CASO DE ESTUDIO FINANZAS

Piedrahita Allison; Ramírez Anyi; Vergara María

Departamento de Ingeniería Industrial, Universidad de Antioquia.

Medellín, Colombia.

RESUMEN

La tarificación de los seguros médicos presenta diversos desafíos; Sin embargo, este artículo presenta una metodología que no solo permite la segmentación de usuarios y la predicción del costo médico mediante modelos analíticos, sino también la identificación de variables relevantes para la tarificación de seguros de salud. Para lograrlo, se evaluaron los algoritmos de regresión lineal múltiple, árbol de decisión y bosque aleatorio, dada la importancia de la interpretabilidad en la solución analítica. Aquí, el criterio y conocimiento de los analistas jugaron un papel fundamental al realizar una selección adecuada de variables para alimentar los modelos. Finalmente, se destaca la relevancia de la edad en la determinación de la tarifa, así como la cantidad de expuesto, así como cuenta información que se le pueda preguntar al usuario para obtener una correcta tarifación del seguro.

1. INTRODUCCIÓN

En la actualidad, el sector salud se han venido presentando dificultades con respecto a la prestación del servicio, tanto así que las personas que desean ser atendidas deben hacer largas filas y desde muy temprano para conseguir atención médica, adicional a esto se presentan numerosas quejas por las demoras en la atención y sobre todo que muchos servicios no pueden llevarse a cabo; por tal razón aparecen los planes voluntarios de salud, los cuales son presentados como una alternativa para ampliar la cobertura asistencial, la cual va más allá del plan obligatorio de salud.

En el mundo de los seguros, el riesgo es el factor principal. Este representa la posibilidad de sufrir perdidas o daños la cuales se pueden medir en términos de dinero. El riesgo se vuelven un concepto de suma importancia cuando se habla de seguros de salud, ya que el riesgo se ve reflejado en eventos que requieren de atención médica, tales como enfermedades, accidentes, exámenes médicos o partos.

El cuantificar estas situaciones no solo dice cuanto es el nivel de riesgo, sino que también dice cuál es el riesgo a nivel financiero, esto beneficia la planeación de cómo proteger la salud y el dinero de las personas aseguradas

2. METODOLOGÍA

En las siguientes secciones se describe cada uno de los pasos desarrollados.

2.1.DISEÑO DE LA SOLUCIÓN

Con las bases de datos proporcionadas por el departamento de actuaria, se plantean los siguientes objetivos para desarrollar el caso de estudio:

- Segmentar las personas de las bases de datos para seleccionar aquellas que aun tenga el servicio vigente del seguro de salud, es decir, que no hayan cancelado el plan de seguro de salud, con el fin de enfocar la solución analítica.
- Predecir el costo medico (valor de utilización) por medio de un modelo analítico basado en variables designadas como significativas de cuatro meses históricos.

- Identificar las variables que pueden ser de fácil acceso a la hora de hacer una tarifación para un seguro de salud.

Con base a lo anterior, se plantea la solución analítica de la Figura 1, en la cual se evidencian los objetivos anteriores y las condiciones de tiempo de la información.

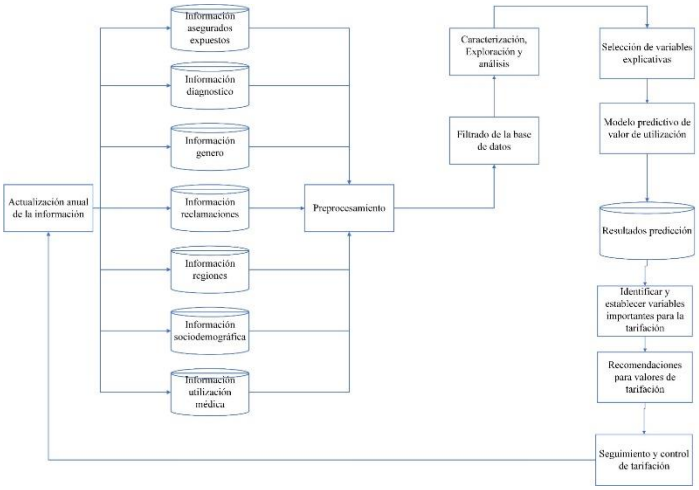

Figura 1. *Diseño de la solución analítica.*

2.2.LIMPIEZA Y TRANSFORMACIÓN

Para el desarrollo de la solución, se cuenta con siete bases de datos:

- *"BD_Asegurados_Expuestos":* contiene la información general el id de la póliza, fecha de inicio, fecha de fin y fecha de cancelación. Esta base de datos tiene 231.520 registros y cinco variables.
- *"BD_Diagnostico":* contiene información de variables médicas que describen un tipo de diagnóstico. Esta base de datos tiene 3.411 registros y dos variables.
- *"BD_Genero":* contiene la información básica de los tipos de géneros. Esta base de datos tiene tres registros y dos variables.
- *"BD_Reclamaciones":* Contiene información sobre los tipos de reclamaciones. Esta base cuenta con 38 registros y dos variables.
- *"BD_Regional":* contiene información acerca de la región. Esta base tiene seis registros y dos variables.
- *"BD_SocioDemograficas":* Contiene información general del paciente que presta el servicio. Esta base tiene 225.776 registros y nueve variables.
- *"BD_UtilizacionesMedicas":* Contiene información acerca del servicio por el cual el paciente hizo la reclamación, además del número de veces que se usó el servicio y el valor de la utilización. Esta base tiene 800.232 y seis variables.

Para realizar la limpieza de la información, se realizan los siguientes pasos para cada una de las bases de datos:

- Separación de variables/columnas según el separador de datos (",")
- Se eliminan categorías de algunas variables que se muestran como "sin información".
- Se reescriben las categorías de la variable "regional_id" dado que algunas están mal escritas.

Se procede a realizar la unión de las bases de datos por partes, dado que no todas tienen la misma variable tipo "key".

- Se une la base sociodemográfica con el género, por medio de "Sexo_cd".
- Luego se une la base sociodemográfica con la base de regional por medio de "Regional_id"
- Se une la base utilizaciones medicas con diagnóstico, por medio de la variable "Diagnostico_codigo"
- Luego se une la base número de utilizaciones y reclamaciones por medio de la variable "Reclamacion_cd"
- Para finalizar, se unen la bases sociodemográficas y utilizaciones medicas para generar una sola base a partir de la variable "Afiliado_Id"

La unión se realiza por medio de la función "merge" bajo el método "inner" el cual lleva a cabo una

intersección de los datos de las bases para extraer la información contenida en cada una de las bases. Teniendo en cuenta lo anterior, se obtiene una base de datos general con 823.617 registros y 12 variables.

Se crean variables que cuantificar ciertos tiempos y al mismo tiempo para poder filtrar la base de datos, presentadas a continuación.

- La variable edad que corresponde al tiempo que transcurre desde la fecha de nacimiento hasta que toma el seguro
- La variable tiempo de reclamación corresponde al tiempo que transcurre desde el inicio del seguro hasta que reclamo.
- La variable tiempo fin corresponde al tiempo que ha trascurrido desde que reclame hasta que se terminó el seguro.

Con esto se filtra la base con la variable edad, para que no haya números negativos, también se aplica este filtrado a tiempo de reclamación y a tiempo fin de reclamación, para que la base no tenga datos inconsistentes. Luego se eliminan los registros que tengan la misma fecha de inicio y de fin, pues corresponde a una venta ficticia, además también se eliminan los registros que corresponden a un numero de utilizaciones en cero. Para finalizar, con la fecha de cancelación se genera otra variable que corresponde a si el usuario canceló o no el servicio (Si/No) de tal forma que eliminamos a las personas que ya no tengan el servicio activo, pues ya no tiene un nivel de expuesto de importancia para la aseguradora. Luego de aplicar todos estos filtros y eliminación de datos que fueran incoherentes se obtiene una base de 460.934 y 18 variables.

2.3.ANÁLISIS EXPLORATORIO

Después de realizar un análisis para comprender las características de los usuarios y de las variables de la base de datos general, se obtiene que:

- Los registros de edad presentan una media de 39 años y una mediana de 40 años. Además, existen registros de personas entre los 0 y 105 años.
- El 70% de los usuarios hace una única utilización o menos y el 30% restante hace una única utilización o más.
- El tiempo promedio de reclamación corresponde a 229 días.
- El 50% de los registros corresponden a personas del género femenino.
- Existen en los registros 439.823 personas que no padecen de cáncer.
- 456.010 personas no padecen de Enfermedad pulmonar obstructiva crónica (EPOC).
- 441.350 personas no padecen de diabetes.
- 390.604 personas no padecen de hipertensión.
- Existen 455.808 personas que no padecen de enfermedad cardiovascular.
- El tipo de reclamación más frecuente es la R10, la cual corresponde a la consulta externa con un total de 181.608 veces.
- El valor de utilización se encuentra entre 10 pesos y 210.263.900.000.000 pesos, con una desviación de 1.688.144.000.000 pesos
- Valor de utilización y numero de utilizaciones presentan valores muy externos y diversos entre sí.
- Se observa una correlación fuerte entre la edad y la hipertensión, la diabetes, enfermedad, cardiovascular, el EPOC, cáncer.
- De las variables anteriores las que tienes mayor correlación son la edad y la hipertensión con un índice de correlación de 0.42.

Para el tratamiento de los datos atípicos asociados a errores de digitalización o registro en las variables valor de utilización y número de utilización, se hace una imputación de valores extremos (Aquellos que se encuentran fuera de tres veces el rango intercuartílico, es decir, entre el primer cuantil (Q1) y el tercer cuantil (Q3)) con la finalidad de evitar sesgos en los modelos analíticos. Para la imputación se toma el valor máximo o mínimo de los registros sin tener en cuenta los atípicos, y se reemplazan dichos valores extremos, se hace la imputación usando este método con la finalidad de no afectar las características que representan los registros de los pacientes. Después de realizar la

imputación de datos para las variables mencionadas, se obtienen los resultados de las siguientes figuras.

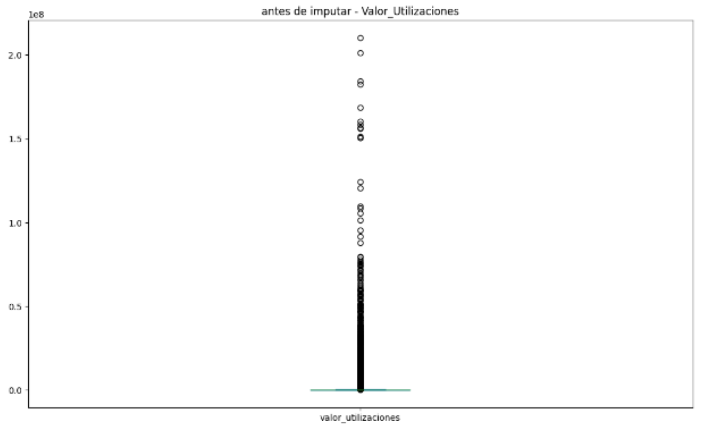

Figura 2. *Antes de imputar datos atípicos.*

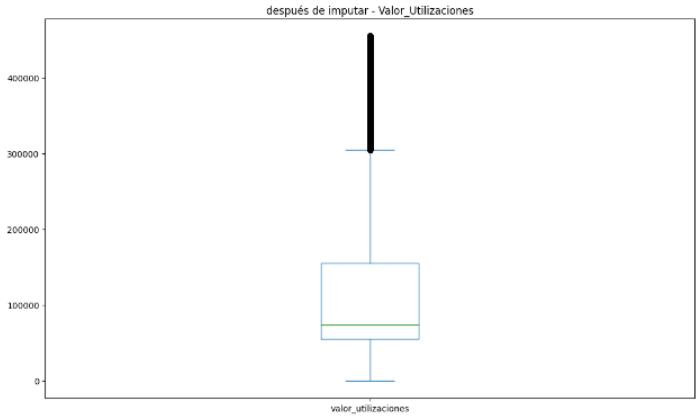

Figura 3. *Después de imputar datos atípicos.*

Al hacer esta imputación, se puede observar que no se cambia la distribución de los datos, y que estas tienen una forma más simple de ser modeladas.

2.4.SELECCIÓN DE ALGORITMOS Y VARIABLES

Teniendo en cuenta que la base datos general, es muy densa se descarta el uso del algoritmo Extreme Gradient Boosting, debido al consumo computacional que en si el modelo solo requiere.

Por otro lado, teniendo en cuenta la importancia de la interpretabilidad de la solución analítica, se plantea el uso de los siguientes algoritmos iniciales para evaluar su desempeño: regresión lineal múltiple, árbol de decisión y bosque aleatorio.

Antes de la construcción de los modelos, se eliminan aquellas variables que aportan información administrativa pero no generan valor en la solución analítica, tales como afiliado_id y póliza_id. además, se elimina la variable tiempo de fin, debido a su alta correlación con tiempo de reclamación. Asimismo, se evidencia que las variables de diagnóstico y reclamación contienen muchas categorías, lo cual, aumenta la dificultad en la construcción del modelo y, por lo tanto, se eliminan de la base de datos.

Posteriormente, se dejan las variables que sean de tipo numérica que permitan obtener un mejor desempeño de los modelos, asimismo, se escalan las variables con el fin de normalizar los datos y evitar sesgos en los resultados.

Teniendo en cuenta lo anterior, se construyen los modelos descritos para dos condiciones especiales: usando todas las variables de la base de datos general y usando selección de variables a criterio de los analistas, teniendo en cuenta que la hora de la tarifación puedan ser variables de fácil acceso al preguntarle a los usuarios.

Para evaluar el desempeño de los modelos construidos, se tienen en cuenta cuatro métricas de desempeño: el MAE para conocer el promedio de los errores en la predicción y para abarcar los resultados de valores atípicos, el MSE para dimensionar la variabilidad de los residuales de las predicciones, el RMSE para dimensionar la desviación estándar de los errores de las predicciones, el MAPE para identificar el porcentaje de error en las predicciones y el R^2 (R cuadrado) que proporciona información sobre la proporción de la variabilidad en la variable dependiente que es explicada por el modelo de regresión. En el análisis, se le da prioridad a la métrica MAE, debido a que permite obtener una mayor interpretabilidad de los resultados de los modelos.

Para la construcción de los modelos, se tiene en cuenta una división de datos de 70% entrenamiento y 30% validación, debido a se tienen 460.934 registros y el modelo debe tener la mayor cantidad de ejemplos

posibles para realizar una buena predicción. Teniendo en cuenta lo anterior, se obtienen las siguientes métricas de desempeño en el entrenamiento y en la validación para los modelos inicialmente construidos.

	Entrenamiento			Validación		
Modelo	**RL**	**Árbol**	**Bosque**	**RL**	**Árbol**	**Bosque**
MAE [$]	99.144	84.959	84.967	98.764	84.590	84.540
RMSE [$]	131.18	142.454	141.961	130.738	142.005	141.488
MAPE [%]	178.37	92,9	94.10	172.5	89.6	90.93
R^2 [%]	3,6	-13,7	-12,9	3,7	-13.6	-12,8

Tabla 1. *Desempeño de los modelos sin selección de variables.*

Se observa que el desempeño computacional del árbol de decisión y el bosque aleatorio es alto, ya que los resultados se obtienen en aproximadamente 30 minutos.

De manera adicional, se observa que los modelos presentan un porcentaje de error superior al 89%, lo cual puede asociarse a los valores de utilización atípicos identificados desde el análisis exploratorio. No obstante, se identifica que el promedio de error en la regresión lineal es 99.144,874 en entrenamiento y 98.764,27 en la validación, en el árbol de decisión los resultados son de 83.323,317 y 83.471,474 para entrenamiento y validación respectivamente, por último, en el bosque aleatorio, se obtienen promedios de error de 83.503,211 en entrenamiento y 83.621,176 en validación.

2.5.SELECCIÓN DEL MODELO

Teniendo en cuenta lo anterior, se selecciona la regresión lineal como modelo de implementación para el desarrollo del tarifario, debido a que tiene un desempeño computacional superior y unas métricas no muy lejanas a las de los demás modelos.

Teniendo en cuenta, que la idea del tarifario es poder obtener información de importancia que ayude a predecir el valor de la prima comercial, se hace una selección de variables para determinar el nivel de riesgo del usuario, para ello se toman en cuenta las siguientes variables: Regional, cáncer, EPOC, diabetes, hipertensión, enfermedad cardiovascular y la edad. No se toma en cuenta la variable de sexo, debido a que como el costo se mutualiza, el sexo queda discriminado

	Entrenamiento	Validación
Modelo	**RL**	**RL**
MAE [$]	101.816	101.472
RMSE [$]	132.920	132.541
MAPE [%]	184,81	179,01
R^2 [%]	1	1

Tabla 2. *Desempeño de los modelos con selección de variables.*

Al bajar el número de variables, vemos que no hay mucha diferencia en las métricas de desempeño, lo cual hasta resulta beneficioso, puesto que tenemos un modelo que se explica más fácil en un número pequeño de variables.

2.6.AFINAMIENTO DE HIPERPARÁMETROS

No se hace afinamiento de hiperparámetros, puesto que en la regresión lineal los parámetros son los coeficientes asociados a cada variable independiente, y estos son estimados directamente a través del modelo.

2.7.ANÁLISIS DEL MODELO

Con el modelo de regresión lineal construido, se evalúan las métricas por medio de la validación cruzada con un divisor CV de 100, y se obtienen los resultados de la Tabla 5 y la Figura 3.

Etapa	Entrenamiento	Validación
MAE [$]	101.636,75	101.656,57
RMSE [$]	132.806,52	132.611,18
MAPE [%]	182,77	182,81

Tabla 5. *Desempeño del modelo regresión lineal*

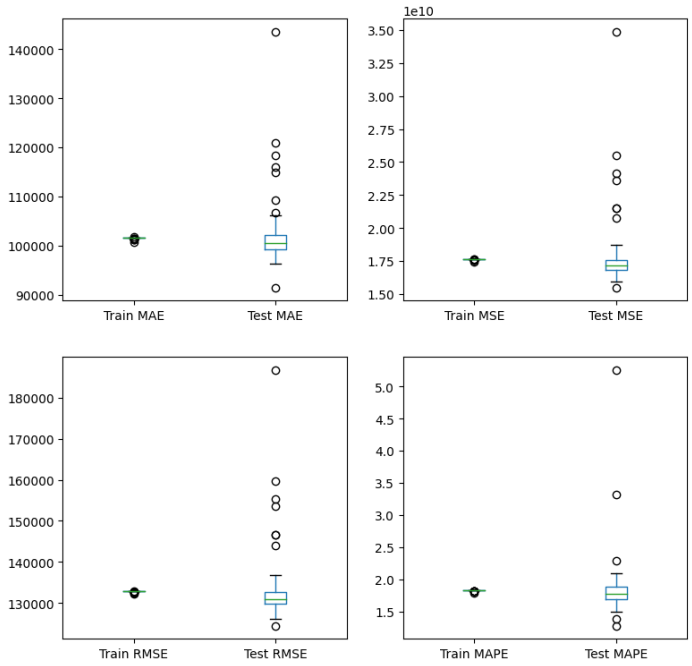

Figura 4. *Desempeño del modelo regresión lineal en gráficos de bigotes.*

En los resultados se evidencia un desempeño del modelo con tasas de entrenamiento y validación cercanas, lo cual indica que el modelo no presenta sobreajuste ni subajuste. No obstante, se evidencian métricas de porcentaje de error muy altas debido a los valores atípicos de la variable objetivo, a pesar de esto, se evidencia que el promedio de error es de 101.636,75 pesos en entrenamiento y 101.656,57 pesos en validación, lo cual se considera aceptable desde el punto de vista de las autoras.

Los coeficientes obtenidos en el modelo de regresión lineal son los siguientes para cada una de las variables:

Variable	Coeficiente Beta [$]	Interpretación
regional_id	-$3.024,342	Dejando las demás variables intactas, un incremento de código de la regional disminuye el costo médico en –$30.243,42.
cancer	$2.875,981	Dejando las demás variables intactas, el padecimiento de cáncer por parte de un paciente (1) incrementa el costo médico en $2.875,98.
epoc	$418,912	Dejando las demás variables intactas, el padecimiento de EPOC por parte de un paciente (1) incrementa el costo médico en $418,91.
diabetes	$1.430,525	Dejando las demás variables intactas, el padecimiento de diabetes por parte de un paciente (1) incrementa el costo médico en $1.430,53,91.
hipertension	-$2.029,326	Dejando las demás variables intactas, el padecimiento de hipertensión por parte de un paciente (1) disminuye el costo médico en $2.029,33.
enf_cardiovascular	$393,145	Dejando las demás variables intactas, el padecimiento de una enfermedad cardiovascular por parte de un paciente (1) incrementa el costo médico en $393,14.
edad	$12.449,589	Dejando las demás variables intactas, el incremento de un año en la edad del paciente incrementa el costo médico en $12.449,59.
Intercepto	$134.089,911	Dejando las demás variables nulas, se obtiene que el mínimo valor de costo médico es de $134.089,91.

Tabla 5. *Desempeño del modelo regresión lineal*

Teniendo en cuenta lo anterior, se observa la importancia de la edad en la determinación de la tarifa, debido a que es el coeficiente más alto en comparación con las demás variables, un incremento en la edad impacta directamente el valor del seguro, por ejemplo, una persona con 50 años tendría un costo basado solo en la edad de 622.479,45.

2.8.DESPLIEGUE DEL MODELO

Para la implementación del modelo construido para la tarifación se tienen en cuenta los siguientes aspectos:

- El modelo se entrenará con los 12 meses anteriores, para predecir la tarifa del año siguiente y así de forma consecutiva, la actualización de los datos será de forma anual. Esta contiene información de los usuarios del año anterior y además del valor de utilización, con el fin de hacer predicciones para los próximos años de los futuros usuarios del seguro médico.

- Se entregará un tarifario anual donde se va a predecir el valor del costo médico, donde se tendrá en cuenta la cantidad de usuarios expuesto a las condiciones determinadas por las variables seleccionadas en el año anterior.
- De manera adicional, el tarifario determina la prima pura teniendo en cuenta la cantidad de personas que están expuestas a las mismas condiciones dentro de un rango de edad que se establece de cada 5 años (prima pura = costo médico / expuestos).
- Posteriormente, teniendo en cuenta el porcentaje que representa la prima pura y los otros gastos, se determina la prima comercial para el usuario que desea acceder a los seguros de salud (prima comercial = prima pura / (1- porcentaje otros gastos).

3. RESULTADOS

Se presenta varios escenarios para predecir el valor de la prima comercial en el tarifario:

Variables condicionantes	
regional_id	40
cancer	0
epoc	0
diabetes	1
hipertension	0
enf_cardiovascular	0
rango_edad	[70,75)
Edad	**73**
Suma de expuestos	3

% Prima Pura	70%
% Otros Gastos	30%

Costo Médico	$	923.366,77
Prima Pura	$	307.788,92
Prima Comercial	$	439.698,46

Figura 3. *Escenario 1 en el tarifario*

Variables condicionantes	
regional_id	10
cancer	0
epoc	0
diabetes	1
hipertension	1
enf_cardiovascular	0
rango_edad	[55,60)
Edad	**56**
Suma de expuestos	72

% Prima Pura	70%
% Otros Gastos	30%

Costo Médico	$	800.424,68
Prima Pura	$	11.117,01
Prima Comercial	$	15.881,44

Figura 4. *Escenario 2 en el tarifario*

4. CONCLUSIONES

De los resultados obtenido, se evidencia que:

- La edad es un factor de vital importancia vemos que a mayores rangos de edad el costo medico aumenta de forma considerable
- Es importante tener en cuenta que la cantidad de expuesto, debido a que si hay pocos expuestos la mutualización de costo es baja
- Tener en cuenta información que se le pueda preguntar al usuario a la hora de fijar un valor de tarifación debe ser determinante, puesto que, a partir de preguntas simples como condiciones de salud, pueden dar diferencias significativas en el costo médico.
- Es importante resaltar que, aunque las métricas de los modelos no son las mejores del tarifario se pueden obtener valores coherentes a la hora de fijar un valor para el seguro.
- La regresión lineal es un modelo sencillo, sin embargo, muy poderoso porque en estos casos donde la cantidad de información es alta su procesamiento es rápido.

5. REFERENCIAS

[1]. *Fundación mapfre. (s.f). Diccionario de Seguros.* *https://www.fundacionmapfre.org/publicaciones/diccionario-mapfre-seguros/*

[2]. Fasecolda. (s.f). Seguro de salud. [vida y personas]. https://www.fasecolda.com/ramos/vida-y-personas/los-seguros/seguro-de-salud/