

Más allá del análisis actuarial: Hacia una forma sistemática de abordar el riesgo

Antecedentes

En un mercado competitivo de seguros de vida, una tarificación precisa del riesgo es vital para atraer a nuevos asegurados sin incurrir en antiselección. El análisis de la experiencia actuarial tradicional puede aportar información valiosa sobre el riesgo biométrico al que está expuesta una aseguradora y las tendencias de esta experiencia a lo largo del tiempo. Sin embargo, el análisis tradicional tiene algunas limitaciones, y las técnicas de ciencia de datos más avanzadas pueden ofrecer más información para ayudar a comprender el riesgo.

Tomemos el ejemplo de un análisis de experiencia que muestra una tendencia de mejora de la experiencia biométrica por año de suscripción. Se podría concluir que los negocios más recientes están funcionando mejor que los históricos, lo que permite fijar los precios de los nuevos negocios a un nivel inferior al del agregado según la experiencia histórica.

Tal vez las normas de suscripción han ido mejorando con el tiempo, o el marketing y la distribución han llevado a los asegurados de menor riesgo a favorecer a su organización. Sin embargo, podría observarse una tendencia similar si en realidad la experiencia se deteriora debido a la duración de la póliza, porque en los primeros años de suscripción están las pólizas que contribuyen a la experiencia de duración posterior en nuestro análisis.

Con los métodos tradicionales, puede ser muy difícil distinguir estos efectos, y aún más difícil desglosar el impacto. Sin embargo, los modelos más sofisticados permiten controlar múltiples factores de riesgo y obtener más información sobre los verdaderos impulsores de la experiencia. En el ejemplo anterior, es posible separar el impacto de la duración y el año de suscripción, y ver las verdaderas tendencias subyacentes.

Esto no niega la necesidad de comprender el negocio suscrito y los factores que impulsan la experiencia subyacente. Comprender la razón por la que observamos una tendencia a la mejora por año de suscripción, por ejemplo, es vital para garantizar que se toman las decisiones correctas a la hora de tarificar nuevos negocios, pero las nuevas técnicas son extremadamente útiles para destacar los factores clave de la experiencia histórica.

Limitaciones del análisis tradicional

El análisis tradicional de la experiencia está sujeto a una serie de limitaciones, derivadas de no considerar los datos desde una perspectiva holística y de la pérdida de información debida a la transformación de los datos. Estas limitaciones se agrupan en cuatro categorías que se analizan a continuación.

Limitación 1: No modelizar adecuadamente las variables continuas

Debido a las dificultades para visualizar las relaciones entre las variables predictoras continuas (como la edad) y los resultados, y para aumentar la potencia estadística de cada estimación puntual, las variables continuas suelen agruparse en grupos. Como ya se ha comentado anteriormente¹, esto conlleva una pérdida de información y puede ocultar patrones (como el pico de enfermedad crítica a los 50 años entre las mujeres debido al cribado del cáncer de mama).

La pérdida de información puede ser doble: en primer lugar, se pierde información al combinar diferentes valores en un único grupo y, en segundo lugar, si la variable agrupada se trata como una categoría, la secuencia continua se pierde para cualquier modelo posterior.

¹ <https://www.hannover-re.com/1650132/recent-uk-insights-an-insight-into-accelerated-critical-illness-experience-2021.pdf>

Limitación 2: Riesgo de interpretación errónea debido a factores de confusión

La confusión se produce cuando la relación entre una variable y el resultado se distorsiona debido a la presencia de una segunda variable que está asociada tanto con la primera variable como con el resultado. La confusión puede dar lugar a una interpretación errónea de los resultados del estudio:

Atribución errónea de los efectos: Si no se tiene en cuenta adecuadamente una variable de confusión, ésta puede crear una asociación espuria o enmascarar una asociación real. Por ejemplo, si no se tiene en cuenta el tabaquismo (variable de confusión), puede aparecer una asociación espuria entre el consumo de café y el cáncer de pulmón, ya que los fumadores tienen más probabilidades de tomar café que los no fumadores.

Asociación distorsionada: Los factores de confusión pueden llevar a sobrestimar o subestimar la verdadera asociación entre la variable predictiva y el resultado. Consideremos el conjunto de datos hipotético resumido en

la tabla 1. La tarificación de los hombres al 120%, de las mujeres al 80%, de las vidas en común al 80% y de las vidas individuales al 120% (ajuste 1) no tarificará con exactitud este conjunto de datos ni siquiera en ausencia de una interacción (más información sobre las interacciones en el siguiente apartado) entre el sexo y el estado de vida en común. La razón es que el sexo y la situación de vida en común son factores de riesgo correlacionados en este conjunto de datos (tabla 2): la mayoría de los hombres son solteros y la mayoría de las mujeres viven en común. En el ajuste 1, los tamaños del efecto del sexo y de la situación de vida conjunta están sobreestimados debido a la confusión. En este ejemplo concreto, los ajustes de precios son relativamente fáciles de realizar calculando directamente los cocientes reales/esperados para cada categoría individualmente (ajuste 2).

Las variables predictoras rara vez no están correlacionadas. El ajuste manual de múltiples variables correlacionadas se complica progresivamente, especialmente en el caso de variables numéricas.

Tabla 1: Conjunto de datos hipotético - por sexo y situación vital conjunta univariante

Característica	Categoría	Real	Esperado	Ratio Real / Esperado Tabla estándar
Género	Mujer	600	500	120%
	Varón	400	500	80%
Estado de vida en común	Situación vital conjunta	600	500	120%
	Soltero	400	500	80%

Tabla 2: Conjunto de datos hipotético - por sexo y situación vital conjunta multivariante

Género	Estado de vida común	Real	Esperado	Ajuste 1	Ajuste 2
Varón	Soltero	478	375	$375 * 120\% * 120\% = 540$	$375 * 127,6\% = 478,5$
Varón	Conjunto	122	125	$125 * 120\% * 80\% = 120$	$125 * 97,3\% = 121,6$
Mujer	Soltero	122	125	$125 * 80\% * 120\% = 120$	$125 * 97,3\% = 121,6$
Mujer	Conjunto	278	375	$375 * 80\% * 80\% = 240$	$375 * 74,2\% = 278,3$
Conjunto de datos completo		1.000	1.000		1.000

Generalizabilidad limitada: Los factores de confusión pueden afectar a la generalizabilidad de los resultados del estudio. Si la variable de confusión está desigualmente distribuida entre los grupos de estudio, la asociación observada puede no ser generalizable a otros grupos o contextos.

Por ejemplo, la fijación de precios por canal de ventas derivada de un estudio distorsionado por factores socioeconómicos sólo será válida para poblaciones con el mismo desglose socioeconómico por canal de ventas.

Limitación 3: Se presta poca atención a las interacciones entre variables

Las interacciones entre variables se producen cuando una variable predictiva modifica el efecto de otra variable sobre el resultado. Por ejemplo, la forma seleccionada podría diferir hipotéticamente entre distintos canales de venta.

En estos casos, la determinación precisa de la forma de selección y del efecto del canal de ventas a nivel agregado no tendrá en cuenta la interacción, que requiere una forma de selección diferente para cada canal de ventas. Las interacciones entre variables sólo pueden detectarse considerando explícitamente cada par de variables, lo que puede llevar mucho tiempo (limitación 4).

Algunos modelos, como los basados en árboles y los de aprendizaje profundo, tienen en cuenta de forma inherente las interacciones variables. La contrapartida es la falta de transparencia de estos modelos.

Limitación 4: Falta de un método sistemático para la selección de características

In addition to being vulnerable to confounding, case-by-case analysis of variables can be time consuming. This is even more so if pairs (or even triplets) of variables are considered on a case-by-case basis to explore for interactions. For example, with 10 variables, there are 45 potential two-way interactions to consider.

Flujo de trabajo para abordar las limitaciones del análisis tradicional

Una posible solución a los problemas descritos anteriormente es incorporar una matriz de correlación, así

como una evaluación de la importancia de las variables en el análisis tradicional de la experiencia. Esto se incorpora mejor en la exploración de datos. Ninguna de estas técnicas requiere el anillado de variables continuas, lo que evita la pérdida de información (limitación 1).

Matriz de correlaciones

Una matriz de correlaciones es una tabla que muestra sistemáticamente los coeficientes de correlación entre las variables de un conjunto de datos. Proporciona una visión completa de la fuerza y la dirección de las relaciones entre pares de variables.

La matriz de correlaciones puede:

1. Identificar posibles variables de confusión, lo que puede evitar la atribución errónea de efectos, asociaciones distorsionadas y una generalizabilidad limitada (limitación 2).
2. Proporcionar una comprensión más profunda del conjunto de datos. Por ejemplo, un cambio en la combinación de negocios a lo largo del tiempo. Esto puede incluso dar lugar a la exclusión de ciertos datos del análisis (por ejemplo, un canal de ventas que ya no proporciona nuevos negocios) con el fin de mejorar la generalizabilidad.
3. Identificar las variables redundantes y evitar la multicolinealidad. Las variables predictoras fuertemente correlacionadas pueden estar captando las mismas características y algunas podrían eliminarse con seguridad del análisis sin pérdida de información. Evitar variables predictoras fuertemente correlacionadas también puede prevenir problemas de multicolinealidad.
4. Identificar posibles errores en los datos. Los errores en la introducción de datos o la falta de datos no aleatorios pueden provocar correlaciones fuertes e inesperadas entre las variables. Por ejemplo, si no se notifican las vidas nominales para determinados canales de venta, se produciría una correlación fuerte e inesperada entre las vidas nominales y el canal de venta.

Importancia de las variables

Una tabla de importancia de las variables es un resumen que clasifica las variables predictoras de un modelo en función de su importancia o contribución al rendimiento del modelo. La importancia de la variable proporciona una medida de la importancia relativa de cada variable en la predicción del resultado. Las tablas de importancia de las variables pueden ser generadas por varios modelos, incluidos los modelos basados en árboles, como los bosques aleatorios y las máquinas de aumento de gradiente, así como la regresión LASSO.

Las tablas de importancia de las variables proporcionan un método sistemático para evaluar la importancia relativa de todas las variables predictoras en un modelo único (holístico) (limitación 4). Además, todas las interacciones potenciales entre variables pueden generarse explícitamente utilizando la matriz de diseño del modelo y considerarse utilizando la tabla de importancia de variables (limitación 3). El modelo subyacente a la tabla de importancia de variables también separará los efectos de confusión (limitación 2) y asignará importancia a la causa raíz de los efectos. Las tablas de importancia de las variables permiten determinar con rapidez qué variables predictoras determinan un resultado.

Flujo de trabajo descendente

La matriz de correlaciones y la importancia de las variables proporcionarán información sistemática sobre los datos, las variables y las interacciones de las variables que deben tenerse en cuenta en un modelo de producción. La incorporación de la matriz de correlaciones y la importancia de las características en el flujo de trabajo de exploración de datos reducirá en gran medida el riesgo de factores de confusión, interacciones de variables no identificadas o modelos poco generalizables.

Limitaciones

Los factores de confusión pueden no estar recogidos en el conjunto de datos. Por ejemplo, en el ejemplo de generalizabilidad limitada del canal de ventas confundido por factores socioeconómicos, la información socioeconómica puede no estar capturada en el conjunto de datos. En estos casos, el análisis de datos no podrá detectar ni ajustar la variable de confusión. La opinión de los expertos es la única herramienta disponible para alimentar cualquier análisis en estos casos.

Aunque las tablas de importancia de las variables proporcionan una visión sistemática de qué variables impulsan el resultado, la selección de características en sí debe tener lugar en el marco del modelo de producción (por ejemplo, mediante selección por pasos o regresión LASSO). En caso de que haya una gran cantidad de variables, durante la exploración de los datos se pueden excluir todas aquellas en las que las puntuaciones de importancia de las variables sean similares o inferiores al ruido aleatorio.

Conclusión

La incorporación de una matriz de correlación y el cálculo de la importancia de las variables en el flujo de trabajo de exploración de datos del análisis de la experiencia es una forma eficaz de reducir en gran medida el riesgo de confusión, las interacciones entre variables omitidas o los modelos poco generalizables.

Hannover Re utiliza los conocimientos derivados del flujo de trabajo descrito para establecer nuestra mejor estimación de las tasas de mortalidad/morbilidad y podemos apoyar una evaluación/tarifación del riesgo más granular. Hannover Re también puede proporcionar información específica del cliente utilizando métodos similares.

Autores



Michiel Luteijn Ph.D.
Senior Data Scientist
Hannover Re UK Life Branch
Tel. +44 20 3206 1828
michiel.luteijn@hannover-re.com



Tim Smith
Head of Protection
Hannover Re UK Life Branch
Tel. +44 20 3206 1811
tim.smith@hannover-re.com

Síguenos en [LinkedIn](#) para estar actualizado con las últimas noticias de Vida y Salud.



La información proporcionada en este documento no constituye en modo alguno asesoramiento legal, contable, fiscal u otro tipo de asesoramiento profesional. Si bien Hannover Rück SE se ha esforzado por incluir en este documento información que considera segura, completa y actualizada, la compañía no hace ninguna representación o garantía, expresa o implícita, en cuanto a la precisión, integridad o estado actualizado de dicha información. Por lo tanto, en ningún caso Hannover Rück SE y sus compañías afiliadas o directores, funcionarios o empleados serán responsables ante nadie por cualquier decisión o acción tomadas, o por cualquier daño relacionado con la información de este documento.

© Hannover Rück SE. Todos los derechos reservados. Hannover Re es la marca de servicio registrada de Hannover Rück SE.