Ajuste de Riesgo en el Sector de la Salud usando Técnicas de Minería de Datos

Álvaro J. Riascos Villegas
Mauricio Romero
Natalia Serna [1]

[1]Universidad de los Andes - UCSD - Quantil

Contenido

Contenido

Introducción

- La ley 100 de 1994 transformó el sistema colombiano de la salud en un mercado de aseguramiento competitivo.
- Tres factores fundamentales en la organización del mercado: POS, EPS, UPC.
- Dos sistemas: Contributivo y Subsidiado.
- Los afiliados del régimen contributivo aportan mensualmente al sistema una proporción de sus ingresos más un subsidio para financiar a los afiliados del régimen subsidiado.
- El FOSYGA recauda las contribuciones y luego las redistribuye.

Contenido

Ajuste de riesgo

- La redistribución de los recursos entre las EPS depende del riesgo de los individuos que afilian.
- Cuánto más riesgosos los afiliados, mayores deben ser los recursos dirigidos a su aseguradora.
- El pago per cápita que el gobierno hace a las EPS se conoce como Unidad de Pago por Capitación (UPC).

Ajuste de riesgo

- La UPC es la predicción del gasto anual en salud ponderado por los días compensados y ajustado a variables neutrales al ingreso que describen el perfil de riesgo de los asegurados y que se le paga a las EPS por cada uno de sus afiliados.

$$UPC_i = 360 \times \frac{x_i}{d_i} \tag{1}$$

donde x_i es la suma del costo de todos los servicios que el afiliado i reclama en el año y d_i son los días que estuvo afiliado o que compensó al sistema.

Ajuste de riesgo

- Entre la estimación de la UPC y la aplicación de los pagos hay un rezago de 2 años:
  - $t-2$: los prestadores de servicios atienden la demanda en salud.
  - $t-1$: las aseguradoras compilas la información de las reclamaciones y la envían al Ministerio.
  - t: el Ministerio estima la UPC y le paga a las EPS.
- El modelo de predicción del gasto anual en salud en el cual se basan los pagos de UPC debe tener un alto poder predictivo y ser estable en el tiempo.

Ajuste de riesgo

- Nótese que de este modelo dependen los ingresos de las aseguradoras y, por lo tanto, su solvencia financiera.
- Cuanto mayor sea la subestimación del gasto, mayor es el riesgo de insolvencia financiera del sistema.
- La sobreestimación del gasto también genera incentivos a la selección de riesgos: las EPS van a afiliar solo a aquellos individuos para los que el ingreso es alto pero el gasto es bajo (los más sanos).

Ajuste de riesgo

- El modelo actual de ajuste de riesgo del Ministerio es una regresión lineal:

$$\log(UPC_i) = \beta_0 + \sum_{j=1}^{K} \beta_j D_j + \varepsilon_i \tag{2}$$

  donde D es un conjunto de K variables dummy que indican algunas características de los afiliados como el sexo, el grupo de edad al que pertenecen, o la zona UPC donde residen; y ε_i es un choque aleatorio específico al individuo.
- El objetivo es comparar el poder predictivo del modelo del Ministerio contra los modelos de aprendizaje de máquinas.

Contenido

Marco empírico

- Datos: panel de atenciones de los afiliados activos en 2010 y 2011.
- Variables: demográficas como sexo, grupo de edad, y ubicación ($Demog$); 29 grupos de enfermedades de larga duración (Dx); indicador de hospitalización (H), consulta con especialista (E), e ingreso a UCI (U).
- La función que queremos estimar es:

$$y_{i,2011} = f_{2010}(Demog, Dx, H, E, U; \theta) \tag{3}$$

donde $y_{i,2011}$ es el gasto anual del individuo i durante 2011 ponderado por sus días compensados en ese mismo año y f_{2010} es una función de las características del afiliado durante 2010.

Marco empírico: Regresiones lineales

El primer ejercicio es la estimación del modelo (3) por mínimos cuadrados ponderados tal que:

$$y_{i,2011} = w_{i,2011} Y_{i,2011} \tag{4}$$

y,

$$w_{i,2011} = \frac{d_{i,2011}}{360} \tag{5}$$

Marco empírico: Aprendizaje de máquinas

1. Redes neuronales.
2. Bosques aleatorios.
3. Boosting de árboles.

Marco empírico: Aprendizaje de máquinas

Problema de regresión

Una de las técnicas más comunes para entrenar modelos basados en el aprendizaje de máquinas cuando el problema es de regresión y la variable a predecir es altamente dispersa, consiste en transformar la variable a una escala $\{0,1\}$.

En este caso se divide el gasto anual en salud entre 10^{10} que es mayor al máximo gasto observado y luego las predicciones se multiplican por ese mismo valor para devolverlas a su escala original.

Marco empírico: Aprendizaje de máquinas

Usando validación cruzada se calculan los siguiente parámetros:

- Para la red neuronal: el número de neuronas y el decaimiento de los ponderadores de las variables al pasar de una capa interna a una externa.
- Para el modelo de bosques aleatorios: el número de árboles.
- Para el modelo de boosting de árboles: el mínimo de observaciones en los nodos no terminales de los árboles, la contracción de los árboles entrenados secuencialmente, el grado de interacción entre las variables y el número de árboles.

Marco empírico: Aprendizaje de máquinas

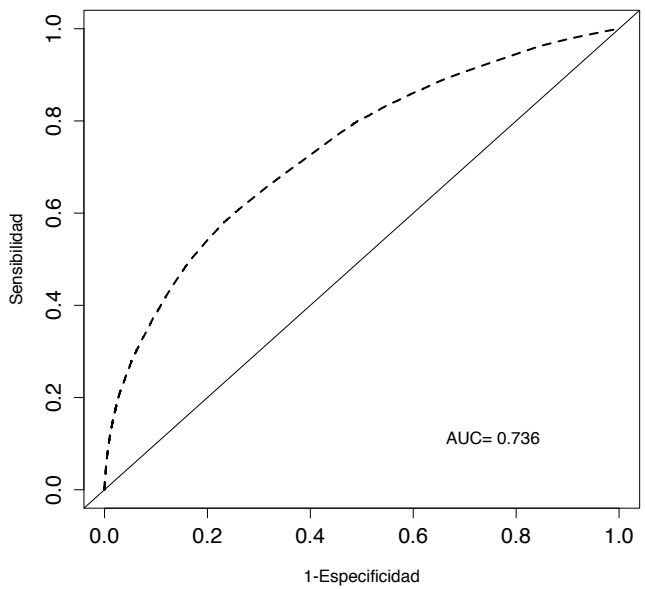
En los modelos de aprendizaje de máquinas se siguen dos especificaciones adicionales:

- Se controla por la probabilidad de reclamar un servicio.
- Se hace un proceso de selección de variables.

Marco empírico: Aprendizaje de máquinas

El modelo para la probabilidad de reclamar un servicio es una red neuronal de 3 capas y 5 neuronas ajustada sobre (Demog, Dx, H, E, U). La variable dependiente toma el valor de 1 si el gasto anual es menor a 200,000 pesos y 0 en caso contrario. La curva ROC se muestra a continuación:

Figura: Curva ROC - Modelo de reclamación



Marco empírico: Aprendizaje de máquinas

El proceso de selección de variables consiste en fijar un umbral para la importancia relativa a la Breiman (1993) en el modelo de boosting de árboles. El conjunto final contiene 21 variables y se excluye la probabilidad de reclamar un servicio. Este conjunto es 90 % más pequeño que el conjunto sobre el que se estima el modelo del Ministerio.

Marco empírico: Medidas de ajuste

- Error cuadrático medio (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i(y_i - \hat{y}_i)^2}$$

- Error medio absoluto:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} w_i|y_i - \hat{y}_i|$$

- Ratio de predicción para el gasto anualizado (PR - anualizado):

$$PR = \frac{\sum_{i=1}^{N} \hat{y}_i}{\sum_{i=1}^{N} y_i}$$

Marco empírico: Medidas de ajuste

- Ratio de predicción para el gasto no anualizado (PR - no anualizado):

$$PR = \frac{\sum_{i=1}^{N} w_i \hat{y}_i}{\sum_{i=1}^{N} w_i y_i}$$

- R-cuadrado:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} w_i (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} w_i (y_i - \overline{y}_i)^2}$$

Marco empírico: Medidas de ajuste

- Se construyen dos bases de datos mutuamente excluyentes, cada una con 500,000 afiliados: entrenamiento y prueba.
- En la base de entrenamiento se ajustan todos los modelos y en la de prueba se calculan las métricas de ajuste.

Contenido

Resultados

- Dos etapas: incluye probabilidad de reclamar servicios.
- FS: incluye variables del proceso de selección.

Resultados

Cuadro: Ajuste fuera de muestra distribución completa

Parámetros	Modelo	RMSE	MAE	PR anual	PR - no anual	R^2
	WLS UPC	3,506,658	720,587	0.896	0.999	1.57
	WLS UPC + Dx	3,440,928	694,404	0.892	0.999	5.23
	WLS UPC + Dx + H	3,437,175	694,005	0.894	1.000	5.45
	WLS UPC + Dx + H + E	3,435,470	691,169	0.892	0.999	5.53
	WLS UPC + Dx + H + E + U	3,431,842	688,771	0.892	0.999	5.73
	WLS UPC x H x E x U + Dx	3,432,097	683,209	0.893	0.999	5.71
5 + 0.05	ANN Demog + Dx (Dos etapas)	3,470,597	816,431	1.072	1.203	3.59
5 + 0.05	ANN Demog + Dx + H (Dos etapas)	3,455,518	785,698	1.058	1.189	4.42
5 + 0.05	ANN FS	3,455,366	774,190	1.064	1.179	4.43
680	RF Demog + Dx + H + E + U (Dos etapas)	3,607,413	735,027	0.981	1.097	-4.16
680	RF FS	3,465,301	712,820	0.975	1.087	3.88
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Dos etapas)	3,436,299	719,029	0.997	1.111	5.48
680 + 3 + 0.1 + 50	GBM FS	3,431,044	721,168	1.002	1.115	5.77

Resultados

Cuadro: Ajuste fuera de muestra en el primer quintil

Parámetros	Modelo	RMSE	MAE	PR anual	PR no anual
	WLS UPC	585,387	423,863	344.0	285.9
	WLS UPC + Dx	544,839	344,251	285.0	232.5
	WLS UPC + Dx + H	549,384	331,398	275.4	223.8
	WLS UPC + Dx + H + E	541,922	314,979	272.6	212.6
	WLS UPC + Dx + H + E + U	560,868	317,613	261.6	214.4
	WLS UPC x H x E x U + Dx	552,219	319,218	267.2	215.6
5 + 0.05	ANN Demog + Dx (Dos etapas)	674,653	401,014	331.3	270.3
5 + 0.05	ANN Demog + Dx + H (Dos etapas	633,907	362,857	301.3	244.7
5 + 0.05	ANN FS	646,313	439,889	369.8	296.7
680	RF Demog + Dx + H + E + U (Dos etapas)	792,871	347,280	291.2	234.5
680	RF FS	658,254	361,954	304.7	244.4
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Dos etapas)	630,485	364,222	307.6	245.9
680 + 3 + 0.1 + 50	GBM FS	652,681	376,233	317.1	253.9

Resultados

Cuadro: Ajuste fuera de muestra en el quintil superior

Parámetros	Modelo	RMSE	MAE	PR anual	PR no anual
	WLS UPC	7,749,235	1,920,486	0.291	0.335
	WLS UPC + Dx	7,580,659	1,983,269	0.367	0.426
	WLS UPC + Dx + H	7,569,594	1,974,414	0.380	0.440
	WLS UPC + Dx + H + E	7,565,491	1,963,559	0.388	0.450
	WLS UPC + Dx + H + E + U	7,555,009	1,965,934	0.390	0.452
	WLS UPC x H x E x U + Dx	7,555,760	1,981,378	0.393	0.456
5 + 0.05	ANN Demog + Dx (Dos etapas)	7,576,299	2,045,276	0.444	0.517
5 + 0.05	ANN Demog + Dx + H (Dos etapas)	7,558,045	2,000,360	0.454	0.526
5 + 0.05	ANN FS	7,582,293	1,962,318	0.412	0.474
680	RF Demog + Dx + H + E + U (Dos etapas)	7,780,452	2,118,572	0.445	0.520
680	RF FS	7,580,672	1,988,824	0.424	0.490
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Dos etapas)	7,532,498	1,988,610	0.436	0.505
680 + 3 + 0.1 + 50	GBM FS	7,517,520	1,961,026	0.430	0.500

Incentivos a la selección de riesgos

Cuadro: Incentivos a la selección de riesgos

Parámetros	Modelo	Afiliados (%)	Ganancias ($)*	Ganancias (%)
	WLS UPC	70.39	180,046	73.39
	WLS UPC + Dx	67.82	173,432	72.32
	WLS UPC + Dx + H	65.66	173,464	72.29
	WLS UPC + Dx + H + E	64.02	172,739	71.93
	WLS UPC + Dx + H + E + U	64.50	172,051	71.92
	WLS UPC x H x E x U + Dx	67.52	175,454	71.55
5 + 0.05	ANN Demog + Dx (Dos etapas)	57.96	234,627	76.19
5 + 0.05	ANN Demog + Dx + H (Dos etapas)	60.15	224,829	74.62
5 + 0.05	ANN FS	57.01	251,677	76.68
680	RF Demog + Dx + H + E + U (Dos etapas)	67.85	198,359	73.97
680	RF FS	69.77	191,198	72.33
680 + 3 + 0.1 + 50	GBM Demog + Dx + H + E + U (Dos etapas)	70.53	196,517	72.38
680 + 3 + 0.1 + 50	GBM FS	70.21	197,569	72.58

Contenido

Conclusiones

- Los modelos basados en el aprendizaje de máquinas permiten hacer una mejor redistribución de los recursos del sistema de salud colombiano.
- El modelo lineal del gobierno puede mejorarse significativamente si se condiciona ex ante a la morbilidad de los afiliados con los 29 grupos de enfermedades de larga duración.
- El ajuste ex ante por morbilidad con estos grupos de enfermedades es mejor que el ajuste ex post con la Cuenta de Alto Costo.