importantes están relacionados con las finanzas de la escuela, que se publican a través del censo de fuerza laboral de la escuela. Un predictor interesante es la falta de datos. Si una escuela no entregaba cierto tipo de información a, por ejemplo, el Departamento de Educación, era más probable que no aprobara las inspecciones, a diferencia de las escuelas que sí proporcionaron esta información. Esto requiere una mayor investigación, pero podría indicar una posible insuficiencia administrativa en estas escuelas.

Figura 4: Proporciones de poder predictivo contribuidas al modelo por las diversas fuentes de datos.



Asistencia en la toma de decisiones

Asistir a los trabajadores sociales, especializados en infancia, para procesar casos con precisión

Los trabajadores sociales tienen uno de los trabajos más difíciles en el sector público. Los trabajadores sociales que realizan evaluaciones de manera individual pueden manejar hasta 50 casos a la vez. Ellos se encargan de evaluar rápidamente si un niño está en riesgo y en necesidad de protección, y en última instancia, con el apoyo de los tribunales, si un niño necesita ser internado en una institución.

Ellos deben cumplir con esto haciendo frente a una escasez feroz de recursos y tiempo. A diferencia de otros campos en que se trabaja con niños, como la educación, acá no existe una medida de resultados clara (como las calificaciones) mediante la cual evaluar a los trabajadores sociales, es por ello que las fallas individuales en su toma de decisiones están sujetas a un profundo escrutinio.

Nuestro trabajo en el servicio social infantil

Hemos estado desempeñándonos en servicio social infantil durante varios años, centrándonos en su "puerta de entrada": las decisiones que se toman cuando se contacta al Concejo por primera vez acerca de un posible incidente. Nuestro principal hallazgo, desde la perspectiva de las ciencias del comportamiento, fue que los trabajadores sociales de los equipos de evaluación toman cientos o miles de decisiones a lo largo de su carrera, pero reciben relativamente poca información acerca de lo que sucede posteriormente; sus casos o salen del sistema de servicio social o son derivados a otro equipo. Esta falta de retroalimentación hace que sea difícil para los trabajadores sociales aprender de manera eficiente acerca de sus decisiones anteriores.

También hemos investigado los factores asociados con la trayectoria del niño a través de la atención: ¿recibió evaluaciones múltiples? ¿Pudo cerrarse su caso? ¿Ha ingresado nuevamente al sistema de servicio social? Un hallazgo clave fue que, en los casos de las tres autoridades locales con las que trabajamos, cuando el primer contacto con el niño se realizó durante un fin de semana, resultó ser menos probable que este progresara en el sistema de servicio social que cuando se hacía en un día hábil.

Al evaluar el Project Crewe del Fondo de Innovación para la Obra Social del Departamento de Educación[8], realizamos nuestro primer ECA en servicio social, que se publicó en 2017. A pesar de que esta prueba fue a pequeña escala, pudimos extraer ideas a través del análisis cualitativo realizado sobre anotaciones de casos de trabajadores sociales, que fueron codificados manualmente en función de factores protectores y factores perjudiciales dentro del texto. Esto proporcionó un puntaje de riesgo basado en la cantidad de factores presentes, y nos permitió comprender cómo cambió el riesgo en el transcurso de la evaluación.

¿Es posible predecir el agravamiento de casos ya cerrados?

En 2017, nos embarcamos en un proyecto para mejorar nuestro trabajo previo en el servicio social infantil utilizando técnicas de análisis de datos de vanguardia tomadas de los campos del "machine learning" (aprendizaje automático de inteligencia artificial, en adelante "aprendizaje automático") y el procesamiento del lenguaje natural. Sobre la base de nuestro análisis anterior, investigamos el proceso de toma de decisiones de los trabajadores sociales para cerrar un caso y recomendar "no adoptar medidas ulteriores". Nuestro objetivo principal era explorar hasta qué punto estas técnicas y herramientas podrían proporcionar conocimientos que pudiesen resultar prácticos para los trabajadores sociales en terreno. Fue por ello por lo que consultamos a los trabajadores sociales acerca de los hallazgos de la ciencia de datos, realizando seis entrevistas semiestructuradas con los trabajadores sociales actuales para comprender su interpretación de los datos.

El principal problema predictivo fue el siguiente: dado el texto referencial inicial y la evaluación, así como los datos estructurados relacionados con el caso, ¿podíamos predecir si este sería vuelto a remitir, agravándose, en caso de que se cerrara?

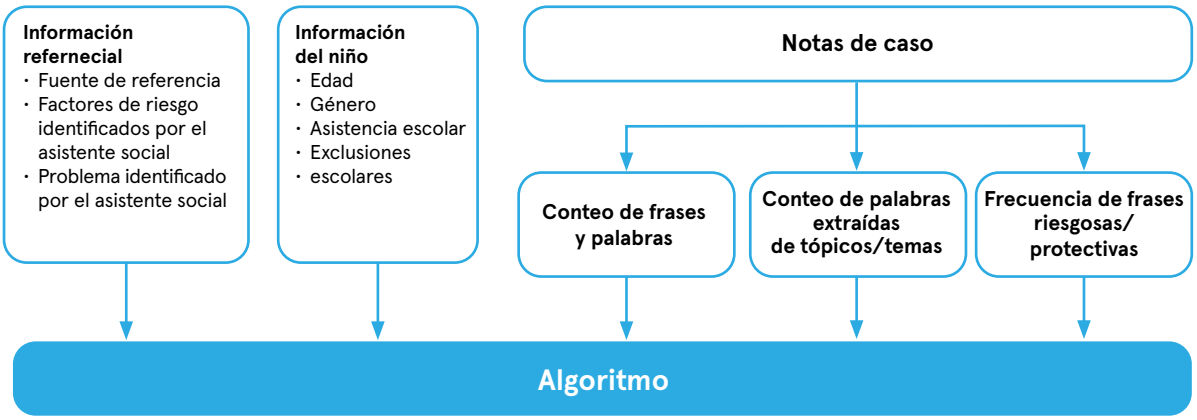
De los 11.000 casos de niños que fueron remitidos al sistema durante los dos años que analizamos, 5.117 se cerraron inmediatamente, sin más tramitación. Tratamos de determinar cuáles de estos casos tendrían propensión a volver al sistema de servicio social más tarde (1.693 niños) y, de aquellos, cuáles regresarían convertidos en casos graves (583) como los casos en que exista un requerimiento de protección infantil (niño identificado en situación de riesgo vital, o en riesgo de daño inmediato).

Cómo procedimos

Si bien parte de la información sobre la remisión inicial y el niño estaba disponible para el análisis, los datos más importantes fueron las notas de los casos de los trabajadores sociales, que son notas en formato libre hechas por trabajadores sociales bajo rúbricas como "Análisis de evaluación" para documentar sus hallazgos y los detalles de cada caso. Las notas de casos no están, en gran medida, estructuradas: existen muchos usos y costumbres particulares con respecto a cómo se registra la información y cómo se describen los problemas. Esto las hace difíciles de analizar utilizando técnicas tradicionales de análisis de texto, como el conteo de la frecuencia con la que aparecen ciertas palabras o frases en el texto.

Por lo anterior, analizamos el texto utilizando modelos temáticos (ver recuadro a continuación) para extraer tópicos del texto. Estos tópicos, junto con datos estructurados más tradicionales relevantes al caso, se incorporaron a un algoritmo de aprendizaje automático (una máquina de potenciación de gradiente; ver recuadro en la página 11). Este algoritmo de aprendizaje automático se utilizó para identificar los casos que tenían un alto riesgo de regresar al sistema de servicio social después de ser cerrados. Las diversas entradas están ilustradas en la Figura 5.

Figura 5: Las entradas del algoritmo de aprendizaje automático utilizado para detectar casos cerrados agravados.



Modelación de tópicos

La modelación de tópicos es el proceso automatizado mediante el cual se busca encontrar grupos de palabras que comúnmente aparecen juntas (dentro de tópicos). Por lo general, generan hallazgos más fáciles de comprender para los formuladores de políticas y otros profesionales que otros métodos de análisis de texto, porque los tópicos tienden a ser más coherentes que listas de palabras o frases consideradas de manera independiente.

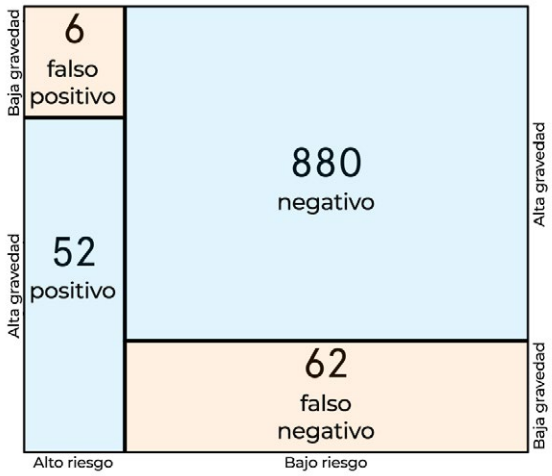
Nos apoyamos en una técnica llamada "modelación de tópicos estructurales"[9], que mejora la coherencia de estos grupos basándose en el hecho de que la prevalencia del tópico varía entre los hablantes y lo hablado, por ejemplo, el estudio original descubrió que la propensión a discutir acerca de la guerra de Irak variaba según la ideología política de cada blog. En nuestro contexto de trabajo social, los temas podían variar según las características del niño, la experiencia del trabajador social, los riesgos identificados y quién remitía el caso a los servicios sociales.

Para ayudarnos a interpretar estos datos y situar los hallazgos del algoritmo en experiencias del mundo real, buscamos retroalimentación realizando una serie de entrevistas semiestructuradas con cuatro trabajadores sociales y dos gerentes de equipo de la autoridad local. Esto era importante para este proyecto porque los trabajadores sociales necesitaban comprender las razones detrás de las sugerencias del algoritmo para cada caso en particular, de manera que pudieran combinar estos conocimientos con su propia experiencia.

Lo que encontramos

Utilizando esta combinación de datos estructurados y no estructurados, nuestro algoritmo pudo identificar un pequeño grupo de casos (6 por ciento) que fueron cerrados en situación de "alto riesgo". Este conjunto de alto riesgo contenía casi la mitad de los casos que luego regresarían agravados, con muy pocos (0,6 por ciento) falsos positivos (en este caso, un falso positivo es un caso de alto riesgo que en realidad no regresa ni se agrava). La Figura 6 ilustra el rendimiento esperado del modelo en 1.000 casos cerrados. Un caso es de "alta gravedad" si es vuelto a ser referido y se agrava; de lo contrario, es de "baja gravedad".

Figura 6: Distribución esperada de positivos y falsos positivos y negativos para 1.000 casos no clasificados previamente.



'Riesgo' se refiere a la predicción y 'severidad' del modelo con respecto al resultado real. Por ejemplo, un "falso negativo" es un caso que se agravó pero que el modelo clasificó como de "bajo riesgo".

Siempre existe una compensación entre la tasa de falsos positivos y la tasa de falsos negativos; sin embargo, en este caso es difícil disminuir los falsos negativos sin que ocurra un gran aumento de falsos positivos; lo que crearía un exceso de trabajo innecesario. Existen muchas razones por las que este grupo de resultados es difícil de predecir. Algunas familias, temerosas de los trabajadores sociales, encubren problemas o pretenden ser más obedientes de lo que son para evitar ser asociadas con el estigma de estar involucradas con servicios sociales infantiles. Además, las entrevistas con los trabajadores sociales revelaron que encontrar evidencia real del problema también puede ser un desafío dentro del breve marco de tiempo que les es asignado.

El tópico extraído del texto que fue más útil para identificar casos que probablemente reincidirían, se ejemplifica con el siguiente extracto anónimo de las notas de caso:

> "Yo, [nombre del trabajador social], no recomiendo que la CSC tome medidas adicionales [servicio social infantil]. Siento que los niños no han mostrado ningún tipo de preocupación o cambios de comportamiento a partir de la discusión verbal. Siento que, aunque la CSC se ha involucrado con la familia a lo largo de los años, [Nombre] puede ser una persona de confianza para [niño, medio hermano y hermano]".

Al revisar múltiples ejemplos, este tópico parece corresponder a notas de casos donde el trabajador social siente que es necesario invertir tiempo en justificar por qué se está cerrando un caso, a menudo debido a una evidencia insuficiente o al consentimiento de la familia. Esto refleja algunos de los desafíos que enfrentan los trabajadores sociales, donde es posible que no exista suficiente evidencia para fundamentar los problemas que el trabajador social sospecha pueden estar presentes, o la evidencia no es lo suficientemente clara, o las familias pueden estar ocultando el alcance de los problemas.

> *"Mi intuición, y la de otros trabajadores sociales, es que esto volverá a ocurrir, porque no hemos cambiado efectivamente la dinámica de la situación... Pero sabemos que por el momento no es suficiente, por ejemplo, para acudir a la corte."*
>
> *(Un trabajador social entrevistado)*

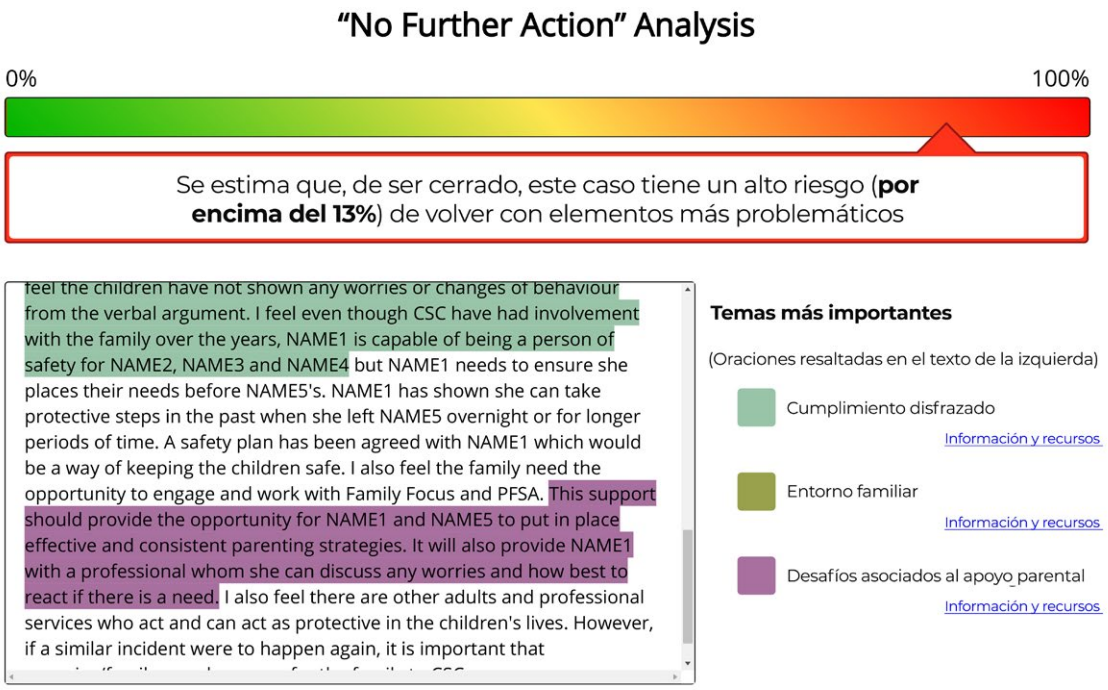
En otros casos, los trabajadores sociales utilizan este lenguaje cuando creen que una familia puede beneficiarse de su ayuda, pero la familia se niega a cooperar o aceptar ayuda del servicio social. Los trabajadores sociales entrevistados tenían claro que tenían que tomar una decisión basada en la evidencia disponible en ese momento, pero apreciarían una herramienta que los ayudara a superar sus sesgos internos o a identificar si los casos tenían la impronta de estar en alto riesgo de volver a ser remitidos.

¿Qué es lo que sigue? Un asistente de toma de decisiones digital

Actualmente, estamos desarrollando una herramienta digital que permitirá a los trabajadores sociales identificar el riesgo estimado, para un caso particular, según el algoritmo. Esto nos permitirá lograr que nuestros hallazgos sean accesibles y útiles para los trabajadores sociales en su práctica cotidiana. Siguiendo los comentarios de los trabajadores sociales y gerentes, la mejor utilización del algoritmo podría ser proporcionar una base de evidencia para justificar una mayor inversión de tiempoen casos potencialmente riesgosos cuando la decisión no es lo suficientemente clara, y el trabajador social normalmente no tendría suficientes motivos para mantenerlos abiertos.

La primera versión de esta herramienta, que esperamos sea probada con la autoridad local participante a principios de 2018, permitirá que un asistente social o gerente copie el texto de las notas de casos de un individuo y lo pegue en la herramienta, que luego analizará el texto y proporcionará una estimación de riesgo. Además de la calificación de riesgo, que será de color rojo, ámbar o verde, la herramienta identificará qué fragmentos de oraciones y tópicos en el texto son más indicativos de riesgo. El diseño básico de la herramienta se puede apreciar en la Figura 7.

Figura 7: Diseño de la herramienta prototipo de evaluación de riesgos para servicio social para niños.



El diseño de esta herramienta se basa en nuestras entrevistas previas con trabajadores sociales, la herramienta estará sujeta a etapas adicionales de retroalimentación y adaptación. Es clave que esta herramienta ofrezca una visión práctica y sea extremadamente fácil de incorporar en el trabajo diario de los trabajadores sociales, por lo que estamos siendo particularmente cuidadosos en incorporar a los trabajadores sociales y sus gerentes en el proceso de diseño.

Diseño de intervenciones focalizadas

En el espíritu de "qué funciona", podemos usar la ciencia de datos para comprender qué hace que una persona realice o no un comportamiento indeseable. Esto nos permite diseñar intervenciones enfocadas en aquellos individuos que presentan un mayor riesgo, así como a maximizar la eficiencia con la que efectuamos nuestras intervenciones. Por otro lado, podemos evitar la focalización en personas que difícilmente realizarán el comportamiento que queremos detener.

Reducción de la tasa de accidentes de tráfico en East Sussex

Si bien las carreteras del Reino Unido son muy seguras según los estándares internacionales, desde 2015 hasta 2016 se registró un aumento del 4 por ciento en los accidentes fatales[10], donde los accidentes de transporte son una de las principales causas de muerte en personas de entre 20 y 34 años.

Existe una gran cantidad de datos disponibles con los cuales diseñar intervenciones, además de los de los conductores y las víctimas, un oficial presente en una colisión forma y registra una opinión acerca de cuáles fueron los "factores contribuyentes" que condujeron a la colisión. Nosotros podemos utilizar estos datos para formar una imagen de qué comportamientos provocan los accidentes más graves y dónde alguien fallece o se lesiona gravemente (KSI, por su sigla en inglés). Los factores contribuyentes más comunes incluyen "descuidado, temerario o apresurado", "bajo el efecto del alcohol" y "superar el límite de velocidad".

Utilizamos todas estas fuentes de datos como base para un modelo predictivo de la gravedad de una colisión. El objetivo no era producir predicciones en sí mismas: predecir la gravedad de una colisión una vez que ha sucedido no es de gran ayuda; en cambio, este modelo puede decirnos la importancia de cada uno de sus datos para predecir la gravedad de una colisión, de modo que podamos enfocarnos en aquellas conductas que tienen un impacto real.

En especial, ahora podemos cuantificar hasta qué punto la gravedad de la colisión se ve influenciada por el comportamiento que causó la colisión. Esto se muestra en la Figura 8.

Figura 8: Importancia predictiva relativa de categorías de datos.

El total de las barras suma un 100 por ciento, y cada una indica el grado total de importancia (similar a la correlación, pero más generalizada) para todas las variables dentro de esa categoría. Los colores son solo para fines cosméticos.

El comportamiento del conductor es el elemento más predictivo en la gravedad de la colisión; más que las características del camino (incluyendo el límite de velocidad y las condiciones climáticas predominantes), la edad y el sexo del conductor, y qué tan lejos estaba este de casa.

Otra revelación importante es que el propósito del viaje (desplazamiento, trabajo o personal) no es un predictor fuerte de la gravedad de la colisión[11]. Esto significa que es probable que no sea tan efectivo enfocarse en conductores ocupacionales, como lo sería abarcar todos los tipos de viajeros.

Armados con el conocimiento de cuáles son los comportamientos más peligrosos, podemos enfocarnos en las personas que incurren en estos con mayor frecuencia. Dentro de East Sussex, estos se dividen en tres categorías.

La primera categoría son los conductores jóvenes. Se sabe que los conductores menores de 25 años son un grupo de alto riesgo a nivel nacional, y en East Sussex encontramos que los conductores varones y jóvenes estaban significativamente sobrerrepresentados en los datos de colisión en comparación con el número de licencias de conducir: el 16,8 por ciento de las colisiones KSI fueron provocadas por ellos, pero estos solo poseían el 5,6 por ciento de las licencias. Los conductores jóvenes causaron accidentes graves de forma desproporcionada cuando se encontraban a menos de tres millas de su hogar, y estar bajo la influencia del alcohol fue mucho más común para este grupo de edad que para otros.

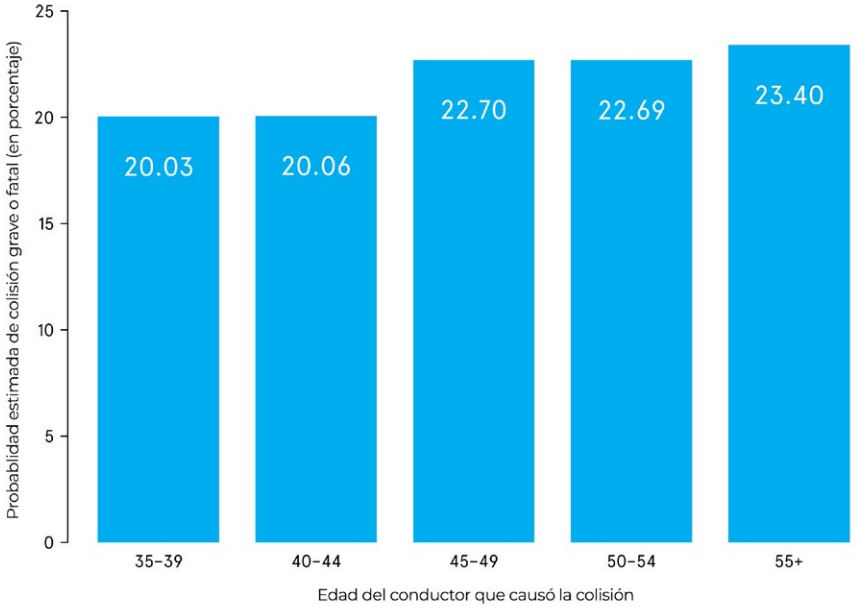
Si un conductor joven provoca una colisión, el límite de velocidad y la gravedad de la colisión son casi estadísticamente independientes. Esto no es cierto para ninguna otra categoría de conductor, y sugiere fuertemente que la alteración de los límites de velocidad no es una táctica efectiva para reducir la severidad de colisión entre conductores jóvenes.

La segunda categoría son los motociclistas. A nivel nacional, una milla recorrida en una motocicleta tiene doce veces más probabilidades de resultar en fallecimiento o lesiones graves que una milla recorrida en un automóvil.[12] En East Sussex, esta cifra es 26 veces mayor.[13] Cuando ocurre una colisión, la gravedad esperada del accidente es previsiblemente alta, pero en las zonas de límite de alta velocidad esta confluencia de riesgos conduce a un fenómeno muy raro: si un motociclista en East Sussex está involucrado en una colisión en una zona de límite de velocidad de 60 millas por hora o superior, es más probable que no haya una lesión grave o resultado de muerte.

Sabemos por entrevistas con profesionales de seguridad vial que los motociclistas a menudo se organizan para evitar las medidas de control de velocidad, y que exceder el límite de velocidad es un factor común que contribuye a las colisiones entre los motociclistas. Esto no significa que la reducción de los límites de velocidad no es efectiva, ya que puede actuar como un “cable a tierra”[14] o punto de referencia que afecta nuestra percepción de las velocidades altas y bajas; sin embargo, debemos reconocer que los motociclistas son más propensos a verse involucrados en conductas intencionalmente riesgosas y que no solamente se puede culpar a fallas en su concentración, lo cual hace que combatir este problema requiera el uso de técnicas variadas.

La tercera categoría de focalización contiene más matices: conductores de entre 45 y 65 años que interactúan con otros usuarios viales vulnerables. Si una colisión involucra, pero no es causada por, un usuario vial vulnerable, entonces existe una posibilidad aproximada de mortalidad o lesiones graves en uno de cada cinco casos si es que la otra parte está en el rango etario de 35-44 años (ver Figura 9). Si se aumenta esa edad a 45 años y más, habrá un aumento repentino de esta posibilidad a casi uno de cada cuatro casos. Estas colisiones ocurren a menudo cerca del hogar del conductor, y es muy común que sean provocadas por la falta de concentración o por una disposición “descuidada, imprudente o apresurada”.

Figura 9: Aumento repentino de colisiones graves/fatalidades con usuarios viales vulnerables para conductores mayores de 45 años.



Posibilidad de colisión grave o fatal (donde una persona muere o resulta gravemente herida: KSI) cuando esa colisión es causada por un conductor de automóvil e involucra a otro usuario vial vulnerable. La edad es la del conductor.

Las conductas exhibidas por estos tres grupos presentan una naturaleza muy distinta y no podemos esperar abordarlos de manera efectiva utilizando las mismas técnicas. Como parte de nuestro compromiso de comprender "qué funciona", probaremos variaciones de dos letras para conductores de automóviles en riesgo de conducción peligrosa, al igual que los efectos de las comunicaciones focalizadas durante los aniversarios de infracciones de conducción o colisiones menores para los tres grupos.

El valor de nuestro enfoque de ciencia de datos es que hemos evitado diseñar intervenciones que están condenadas al fracaso, debido a que el público objetivo no adopta el comportamiento que las intervenciones pretenden inhibir, o porque las intervenciones están dirigidas a un grupo de personas que simplemente no muestran conductas riesgosas.

Mejorando los ECA: Colaboración con KCL×BIT

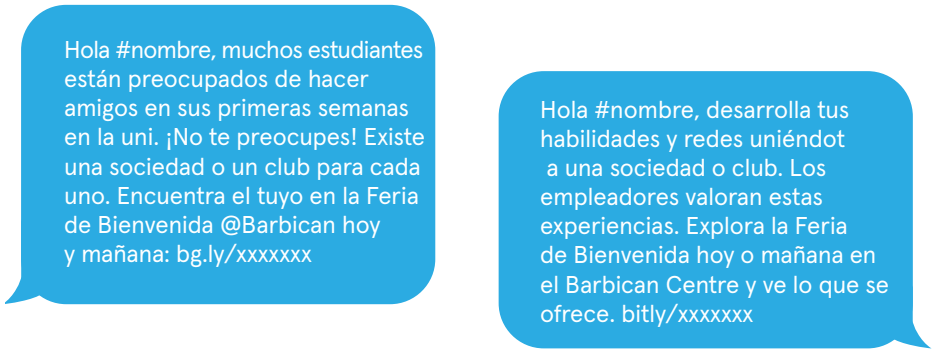
Hemos realizado más de diez ECA con KCL para saber qué es lo que funciona para aumentar la participación en varias actividades universitarias y el sentido general de pertenencia, utilizando conjuntos de datos institucionales para medir la eficacia. KCLxBIT es un proyecto colaborativo de dos años entre el "Departamento de Ampliación de la Participación", Policy Institute en KCL y BIT[15]. Reconociendo que el ingreso a la universidad no es un último paso para los estudiantes, especialmente para aquellos que forman parte de grupos poco representados, este proyecto se enfocó tanto en el éxito como en el acceso estudiantil.

Prueba: hacer que los estudiantes se inscriban en sociedades estudiantiles.

En el 2016 realizamos un ensayo que apuntaba a aumentar la probabilidad de que los jóvenes asistieran a la Feria de Bienvenida que realizaba el Centro de Alumnos de KCL al inicio del semestre. En estas ferias, los estudiantes pueden aprender acerca de los clubes y sociedades estudiantiles y luego inscribirse, así como también conocer a otros estudiantes.

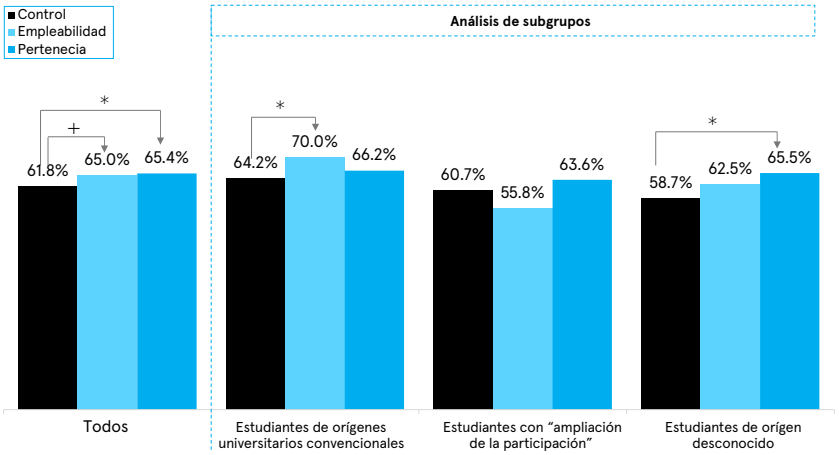
Realizamos un ECA para evaluar el efecto de distintos mensajes acerca de si los estudiantes asistieron a la feria y si es que se inscribieron en sociedades. Un tercio de los estudiantes elegibles no recibió ningún texto antes de la feria. Otro tercio de los estudiantes integraban el grupo de "pertenencia", recibieron tres textos centrados en reducir las barreras percibidas y asociadas con el sentido de pertenencia, mientras que el tercio final (en el grupo de "empleabilidad") recibió tres mensajes que enfatizaban los beneficios laborales de las sociedades. La Figura 10 muestra estos dos textos.

Figure 10: Textos enviados para la prueba durante la Feria de Bienvenida de KCL ("pertenencia" a la izquierda y "empleabilidad" a la derecha).



Descubrimos que los mensajes hicieron que los estudiantes tuvieran una mayor probabilidad de asistir a la Feria de Bienvenida, aunque el mensaje más efectivo parecía depender de las características de los estudiantes. En relación a la masa de estudiantes, los mensajes centrados en el sentido de pertenencia tuvieron el mayor impacto en la asistencia, resultando en un aumento del 6 por ciento. Los mensajes de empleabilidad, en cambio, aumentaron la asistencia en poco más del 5 por ciento, y el aumento no fue significativo a niveles estadísticos convencionales (ver Figura 11). Los mensajes de empleabilidad aumentaron las suscripciones a las sociedades, pero los mensajes de pertenencia no lo hicieron.

Figura 11: SAnálisis de subgrupos del efecto de los mensajes conductuales sobre la asistencia a la Feria de Bienvenida de 2016 por Ampliación del Estado de Participación.



**: p < 0.05 +: p < 0.1 ^: Los estudiantes fueron agrupados por su categorización Acorn, una medida de los antecedentes sociales. Los de orígenes convencionales pertenecían a las categorías 1 a 3 de Acorn (que refleja un mayor nivel socioeconómico), mientras que los estudiantes con "ampliación de la participación" se situaban en las categorías 4-5 de Acorn. Algunos estudiantes no entraban en una categorización Acorn, probablemente porque eran estudiantes internacionales, y estos habían sido agrupados en la categoría "origen desconocido".*

Prueba de seguimiento: utilización del aprendizaje automático para adaptar los mensajes

Debido a que los resultados del ensayo de la Feria de Bienvenida fueron prometedores, y que los efectos de las dos intervenciones parecían variar en distintos subgrupos, decidimos ejecutar una versión del ensayo para la recepción de 2017 usando un enfoque de aprendizaje automático. Utilizamos los resultados del ensayo de 2016 para predecir a qué condición deberían ser asignados los nuevos estudiantes en función de sus características.

Hubo dos condiciones en este ensayo. El primer grupo fue asignado aleatoriamente para recibir ya fuera el mensaje de sentido de pertenencia o de empleabilidad. En el segundo grupo, los estudiantes recibieron el mensaje que el algoritmo predijo que sería más eficaz para motivarlos a asistir. Eliminamos a los estudiantes que no deseaban recibir mensajes o cuyos números rebotaban.

Asignación aleatoria (N = 2.085)	Asignación algorítmica (N = 2.085)
50% recibió mensajes de sentido de pertenencia; 50% recibió mensajes de empleabilidad	La asignación a las condiciones de pertenencia o empleabilidad se basó en el algoritmo.

Este algoritmo se entrenó prediciendo qué estudiantes recibirían de mejor manera el mensaje de empleabilidad, en lugar del mensaje de sentido de pertenencia (más efectivo, en promedio) como el predeterminado.

En la condición de asignación aleatoria, el 60,1 por ciento de los estudiantes asistieron a la Feria de Bienvenida; el 59,1 por ciento para la condición de empleabilidad y el 61 por ciento para la condición de sentido de pertenencia. Por su parte, el 60,5 por ciento de los estudiantes asistieron a la Feria de Bienvenida bajo la condición de asignación algorítmica, la cual no difiere significativamente de la condición de asignación aleatoria.

Podría ser que, para algunos de los participantes que fueron asignados (por defecto) por el algoritmo al grupo de "sentido de pertenencia", la condición de control en realidad hubiera tenido un mejor rendimiento. Estamos iterando nuestra elección de algoritmos en este punto, pero una conclusión importante es que diseñar con exactitud los grupos objetivos de enfoque es importante para tener un mayor impacto a nivel general. Debido a que este se trata, según nuestro entendimiento, del primer ECA en probar este algoritmo, también hemos contribuido a la comprensión de cómo pueden probarse los algoritmos de aprendizaje automático para el análisis de subgrupos.

¿Qué es lo que conforma un buen proyecto de ciencia de datos?

Hemos encontrado que los siguientes cuatro ingredientes son clave en un proyecto de ciencia de datos:

1. Un problema predictivo, o la necesidad de comprender datos no estructurados;
2. Datos apropiados, de alta calidad y a gran escala;
3. La capacidad y la disposición departamental para implementar los hallazgos en la práctica;
4. Claridad ética y legal.

¿Qué es lo que conforma un buen problema predictivo? Es necesario que exista un comportamiento medido o una clasificación que podamos predecir (idealmente a nivel individual; desafortunadamente, una desventaja de las estadísticas oficiales es que a menudo se encuentran en un nivel general), y debe ser directamente relevante para las prioridades de nuestros asociados. Por ejemplo, con las inspecciones escolares, el problema predictivo fue predecir qué escuelas recibirían una calificación de "requiere mejoras" o de "inspección inadecuada", lo cual es claro y medible a partir de los datos históricos.

La siguiente cuestión es la calidad de los datos, ya que debe haber un buen ajuste entre el problema y los datos que podemos utilizar para resolverlo. Si bien cada proyecto es diferente, dos temas comunes son:

- Los datos que están en el mismo nivel que el comportamiento que nos interesa son más útiles que los datos que cubren un área o grupo más amplio.
- Si se requieren datos de texto o imágenes, vale la pena tener especial cuidado para garantizar que los datos sean de alta calidad, razonablemente completos y fácilmente extraíbles.

La disposición departamental es crucial para que la toma de decisiones no vuelva a caer en las formas tradicionales de pensar.

Sin embargo, también es importante pensar desde el comienzo acerca de la ética del proyecto:

1. La ciencia de datos tiene consideraciones particularmente fuertes en torno a la privacidad.
2. La legislación exige que cualquier decisión tomada por medios automáticos pueda ser solicitada para ser reconsiderada únicamente sobre esa base (con algunas exenciones).
3. Los mismos datos que se utilizan para entrenar algoritmos pueden estar sujetos a sesgos (por ejemplo, si solo los afroamericanos han sido históricamente y desproporcionadamente acusados de ciertos delitos debido al prejuicio policial, entonces el algoritmo exacerbará el prejuicio al predecir que los afroamericanos están en mayor riesgo de cometer esas ofensas).[16]
4. Los algoritmos que utilizan características protegidas, como la etnicidad o el sexo, deben diseñarse con mucho cuidado para que no sean menos precisos con respecto a los grupos vulnerables, incluso si los datos no son sesgados.

El futuro de la ciencia de datos en las políticas públicas

Pronóstico

Tradicionalmente, el pronóstico y su tarea complementaria, la asignación de recursos, han sido el dominio de meteorólogos y de personal militar. Sin embargo, debido a la disponibilidad de macro datos ("big data"), podrían ser aplicados de manera útil en todo el NHS y otros sistemas gubernamentales extensos.

Actualmente, estamos trabajando con Transforming Systems (una compañía de análisis de datos de salud especializada en datos de sistemas integrales) y Medway Hospital en un proyecto para predecir los tiempos de espera para pacientes de accidentes y emergencias mediante el aprendizaje automático. Los hospitales necesitan una forma precisa de conocer los tiempos de espera tanto para la planificación a corto plazo como para la gestión de turnos a mediano plazo, los cuales tienen un impacto real en los resultados de salud, graves y agudos. En este proyecto, estamos trabajando con los analistas de Transforming Systems utilizando una metodología similar a los métodos de Google para predecir el éxito publicitario con base en miles de consultas de búsqueda en el tiempo. Esperamos obtener resultados dentro de los próximos seis meses.

Análisis de costo-beneficio y qué funciona para quién

Consideramos que es importante ir más allá de "qué funciona" para ver "lo que funciona para quién". Trabajaremos con departamentos y agencias gubernamentales para examinar cómo los diferentes componentes de los programas nacionales, en particular aquellos centrados en grupos vulnerables, funcionan para diversos grupos de personas. Esto permitirá que las intervenciones estén mejor focalizadas y permitirá a los gobiernos cosechar más información de los costosos ejercicios de recopilación de evidencia.

El análisis de costo-beneficio es rutinario en el gobierno, sin embargo, la disponibilidad de nuevas herramientas y un mayor poder de procesamiento informático han significado que ahora podemos responder preguntas más interesantes que simplemente "en promedio, ¿fue mayor el beneficio que el costo?".

Estamos interesados en examinar cuándo debemos dejar de recopilar nuevos datos o hacer nuevas evaluaciones de las intervenciones y, en su lugar, avanzar hacia su ampliación. Esto depende, en parte, de la cantidad de información sobre suficientes partes del país o de subgrupos de personas.

Acerca de los autores

Michael Sanders
Científico Jefe y Resposable de Investigación y Evaluación

El Dr. Michael Sanders es Científico Jefe y Responsable de Investigación y Evaluación en el Equipo de Aprendizajes Conductuales (BIT). Junto a su trabajo en BIT, Michael trabaja como miembro asociado en Blavatnik School of Government y como investigador asociado superior en la University College de Londres, donde codirige el programa de doctorado en Ciencias del Comportamientos y Políticas. Michael completó su doctorado en Economía en la Universidad de Bristol, y sus estudios posdoctorales en la John F. Kennedy School of Government de Harvard.

James Lawrence
Jefe de Ciencia de datos

James Lawrence es el Jefe de Ciencia de datos de BIT. Antes de llegar a BIT y dirigir el trabajo descrito en este informe, James trabajó en investigación y desarrollo para una importante compañía de seguros del Reino Unido. James tiene antecedentes en matemáticas y estadística, con un MMath de la Universidad de Cambridge.

Daniel Gibbons
Consejero de investigación

Daniel Gibbons ejerce como Consejero de investigación para BIT trabajando en proyectos de evaluación y ciencia de datos con un interés particular en análisis de texto y modelos predictivos. Tiene un MPhil en Investigación Económica de la Universidad de Cambridge y una MSc. en Matemáticas y Estadística de la Universidad de Queensland.

Paul Calcraft
Jefe técnico y Científico de datos

El Dr. Paul Calcraft es Líder Técnico de BI Ventures y Científico de Datos en el Equipo de Investigación y Evaluación. Antes de unirse a BIT, Paul ejercía como programador en la industria de desarrollo web, últimamente desplazándose hacia el aprendizaje automático y la investigación. Ostenta un doctorado en Informática de la Universidad de Sussex.

Notas finales

1. Esto se basa en nuestra propia modelación, utilizando las mismas técnicas, pero con un conjunto limitado de datos. CQC de hecho utiliza más datos que estos en su propia focalización, y esta no se basa únicamente en un modelo. http://dsapp.uchicago.edu/resources/datamaturity
3. Nótese que el árbol de decisiones es un árbol ficticio y no refleja el modelo estadístico real.
4. Skills for Care (http://www.skillsforcare.org.uk/Home.aspx) es el organismo estratégico para el desarrollo de la fuerza de trabajo y capacitación para el servicio social para adultos en Inglaterra, y es la sede de la Academia Nacional de Habilidades para el Servicio Social.
5. Esto es cierto si el puntaje es menor a −0.5 y el intervalo de confianza es completamente inferior a cero.
6. Allen, R., & Burgess, S. (2012). How should we treat under-performing schools? A regression discontinuity analysis of school inspections in England. Working Paper No. 12/287. Centre for Market and Public Organisation. Disponible en http://www.bristol.ac.uk/media-library/sites/cmpo/migrated/documents/wp287.pdf
7. Debemos enfatizar que Ofsted actualmente no adopta este enfoque para su focalización de inspecciones, y que este trabajo no ha provocado ninguna inspección Ofsted al momento de la publicación.
8. Department for Education. (2017). Children in need: Project Crewe. Disponible en https://www.gov.uk/government/publications/children-in-need-project-crewe
9. Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. Journal of the American Statistical Association, 111(151), 988–1003.
10. Department for Transport. (2017). Reported road casualties in Great Britain: 2016 annual report. Statistical Release. Disponible en https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/648081/ rrcgb2016-01.pdf
11. Esto no quiere decir necesariamente que, por ejemplo, los trayectos ocupacionales causen menos colisiones que otros. Estamos haciendo afirmaciones acerca de la gravedad, a saber, que estas colisiones no son ni inesperadamente graves ni inesperadamente leves.
12. Department for Transport. (2016). Road traffic estimates in Great Britain: 2016. Disponible en https://www.gov.uk/ government/statistics/road-traffic-estimates-in-great-britain-2016
13. Department for Transport. (2016). Traffic counts: East Sussex traffic profile for 2000 to 2016. Disponible en https:// www.dft.gov.uk/traffic-counts/area.php?region=South+East&la=East+Sussex
14. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science (New Series), 185(4157), 1124–1131.
15. Para obtener más información acerca de esta colaboración más amplia, consulte el blog del proyecto: https://blogs.kcl.ac.uk/ behaviouralinsights
16. Para una serie de ejemplos, algunos de los cuales no se relacionan con las características protegidas (ver el punto 4 anterior), ver Sample, I. (2017). AI watchdog needed to regulate automated decision-making, say experts. The Guardian, 27 January. Disponible en https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence- watchdog-needed-to-prevent-discriminatory-automated-decisions

THE
BEHAVIOURAL
INSIGHTS TEAM

Equipo de Ciencia de Datos:

Michael Sanders, James Lawrence, Dan Gibbons, Paul Calcraft.

Colaboradores:

Doireann O'Brien, Clare Delargy, Edward Flahavan, Jessica Heal, Min-Taec Kim, Lucy Makinson, David Nolan, Sean Sheehan, Handan Wieshmann.

Equipo de Aprendizajes Conductuales 4
Matthew Parker Street
Londres
SW1H 9NP