No. 1, 2018

RISK INSIGHTS®

Rendimiento de los Modelos de Clasificación

por Louis Rossouw, Gen Re, Ciudad del Cabo

Las aseguradoras están desarrollando cada vez más modelos de predicción para utilizarlos en sus procesos de seguros. A menudo, estos modelos utilizan técnicas tradicionales, pero vemos cómo está aumentando la aplicación de técnicas de aprendizaje automático.

En la práctica, estas técnicas se aplican a casos de suscripción, solicitudes de póliza o incluso siniestros. Un ejemplo podría ser un modelo que se está utilizando para predecir qué casos serán evaluados como «estándar» antes de que un suscriptor los vea. Esto se llama «modelo de clasificación», y se utiliza para clasificar puntos de datos en diferentes categorías (sí o no, estándar o no, etc.).

Las técnicas de modelado utilizadas en dichas aplicaciones pueden ser muy sencillas o muy complejas. Entre los ejemplos de estas técnicas se incluyen:

- Regresión logística (normalmente, un modelo lineal generalizado –GLM, por sus siglas en inglés)
- Árboles de decisiones
- Bosques aleatorios
- Máquinas de vectores de soporte
- Técnicas de potenciación de gradientes
- Redes neuronales

Las técnicas tradicionales tales como los modelos basados en la regresión producen modelos de lectura humana. Se puede ver claramente el impacto de cada variable del modelo en el resultado. Sin embargo, muchas de las técnicas de aprendizaje mecánico producen modelos que no resultan fáciles de entender a simple vista. Proporcionan información, pero los procesos internos están ocultos o son demasiado complejos para comprenderlos íntegramente. Son los llamados modelos de «caja negra».

Contenido

Acerca de este boletín

Risk Insights es una publicación técnica elaborada por Gen Re para ejecutivos de seguros de Vida y Salud de todo el mundo. Los artículos se centran en asuntos actuariales, de suscripción, siniestros, médicos y de gestión de riesgos. Entre los productos a los que se les concede una mayor atención se incluyen los seguros de Vida, Salud, Invalidez, Dependencia y seguro de Enfermedades graves.

Con una gama tan amplia de modelos a elegir, ¿cómo evaluamos su exactitud y calidad para determinar cuál es mejor usar? ¿Cómo consideramos el rendimiento de los modelos fáciles de entender, en comparación con varios modelos de caja negra? ¿Cómo evaluamos el valor relativo de estos modelos para la aseguradora?

En el resto del artículo nos centraremos en un clasificador binario simple que predice si una solicitud en particular debería considerarse «estándar» (sin recargos ni condiciones de suscripción) o no (rechazada, con un recargo u otras condiciones de suscripción aplicadas). El uso potencial de un modelo de este tipo es el de evitar la suscripción tradicional para ahorrar tiempo y costes para un subconjunto de casos.

Datos de entrenamiento y prueba

Al crear un modelo de clasificación, normalmente se debería disponer de un conjunto de datos con casos históricos y el resultado registrado de los mismos. En nuestro ejemplo de suscripción, estos podrían ser datos relacionados con el solicitante y la decisión de suscripción registrada (estándar o no).

Es recomendable dividir los datos pasados en al menos dos categorías:

- datos de entrenamiento que se utilizarán para ajustarse a los modelos en cuestión; es decir, estos datos se usan para «entrenar» el modelo
- datos de prueba que se utilizarán para evaluar los modelos y determinar en qué medida es bueno el rendimiento de dichos modelos

A menudo, un conjunto de datos de validación se utiliza también para refinar parámetros de modelado antes de probar un modelo.

La razón principal para la separación entre la prueba y el entrenamiento es garantizar que el modelo tenga un buen rendimiento con datos para los que no se ha entrenado. En particular, esto identifica el problema de sobreajuste, que ocurre cuando un modelo en particular parece predecir de forma excesivamente precisa utilizando los datos de formación. Sin embargo, cuando este modelo se coteja con otros datos de prueba, el rendimiento del modelo disminuye significativamente. Entonces se podría decir que el modelo tiene un sobreajuste por los datos de entrenamiento.

Normalmente, entre el 10% y el 30% de los datos se retienen como datos de prueba. Eso dependería de la disponibilidad general de los datos y del grado de sobreajuste que se pueda hace en los modelos. Podrían seleccionarse datos de prueba como subconjunto aleatorio de sus datos o, por ejemplo, basar la selección de datos en el tiempo (p. ej. tomando el último año de datos como subconjunto de prueba).

La matriz de confusión

Para evaluar la calidad de un clasificador binario, podemos generar una matriz de confusión utilizando nuestros datos de prueba. Sobre estos datos aplicaríamos el modelo para generar y producir resultados predichos, y también dispondríamos de los resultados reales en los datos. Tabularíamos una matriz con los cómputos de casos donde el modelo identificó el resultado como positivo (estándar en nuestro ejemplo) y el resultado real fue positivo (también estándar). Este es el cómputo «Positivo Verdadero». De forma similar, contamos el número de casos en los que el modelo predijo correctamente los negativos como «Verdaderos negativos». También contamos los errores donde el modelo predijo negativo pero el resultado fue positivo y viceversa. Esto genera una matriz de confusión como se muestra:

		Predicho	
Real		Positivo	Negativo
	Positivo	Verdaderos positivos	Falsos negativos
	Negativo	Falsos positivos	Verdaderos negativos

En una matriz de confusión, los falsos positivos también se conocen como errores de Tipo I y los falsos negativos se conocen como errores de Tipo II.

A continuación, se puede utilizar esta matriz de confusión para clasificar el modelo utilizando varios parámetros:

$$Precisión = \frac{Positivos\ Verdaderos + Negativos\ Verdaderos}{Casos\ Totales}$$

$$Sensibilidad = \frac{Positivos\ Verdaderos}{Positivos\ Verdaderos + Negativos\ Falsos} = \frac{Positivos\ Verdaderos}{Positivos\ Reales}$$

$$Especificidad = \frac{Negativos\ Verdaderos}{Negativos\ Verdaderos + Positivos\ Falsos} = \frac{Negativos\ Verdaderos}{Negativos\ Reales}$$

La medición de la exactitud es una medida general de exactitud. La sensibilidad indica cuántos de los positivos están identificados realmente como

positivos. La interacción entre estas variables indica en qué medida es bueno un modelo. Por ejemplo, si tenemos dos modelos que predicen los resultados de si los casos son estándar o no, podemos tabular una matriz de confusión para cada uno de ellos.

El modelo A es un modelo muy simple (y muy impreciso) que predice que cada caso será estándar.

Modelo A		Predicho	
		Estándar	No estándar
Real	Estándar	80	0
	No estándar	20	0

A partir de la matriz de confusión podemos evaluar que hubo 100 casos. Nuestro modelo predice que todos los casos son estándar. Nuestra sensibilidad es por tanto del 100% y nuestra exactitud, de un buen 80%. Sin embargo, nuestra especificidad es deficiente con un 0%, lo que indica un modelo de bajo rendimiento.

Está claro que es necesario considerar varias medidas para comprobar cómo de bueno es un modelo. Uno más realista podría tener un aspecto parecido a este (sobre los mismos datos):

Modelo B		Predicho	
		Estándar	No estándar
Real	Estándar	61	19
	No estándar	8	12

En este caso, la exactitud general es del 73%; la sensibilidad es del 76,25% y la especificidad, del 60%. Este parece un modelo razonable.

Puntuaciones y umbral del modelo

La mayoría de los modelos clasificadores no muestran solamente una clasificación binaria como resultado. Suelen producir una puntuación para clasificar los casos como positivos o negativos. La puntuación se suele convertir en forma porcentual. Esta puntuación no implica necesariamente una probabilidad verdadera, en especial para las técnicas de aprendizaje automático, y a menudo solo indican una clasificación de casos más que una probabilidad estricta.

Dado que cada caso tendría una puntuación, sería necesario asignar un umbral (o corte) en cuyo punto el resultado del modelo se consideraría positivo. Por ejemplo, un modelo podría producir varias puntuaciones para varios casos, y con un umbral del 80%, trataría un caso como caso estándar solo si la puntuación se encontrara por encima del 80%.

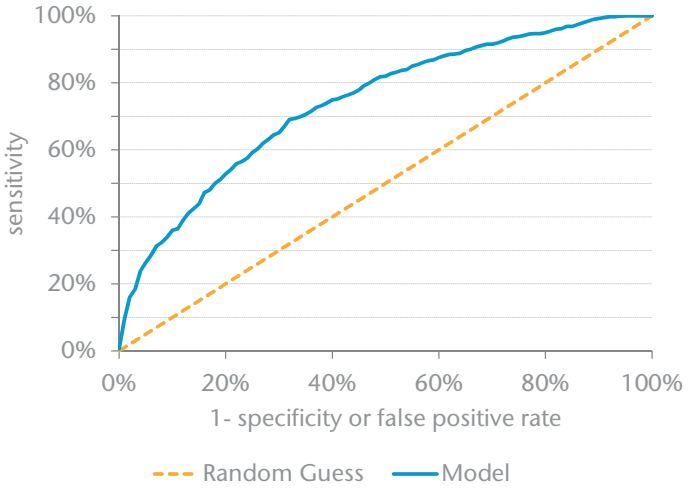
Cada uno de estos umbrales implica entonces una matriz de confusión específica. Por tanto, con un umbral del 0%, terminaríamos en la situación del modelo A mostrado más arriba; es decir, predeciremos cada caso como estándar, con la sensibilidad del 100% pero una especificidad del 0%. De forma similar, un umbral del 100% tiene como resultado la clasificación de todos los casos como no estándar, con un 0% de sensibilidad y un 100% de especificidad.

Curva ROC (Característica Operativa del Receptor)

Cambiar el umbral para un modelo en particular sobre todos los valores entre el 0% y el 100% permite trazar una curva de los diversos valores de especificidad y sensibilidad. La Característica Operativa del Receptor (ROC) traza la sensibilidad (eje y) frente a la especificidad (eje x) para cada valor del umbral. Obsérvese que 1–especificidad es también el índice de falsos positivos. La Figura 1 es un ejemplo de curva ROC de este tipo.

En la parte inferior izquierda vemos el caso en el que el umbral es del 100%. Esto es equivalente a nuestro modelo para predecir todos los casos como no estándar (la sensibilidad es del 0% pero

Figura 1 – Curva ROC



la especificidad es del 100%, o el índice de falsos positivos es del 0%). En la parte superior derecha tenemos la situación en la que todos los casos se predicen como estándar (la sensibilidad es del 100% pero la especificidad es del 0%, o el índice de falsos positivos es del 100%).

La línea diagonal discontinua representa el resultado esperado para un modelo que asigna de forma aleatoria puntuaciones a los casos. La línea azul representa los resultados de un modelo en particular que resulta mejor que las conjeturas al azar. La calidad general del modelo podría determinarse mediante la evaluación del área situada bajo la curva azul. En este caso, se calcula en un 74%, que representa un modelo razonablemente bueno.

Podemos tomar dos casos aleatorios: uno de los casos reales estándar y uno de los casos reales no estándar. Se ha demostrado que el área situada bajo la curva (74% en este ejemplo) es equivalente a la probabilidad de que el caso estándar tenga una puntuación más alta que el caso no estándar. Por ello, esta área proporciona una medida general sobre lo bueno que es el modelo para clasificar los casos estándar de los no estándar.

El modelo de conjeturas al azar presenta un área bajo la curva del 50%, de modo que se trata del requisito mínimo a partir del cual mejorar. Un modelo perfecto tendría un área bajo la curva de uno, y el modelo produciría un punto en la esquina superior izquierda de la Figura 1.

Generalmente, los modelos con un área bajo la curva de entre el 50% y el 60% no se consideran exitosos, y los modelos que superan el 90%, muy buenos. Un resultado en el rango del 70%-80% se considera de aceptable a bueno. Obsérvese que un modelo con un área bajo la curva menor del 50% debería de hecho invertirse, ya que, en efecto, predice el resultado contrario con más éxito; por lo general, el área situada bajo la curva varía entre el 50% y el 100%.

Coeficiente de Gini

El coeficiente de Gini también está relacionado con la evaluación de modelos de clasificación. En realidad, está directamente relacionado con el área situada bajo la curva ROC mencionada más arriba. El coeficiente de Gini se calcula a partir del área situada bajo la curva (ABC) como 2ABC–1.

Un área bajo la curva del 74% se convierte en un coeficiente del 48%, que es aceptable. Por tanto, el coeficiente de Gini oscila efectivamente entre el 0% y el 100%, si bien puede ser negativo, en cuyo caso el modelo debería en realidad invertirse, como en el caso de un área bajo la curva menor del 50%.

Comparación de modelos

Diferentes modelos y técnicas de modelado tendrán como resultado distintos rendimientos sobre datos y contextos de problemas diversos. Para comparar dos modelos diferentes –y seleccionar el mejor–, podemos comparar cada área bajo la curva ROC (o coeficientes de Gini). Podemos decir que un modelo A es mejor si su área situada bajo la curva es mayor que B.

Sin embargo, hay algunas limitaciones. Si trazamos los modelos y la curva ROC de A cruza la curva ROC de B, no podremos decir que A es siempre mejor. Con determinados niveles de sensibilidad y especificidad, B podría ser mejor. Si embargo, si la curva ROC de A no cruza la de B y su área bajo la curva es mayor, esto significa que, en esencia, A es siempre mejor que B.

Para los modelos cercanos en términos de sus curvas ROC y las áreas situadas bajo las mismas, podrían dominar consideraciones prácticas como los detalles de implementación.

Modelos conjuntos

Hay varias maneras de mejorar los modelos en función de la técnica implicada. Una que puede ser interesante considerar es la construcción de modelos conjuntos. Una vez que contamos con múltiples modelos que producen una puntuación para un resultado en particular, podemos comenzar a combinarlos de formas interesantes para producir puntuaciones conjuntas. Éstas pueden utilizarse para mejorar todavía más el área bajo la curva para estos modelos.

Tomemos, por ejemplo, un clasificador de Bosque Aleatorio y un modelo de regresión logística, que en ambos casos predicen riesgos estándar. Se puede calcular una nueva puntuación como el promedio de estos dos clasificadores y a continuación evaluarla como un modelo más. Por lo general, el área bajo la curva mejora para estos modelos conjuntos.

También podríamos construir un modelo de modelos ajustando un modelo de regresión logística combinado a los diversos modelos subyacentes para modelar el resultado final, utilizando esencialmente los datos para sugerir cómo deberían ponderarse los diversos modelos. Por lo general, una muestra de validación posterior de los datos (no utilizados para formación ni para pruebas) se mantendría para producir este tipo de modo conjunto.

Optimización del negocio

Dado un modelo con una curva ROC en particular, podemos decidir ahora la forma de aplicar este modelo en la práctica. Puesto que cada tipo de error (falsos positivos y falsos negativos) tendría costes y beneficios asociados al negocio, podemos estimar el punto en el cual el coste se minimiza o el beneficio se maximiza, estableciendo así el mejor umbral para cada modelo. A partir de estos resultados, podemos comparar el mejor rendimiento comercial de varios modelos para tomar una decisión definitiva sobre un modelo en particular.

En el ejemplo del modelado de decisiones de suscripción estándar, podemos considerar las implicaciones de cada categoría de resultados en el valor actual de beneficios por solicitud:

- Los verdaderos positivos (casos clasificados correctamente por el modelo como estándar) podrían ver un aumento en el valor por póliza, ya que veríamos mayores colocaciones de esta categoría gracias a que tendrían menores gastos de venta por póliza (mayor razón de conversión debido a un proceso armonizado orientado al cliente) y menores costes médicos y de suscripción.
- Los falsos positivos (casos clasificados incorrectamente como estándar) podrían enfrentarse a problemas de aumento de siniestros en relación a las primas. También hay riesgo de antiselección en este caso si los solicitantes comprenden la forma de influir en sus puntuaciones.
- Los falsos negativos (casos clasificados incorrectamente como no estándar) podrían ser muy similares al proceso actual, por lo que podríamos necesitar el uso de una ratio de conversión actual e índices de utilización para evaluar el valor actual de beneficios por solicitud (suponiendo que la suscripción siga nuestro enfoque actual al respecto).
- Los verdaderos negativos (casos clasificados correctamente como negativos) podrán ser de nuevo bastante coherentes con respecto a nuestro tratamiento actual de casos no estándar.

Dados los valores para cada uno de los anteriores, es posible estimar un umbral opcional para el uso en nuestro modelo que aportará un máximo valor por solicitud al negocio. Esto se correspondería con un punto simple en la curva ROC. El resultado seleccionado se debería someter a prueba de sensibilidad a las hipótesis, dado que muchas de las hipótesis realizadas a la hora de determinar el valor de los diversos contextos podrían ser subjetivas.

Si tenemos múltiples modelos (donde las curvas ROC se cruzan como antes), o un modelo para el cual no podemos calcular la curva ROC, podríamos comparar su mejor resultado obtenido por los diferentes modelos para determinar cuál es el mejor. Al comparar estos valores, podría ser necesario asegurarse de que se incluya el coste de implementar el modelo. Algunos modelos podrían implicar costes, basados en los datos que utilicen o bien debido a cuestiones de implementación técnica.

Conclusión

Este artículo ha descrito enfoques para evaluar el rendimiento de los modelos utilizados en problemas de clasificación.

Hemos realizado estas aclaraciones:

- Es necesario contar con muestras de datos representativas con fines de prueba.
- Es posible comparar el rendimiento de estos modelos utilizando medidas en una matriz de confusión.
- Se pueden estimar números, como especificidad y sensibilidad.
- También se podría considerar de forma más general el uso de la curva ROC y el área bajo la misma para obtener una impresión general de la calidad del modelo.
- Las técnicas conjuntas pueden utilizarse para combinar puntuaciones de varios modelos y producir mejores modelos.

- Una vez aplicados los valores de negocio esperados para varios resultados del modelo, puede ser más fácil decidir qué valor de equilibrio entre sensibilidad y especificidad sería el mejor para el negocio.
- Podrían considerarse cuestiones de tipo práctico, incluyendo hipótesis subjetivas a la hora de valorar distintos resultados y detalles de implementación técnica que podrían mostrar diferentes desenlaces.
- Cambiar el proceso podría tener como resultado comportamientos cambiados que invalidarían el modelado. En este contexto hay que evaluar el comportamiento particularmente antiselectivo.

Este resumen relativamente simple de este campo da una idea de la forma de medir objetivamente estos modelos y de cómo evaluar el valor que tienen para el negocio teniendo en cuenta su rendimiento.

Referencias

JAMES, G., WITTEN, D., HASTIES, T. y TIBSHIRANI, R. (2013). An Introduction to Statistical Learning. Springer.

MATLOFF, N. (2017). Statistical Regression and Classification: From Linear Models to Machine Learning. CRC Press.

FAWCETT, T. (2006) An introduction to ROC analysis. Pattern Recognition Letters 27, 861-874.

Acerca del author

Louis Rossouw *es Actuario de Investigación y Análisis en la oficina de Gen Re en Ciudad del Cabo, dando apoyo a Sudáfrica y el Reino Unido. Louis se unió a Gen Re en 2001 y ha trabajado previamente en el "pricing" individual y de grupo, el desarrollo de productos y las reservas. También ha trabajado durante dos años como Actuario Jefe Regional en la oficina de Gen Re en Singapur. Se puede contactar con Louis a través del número Tel. +27 21 412 7712 o del correo electrónico lrossouw@genre.com.*

The difference is...the quality of the promise®

genre.com | genre.com/perspective | Twitter: @Gen_Re

General Reinsurance AG
Theodor-Heuss-Ring 11
50668 Cologne, Germany
Tel. +49 221 9738 0
Fax +49 221 9738 494

General Reinsurance AG
Sucursal en España
Plaza Manuel Gómez Moreno, 2 – Planta 6
Edificio "Alfredo Mahou"
28020 Madrid
Tel. +34 91 722 4700
Fax +34 91 722 2619

General Reinsurance México S.A.
Paseo de la Reforma 350 - 6° Piso
Edificio Torre del Àngel, Col. Juárez
06600 México, D.F.
Tel. +52 55 9171 9200
Fax +52 55 9171 9260