Conceptos de modelización en la formación universitaria de los analistas de datos

Eleonora Bottino
Prensa Ibérica
eleonorabottino@gmail.com

Luis Hidalgo
Prensa Ibérica
luishipe@hotmail.com

Introducción

Un asunto que ha ido surgiendo recurrentemente durante años en conversaciones en los ámbitos profesionales en los que hemos desarrollado nuestras carreras es lo incompletos que son los conocimientos sobre el proceso de modelización que tenemos los titulados universitarios cuando finalizamos nuestros estudios. Quizás hace demasiado tiempo que nosotros terminamos nuestras licenciaturas y la memoria nos está traicionando, pero lo que recordamos de las clases en las que aprendimos a aplicar los algoritmos avanzados de análisis de datos es que el conjunto de datos ya venía dado. Habitualmente se proporcionaba una tabla con una variable dependiente (aquella cuyos valores tenías que predecir) y con una serie de variables independientes (que debían actuar como variables predictivas en el modelo), y nuestro trabajo como estudiante consistía en elegir la técnica o técnicas adecuadas y aplicarlas correctamente para obtener el resultado. Esas conversaciones a las que nos referíamos al comienzo de este articulo han tenido lugar a lo largo de los años desde que comenzamos a dedicarnos profesionalmente al análisis de datos a mediados de la primera década del siglo XXI hasta estos días, y en ellas han participado compañeros de diferentes generaciones, que habían estudiado en diferentes zonas del país y que poseían diferentes tipos de licenciaturas, ingenierías o másteres en los que se impartían asignaturas relacionadas con el análisis de datos. La conclusión casi siempre ha sido la misma: en las tareas de análisis de datos que realizaron durante su etapa de formación académica el conjunto de datos que había que analizar ya estaba construido y las variables que lo formaban ya venían definidas. Sin embargo, en los entornos empresariales el trabajo del analista de datos suele comenzar en un punto anterior del ciclo y el profesional tiene que responsabilizarse también de la construcción del conjunto de datos que sirve como input al proceso de entrenamiento del modelo, incluyendo la definición de todas las variables que forman parte del mismo. Durante ese proceso el analista tendrá que tomar algunas decisiones y realizar algunas asunciones que serán determinantes en la utilidad práctica que tenga el conocimiento proporcionado por el modelo una vez que haya sido entrenado. En nuestra opinión esta fase es crítica para el éxito del trabajo y es la que más diferencias permite generar entre el trabajo de diferentes profesionales. Casi todos nos apoyamos en el uso de las mismas herramientas informáticas para nuestro desempeño laboral y empleamos las versiones de los

algoritmos que están implementadas en ellas. Aprender a utilizar los algoritmos proporcionados por esos programas puede tener mayor o menor dificultad técnica, que una vez superada nos deja a todos los analistas en igualdad de condiciones. Desde ese punto de vista, el planteamiento que se realice durante la fase de modelización del problema y la elección de las variables independientes adecuadas será lo más relevante a la hora de lograr entrenar un buen modelo predictivo. En base a nuestra experiencia pensamos que cuanto mayor sea el conocimiento que el analista posea sobre el modelo de negocio en el que se desea generar las predicciones, más acertada podrá ser la modelización del problema que es objeto del análisis. Y estamos convencidos de que ese conocimiento solo es posible adquirirlo con la práctica del día a día de cada entorno empresarial concreto, ya que incluso dentro del mismo sector de actividad y en el mismo mercado geográfico hay aspectos que son específicos de una empresa y que no es sencillo trasladar a otras. Por eso nos parece complicado que en el ámbito académico se pueda enseñar ese conocimiento experto sobre la materia, ya que se necesitaría demasiado tiempo para tratar todos los sectores de actividad en los que podrían terminar trabajando los alumnos y lo más probable es que los docentes no dispongan de experiencias reales que trasladarles en un porcentaje relevante de esos sectores. Lo que quizás sí sea asumible durante la etapa académica es presentar a los estudiantes los conceptos que habitualmente están presentes durante la fase de modelización y compartir con ellos las diferentes alternativas sobre las que van a tener que elegir para concretar la modelización y las implicaciones derivadas de cada una de ellas. A continuación, vamos a presentar algunos de esos conceptos y para ello utilizaremos el nombre con el que nosotros los aprendimos y que a día de hoy seguimos utilizando, aunque no estamos seguros de que sean un estándar y por tanto es posible que otros profesionales empleen nombres diferentes para referirse a los mismos conceptos, que son lo verdaderamente relevante.

Ventana temporal de respuesta

Entendemos como Ventana temporal de respuesta el periodo del pasado del cual vamos a utilizar sus datos para construir la variable objetivo del modelo. Una de las principales decisiones que deberá tomar el analista en la etapa de modelización es la amplitud de ese periodo ya que determinará el tiempo que se mantendrá vigente cada una de las predicciones que proporcione el modelo. Si elegimos una duración de la Ventana temporal de respuesta de 1 día, 1 semana, 1 mes o 1 año, la vigencia de las predicciones que hagamos, una vez que el modelo esté entrenado, será respectivamente de 1 día, 1 semana, 1 mes o 1 año. Hay que tener en cuenta que cuanto menor sea la duración de la Ventana temporal de respuesta más variabilidad existirá y será más complicado que el modelo ofrezca predicciones precisas. Por otro lado, cuanto más hacia el futuro se alargue la Ventana temporal de respuesta mayor será la incertidumbre sobre cómo evolucionará el entorno del negocio en el tiempo y si se mantendrán o no las condiciones en las que se ha entrenado el modelo. Para decidir la longitud de este periodo también tendremos que tener en cuenta el ámbito de negocio. Si por ejemplo quiero disponer de la probabilidad de que un ascensor se vaya a estropear para realizar tareas de mantenimiento, no será de mucha utilidad que calcule esa predicción para los próximos 5 minutos, ya que aportaría poco valor al negocio comparado con obtener la predicción para 24 horas y a cambio la precisión del modelo descendería considerablemente, ni que obtenga la probabilidad de que se estropee en los próximos 50 años, pues casi seguro que lo va a hacer y ese conocimiento no me ayudará a saber cuál es el mejor momento para las labores de mantenimiento. Además, hay que tener cuenta la profundidad histórica de datos que estén disponibles, ya que en el caso de que se desee utilizar todos habrá que repartir el periodo histórico entre la Ventana temporal de respuesta y la Ventana temporal de análisis (que presentaremos más adelante). Por ejemplo, si solo contamos con información sobre los últimos seis meses, la duración que asignemos a la Ventana temporal de respuesta deberá ser

inferior a seis meses, porque en caso contrario no dispondremos de datos para construir las variables independientes. Una vez fijada la duración de la Ventana temporal de respuesta el analista tendrá que determinar el periodo o periodos concretos del pasado que utilizará para obtener los datos con los que construirá la variable que queremos predecir. En este punto la mayor dificultad radica en ser capaces de seleccionar unos periodos que sean representativos de los diferentes periodos del futuro en los que queramos utilizar las predicciones generadas por el modelo. Si, por ejemplo, queremos que el modelo nos sirva para hacer una única predicción que tenga vigencia durante las 24 horas del día 1 de enero de 2023, podría ser adecuado que, en función de la estacionalidad que tenga el modelo de negocio en el que estamos trabajando, elijamos como Ventana temporal de respuesta las 24 horas del día 31 de diciembre de 2022 (día anterior), las 24 horas del día 25 de diciembre de 2022 (mismo día de la semana anterior), las 24 horas del día 1 de diciembre de 2022 (mismo día del mes anterior) o las 24 horas del día 1 de enero de 2022 (mismo día del mismo mes del año anterior), ya que podrían ser los datos del pasado conocido más similares a la situación en la que vamos a obtener la predicción. Si, por el contrario, pensásemos utilizar el modelo para cada uno de los 365 días del año 2023 (cada uno de ellos es un periodo de 24 horas de duración) es difícil que seleccionando un único día del pasado conocido como Ventana temporal de respuesta recojamos toda la variabilidad de comportamientos a la que se va a enfrentar el modelo en el momento de realizar las predicciones y tendremos que recurrir a mecanismos más sofisticados para definir dicha ventana.

Definición de la variable objetivo (target)

Otra de las tareas que hay que realizar durante la fase de modelización es preparar una definición precisa de la variable cuyos valores futuros se desee predecir y posteriormente construirla. Como explicamos anteriormente, los datos utilizados para la elaboración de la variable objetivo deben pertenecer al periodo indicado en la Ventana temporal de respuesta. Habrá ocasiones en las que para definir esa variable se pueda utilizar un criterio relativamente objetivo, como en el caso de querer predecir la probabilidad de que un cliente de una empresa de banca solicite la baja de su servicio de tarjeta de crédito o estimar los euros que determinada empresa ingresará por la venta de un producto concreto. Sin embargo, otras veces habrá que tomar decisiones subjetivas a la hora de definir dicha variable. En el caso de que el cliente de un supermercado haya decidido dejar de ir a comprar allí, sería muy extraño que se pusiese en contacto con el mismo para avisarles de que ha tomado esa decisión. Si el supermercado quiere calcular la probabilidad que tiene cada uno de sus clientes de dejar de serlo, tendrá que definir con precisión qué considera que alguien deje de ser cliente. "Los clientes que no vuelven a comprar", "Los clientes que tienen un gasto acumulado menor que 10 euros" o "Los clientes que han comprado 3 días distintos o menos en la Ventana temporal de respuesta" son algunas de las opciones que ese supermercado podría utilizar para definir el abandono por parte de sus clientes.

Definición del Universo de análisis

Para definir el Universo de análisis tenemos que especificar quienes de los "individuos" susceptibles de ser objeto del modelo realmente lo serán, ya que en ocasiones buscando aumentar la precisión del modelo, nos interesará que esos "individuos" sean lo más homogéneos posible. Si una empresa de telecomunicaciones quiere entrenar un modelo de abandono que prediga la probabilidad que tiene cada uno de sus clientes de solicitar la baja de los servicios contratados, es muy posible que decida analizar por separado a los clientes del "mercado residencial" de los clientes del "mercado empresarial",

pensando que el tipo de relación que mantienen con la empresa es de diferente naturaleza y que por tanto los patrones que sigan antes de solicitar la baja también puedan ser diferentes. La profundidad histórica de datos disponible para cada uno de los individuos también es un factor muy relevante a la hora de definir el Universo de análisis, ya que, dependiendo de la misma, las variables objetivo y/o independientes del modelo podrían estar representando realidades diferentes para cada uno de los individuos. En el modelo de predicción del abandono que comentábamos, si decidimos trabajar con una Ventana temporal de respuesta de 1 mes y una Ventana temporal de análisis (concepto que presentaremos posteriormente) de 24 meses, ¿incluimos a los clientes con antigüedad inferior a 24 meses? Si lo hacemos, la variable "Importe de gasto histórico acumulado" reflejará la actividad durante 24 meses para los clientes que tengan una antigüedad superior a 25 meses y la actividad de un periodo menor para los clientes cuya antigüedad sea menor. Por ejemplo, para los clientes que tengan una antigüedad de 7 meses esa variable recogería su gasto durante un periodo de 6 meses. Y eso es asumible o no dependiendo del tipo de evento que estemos modelizando y de las particularidades del contexto en el que estemos trabajando.

Definición de la ventana temporal de análisis

La Ventana temporal de análisis es el periodo concreto del pasado del que utilizaremos sus datos para construir las variables independientes o explicativas del modelo. Aunque a algunos les pueda parecer una perogrullada, insistiremos aquí en que la Ventana temporal de análisis debe finalizar estrictamente antes de que comience la Ventana temporal de respuesta, ya que la información relativa a esta última no estará disponible en el momento en el que queramos utilizar el modelo para realizar una predicción de la variable objetivo. En algunas ocasiones nos hemos encontrado en el entorno empresarial con modelos que, tras su entrenamiento, arrojaban indicadores de precisión extrañamente altos y al profundizar en su planteamiento, nos hemos dado cuenta de que no se había respetado esa relación entre la Ventana temporal de respuesta y la Ventana temporal de análisis, y para predecir el valor de una variable se estaba utilizando información generada al mismo tiempo o incluso posteriormente. En nuestro grupo de trabajo utilizamos el término "Futurazo" para referirnos a esta situación. Como indicamos al hablar sobre la Ventana temporal de respuesta, la profundidad histórica de datos que esté disponible limita la flexibilidad a la hora de definir la longitud de la Ventana temporal de análisis. En principio cuánto más amplio sea el periodo utilizado como Ventana temporal de análisis más rica será la información puesta a disposición del modelo para encontrar patrones de comportamiento. Lo que hay que tener muy presente son los cambios relevantes que hayan podido modificar el contexto, tales como cambios en la legislación o en el modelo de negocio de la empresa (promociones especiales, descuentos, cambios tarifarios, etc...). Si trabajamos con una Ventana temporal de respuesta de 12 meses (para realizar predicciones de lo que ocurrirá el próximo año) y hace 18 meses hubo un cambio relevante por parte del regulador, definir una Ventana temporal de análisis de una longitud superior a 6 meses implicaría que parte de la información utilizada para construir las variables independientes empleadas en el entrenamiento del modelo estarían representando una situación regulatoria diferente de la existente cuando se generaron los datos usados para obtener la predicción.

Definición de la ventana temporal ciega

Hay ocasiones en las que puede resultar útil dejar una determinada distancia temporal entre el momento en el que finaliza la Ventana temporal de análisis y el momento en el que comienza la Ventana temporal de respuesta. A ese periodo le llamamos Ventana temporal ciega. Las situaciones en

las que interesa emplear este tipo de ventana temporal suelen responder a consideraciones operativas relativas al uso que se vaya a realizar del conocimiento proporcionado por el modelo. Si tardamos 24 horas en recopilar y consolidar los datos necesarios para puntuar (obtener las predicciones) a los individuos tendrá sentido que en el planteamiento de la modelización dejemos una ventana temporal de 24 horas desde el final de la Ventana temporal de análisis hasta el inicio de la Ventana temporal de respuesta, ya que si no lo hacemos la vigencia de la predicción que generemos comenzará 24 horas antes de que seamos capaces de obtener la predicción. Si vamos a utilizar las predicciones generadas por el modelo para, por ejemplo, enviar una carta postal mejorando las condiciones comerciales a nuestros clientes con alta probabilidad de darse de baja en los siguientes 2 días, y el tiempo desde la generación de las predicciones hasta que los clientes reciben la carta en su domicilio es de 5 días, durante la modelización tendremos que tener esto en cuenta, porque de lo contrario el cliente recibirá la carta cuando ya haya tramitado la baja.

Definición de las variables explicativas

Con respecto a la definición de las variables independientes o explicativas que se utilizarán en el modelo, indicar que los datos utilizados deberán estar contenidos en la Ventana temporal de análisis que se haya definido. En la medida de lo posible la información recogida por esas variables deberá representar cualquier aspecto que pueda tener influencia en el valor de la variable que queremos predecir. Siguiendo con el ejemplo del modelo de predicción del abandono para los clientes de una operadora de telecomunicaciones, tendría sentido incluir variables inherentes al propio cliente (edad, género, poder adquisitivo, zona geográfica de residencia, etc...), variables sobre los productos contratados, variables sobre el uso histórico que ha hecho de ellos, variables sobre las incidencias que haya podido tener en la prestación del servicio o en su facturación durante su relación con la compañía, variables sobre la presión comercial que la competencia esté realizando en la zona geográfica en la que resida, etc.... A priori no sabemos qué variables serán las que efectivamente tendrán capacidad de predecir los valores que tomará la variable objetivo, así que cuanta más información proporcionemos al algoritmo más oportunidades le estaremos dando de encontrar combinaciones de variables explicativas con alto poder predictivo. Siempre dentro de un orden de magnitud razonable, ya que si incluyésemos millones de variables explicativas el tiempo necesario para la ejecución del algoritmo de predicción podría resultar excesivamente alto a cambio de pequeños incrementos en la precisión del modelo obtenido. Añadir que si algunas de las variables explicativas empleadas en el modelo representan magnitudes o acciones que la empresa sea autónoma para modificar, el modelo obtenido será más fácilmente accionable y tendrá más aplicaciones para la compañía. Si el modelo determina que el valor de la variable objetivo depende del precio con el que la empresa comercializa un producto, esta podría decidir rebajar o aumentar dicho precio para influir en los valores futuros de la variable objetivo. Mientras que, si el aprendizaje obtenido del modelo es que el valor de la variable objetivo depende de la tasa de paro a nivel nacional, a la empresa le vendrá bien saberlo, pero será mucho más difícil que por su cuenta pueda lograr que esa magnitud aumente o disminuya.

Agradecimientos

Antes de despedirnos nos gustaría agradecer a esta publicación y a sus editores la oportunidad que nos han brindado para compartir esta reflexión, que esperamos que pueda contribuir a que los alumnos de programas universitarios relacionados con el análisis de datos puedan acceder al mercado

laboral con recursos que les resulten útiles para afrontar algunos de los problemas a los que se van a enfrentar en el ámbito empresarial.

Acerca de la autores

Eleonora Bottino Licenciada en Economía, cuenta con más de 10 años de experiencia en el desarrollo de modelos analíticos en ámbito internacional en empresas de los sectores de banca, administración pública y medios de comunicación. En estos entornos ha contribuido al desarrollo de modelos de venta cruzada, de valor de cliente, detección de fraude, riesgo, propensión a la compra y optimización de campaña segmentadas. Actualmente es Manager de Ciencia de Datos en Prensa Ibérica.

Luis Hidalgo Licenciado en Matemáticas con más de diez años de experiencia en la aplicación de modelos analíticos para la generación de conocimiento útil en la toma de decisiones, en empresas líderes en España de los sectores de Telecomunicaciones, Distribución y Medios de comunicación, principalmente en las áreas de Inteligencia de cliente y Marketing Relacional. Actualmente es Director de Ciencia de Datos en Prensa Ibérica y profesor en ESIC.