Algoritmo de Regresión Lineal para la Predicción de Gastos Médicos de una Persona en una Aseguradora Médica

German López Rodrigo

I. Resumen

El presente artículo pretende mostrar al lector de cualquier área, las bondades de utilizar el algoritmo de regresión lineal para la predicción de datos. Para ello se recurre al desarrollo de un caso aplicado al desempeño financiero de una aseguradora médica la cual desea pronosticar los gastos médicos de la población asegurada con la meta de poder de recolectar un valor superior en sus ingresos.

II. Palabras clave

Regresión lineal, Aprendizaje Automático, Gastos médicos, Algoritmo predictivo.

III. Descripción del problema

Debido a las actuales restricciones presupuestarias en el sector salud, se requiere de sistemas que permitan gestionar el gasto médico de manera más eficiente. Para ello, las aseguradoras médicas deben de optar por implementar algoritmos y técnicas que sean capaces de proporcionar una predicción de los gastos médicos futuros de la población asegurada con fin de recolectar un valor superior en sus ingresos.

- **Pregunta de investigación**

  ¿Qué información registrada de las personas en el sector salud se puede utilizar para predecir sus gastos médicos?

- **Objetivo**

  Determinar una ecuación que permita predecir los gastos médicos de un asegurado mediante la implementación del algoritmo de regresión lineal.

- **Fuente de Datos**

  Insurance [1] es una fuente de datos que consta de 1300 observaciones y 7 características, de las cuales 4 son numéricas (edad, IMC, hijos y gastos) y 3 características nominales (sexo, fumador y región). Esta fuente de datos es la que será utilizada para cumplir con el objetivo principal del presente artículo de investigación. Cabe aclarar que la fuente de datos fue dividida en dos conjuntos: un conjunto de 1000 observaciones para construir el modelo de regresión lineal y otro conjuntó de 300 observaciones para realizar las pruebas del modelo de regresión lineal.

IV. Introducción

Conforme pasa el tiempo, la tecnología avanza cada vez más y junto con esta, el ser humano busca satisfacer sus diferentes necesidades aplicándola en diferentes áreas de interés. Una de estas grandes aplicaciones, sin duda alguna, es la de poder hacer predicciones o pronósticos de eventos futuros de acuerdo al comportamiento observable de un conjunto de mediciones, observaciones o experimentos pasados. Por ejemplo, en el ámbito financiero se utiliza mucho este tipo de técnicas y algoritmos para predecir o pronosticar el comportamiento de sus ingresos con el fin de ayudar en la toma de decisiones. Actualmente debido al incremento del gasto sanitario en los países desarrollados y al limitado presupuesto que se le asigna al sector médico, las aseguradoras médicas se han percatado que requieren implementar este tipo de técnicas y algoritmos para predecir o pronosticar los gastos médicos futuros de su población asegurada con el fin de recolectar un valor superior en sus ingresos.

La regresión es un método de aprendizaje estadístico útil y ampliamente utilizado para la construcción de modelos que explican o representan la dependencia entre una variable o múltiples variables independientes (X) con una variable dependiente (Y). Existen dos tipos de regresión lineal: regresión lineal simple y regresión lineal múltiple; la regresión lineal simple se caracteriza por depender solo de una variable independiente y la regresión lineal múltiple se caracteriza por depender de múltiples variables independientes. El algoritmo de aprendizaje automático que se presenta en este articulo cómo propuesta de solución es el de regresión lineal el cual permite establecer una ecuación matemática que relaciona las variables de estudio.

V. DESARROLLO DE CONTENIDO

Para poder predecir los gastos médicos de un asegurado, se utilizó un modelo de regresión lineal múltiple de 7 variables independientes y una variable dependiente. El gasto médico de un asegurado se estableció como la variable dependiente la cual estaría en función de la edad del asegurado (x_0), de su género (x_1), de su índice de masa corporal (x_2), de su número de hijos (x_3), si es fumador (x_4) y de la región donde vive (x_5). La ecuación número 1 nos muestra el planteamiento ideal del modelo de regresión lineal para predecir los gastos médicos de un asegurado.

$$y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 + \beta_6 x_5 \quad (1)$$

El modelo de regresión lineal propuesto es construido a partir de un algoritmo programado en lenguaje Python, a continuación se explica el proceso que sigue el algoritmo para la construcción de dicho modelo. El algoritmo consta de tres etapas principales:

- **Etapa de Preprocesamiento de Datos**

  El modelo de regresión lineal es un modelo que solo trabaja con variables numéricas es por este motivo que en esta etapa el algoritmo se encarga de verificar si las variables independientes son categóricas, en caso afirmativo lo que realiza el algoritmo es binarizar las variables independientes categóricas con el fin de poder representar de forma numérica dichas variables; este proceso provoca que haya un aumento en las variables independientes al momento de construir el modelo de regresión lineal. En esta etapa también se verifica si la variable dependiente es categórica, en caso afirmativo lo que realiza el algoritmo es calcular su probabilidad con el fin de poder representar de una forma numérica dicha variable, para el cálculo de la probabilidad se utiliza la ecuación número 2.

$$In(\frac{p(y_i)}{1-p(y_i)}) \tag{2}$$

- **Etapa de Procesamiento de Datos**

  En esta etapa el algoritmo se encarga de realizar la construcción del modelo de regresión lineal, para esto el primer paso que realiza el algoritmo es la construcción del polinomio "S", el cual se construye a partir de la ecuación número 3.

$$S = \sum_{i=1}^{n}(y_i - (\beta_0 + \sum_{j=1}^{n}\beta_j x_{i,j}))^2 \tag{3}$$

  Una vez construido el polinomio "S" el algoritmo se encarga de calcular las derivadas parciales de cada una de las variables independientes sobre el polinomio "S" con el fin de obtener un sistema de ecuaciones. Tras haber construido el sistema de ecuaciones el algoritmo se encarga de resolverlo aplicando la técnica por sustitución por matrices, esto con el fin de calcular los valores de los coeficientes del modelo de regresión lineal. Existe el modelo de regresión linear estándar (4) y el modelo de regresión logístico (5), si la variable dependiente es categórica el algoritmo utiliza el modelo de regresión logístico y en caso contrario el modelo estándar.

$$y_i = \beta_0 + \sum_{j=1}^{n}\beta_j x_{i,j} \tag{4}$$

$$p(y_i) = \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^{n}\beta_j x_{i,j}}} \tag{5}$$

- **Etapa de Resultados**

  En esta última etapa el algoritmo se encarga de evaluar el modelo de regresión lineal recién construido, para esto requiere de otra fuente de datos que solo contenga las variables independientes. Lo que realiza el algoritmo es pasar la nueva fuente de datos por la etapa de preprocesamiento para tener el mismo número de variables que tiene el modelo construido en la etapa anterior y una vez finalizada la etapa de preprocesamiento se procede por evaluar las variables independientes en el modelo con la finalidad de poder predecir el valor de la variable dependiente según las variables independientes de entrada.

VI. Resultados

A continuación se presentan los resultados obtenidos al ejecutar el algoritmo anteriormente descrito sobre la fuente de datos Insurance [1]. Debido a que la fuente de datos Insurance [1] cuenta con variables categóricas el planteamiento del modelo de regresión lineal aumento de 7 variables independientes a 11 variables independientes. La ecuación número 6 nos muestra ya el planteamiento real del modelo de regresión lineal para estimar el costo del seguro para una persona.

$$\begin{aligned} Y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4 \\ +\beta_6 x_5 + \beta_7 x_6 + \beta_8 x_7 + \beta_9 x_8 + \beta_{10} x_9 + \beta_{11} x_{10} \end{aligned} \tag{6}$$

Dónde:

- x_0 Representa la edad de la persona asegurada.
- x_1 Representa si es mujer la persona asegurada.
- x_2 Representa si es hombre la persona asegurada.
- x_3 Representa el índice de masa corporal de la persona asegurada.
- x_4 Representa el número de hijos de la persona asegurada.
- x_5 Representa si es fumador la persona asegurada.
- x_6 Representa si no es fumador la persona asegurada.
- x_7 Representa si la persona asegurada vive en la región de "southwest".
- x_8 Representa si la persona asegurada vive en la región de "southeast".
- x_9 Representa si la persona asegurada vive en la región de "northwest".
- x_{10} Representa si la persona asegurada vive en la región de "northeast".

A continuación se muestra el polinomio "S" generado por el algoritmo para este caso de estudio.

$$\begin{aligned}
S = {} & 314494116894.925 - 26151511.7666\beta_0 \\
& -1148066958.4452\beta_1 - 12471080.4387\beta_2 \\
& -13680431.3279\beta_3 - 834418962.1425\beta_4 \\
& -29754521.3646\beta_5 - 12587097.4118\beta_6 \\
& -13564414.3548\beta_7 - 5875959.6244\beta_8 \\
& -7861194.0634\beta_9 - 5631028.771\beta_{10} \\
& -6783329.3079\beta_{11} + 1000\beta_0^2 + 79230.0B_0\beta_1 \\
& +990.0\beta_0\beta_2 + 1010.0\beta_0\beta_3 + 61726.76\beta_0\beta_4 \\
& +2160.0\beta_0\beta_5 + 392.0\beta_0\beta_6 + 1608.0\beta_0\beta_7 \\
& +488.0\beta_0\beta_8 + 556.0\beta_0\beta_9 + 462.0\beta_0\beta_{10} \\
& +494.0\beta_0\beta_{11} + 1769481.0\beta_1^2 + 39914.0\beta_1\beta_2 \\
& +39316.0\beta_1\beta_3 + 2465190.33\beta_1\beta_4 + 86628.0\beta_1\beta_5 \\
& +15478.0\beta_1\beta_6 + 63752.0\beta_1\beta_7 + 19574.0\beta_1\beta_8 \\
& +21706.0\beta_1\beta_9 + 18264.0\beta_1\beta_{10} + 19686.0\beta_1\beta_{11} \\
& +495.0\beta_2^2 + 0.0\beta_2\beta_3 + 30496.45\beta_2\beta_4 \\
& +1044.0\beta_2\beta_5 + 162.0\beta_2\beta_6 + 828.0\beta_2\beta_7 \\
& +240.0\beta_2\beta_8 + 286.0\beta_2\beta_9 + 220.0\beta_2\beta_{10} \\
& +244.0\beta_2\beta_{11} + 505.0\beta_3^2 + 31230.31\beta_3\beta_4 \\
& +1116.0\beta_3\beta_5 + 230.0\beta_3\beta_6 + 780.0\beta_3\beta_7 \\
& +248.0\beta_3\beta_8 + 270.0\beta_3\beta_9 + 242.0\beta_3\beta_{10} \\
& +250.0\beta_3\beta_{11} + 989083.1899\beta_4^2 + 67038.02\beta_4\beta_5 \\
& +12035.8\beta_4\beta_6 + 49690.96\beta_4\beta_7 + 15025.4\beta_4\beta_8 \\
& +18555.9\beta_4\beta_9 + 13538.45\beta_4\beta_{10} + 14607.01\beta_4\beta_{11} \\
& +2602.0\beta_5^2 + 420.0\beta_5\beta_6 + 1740.0\beta_5\beta_7 + 552.0\beta_5\beta_8 \\
& +570.0\beta_5\beta_9 + 486.0\beta_5\beta_{10} + 552.0\beta_5\beta_{11} \\
& +196.0\beta_6^2 + 0.0\beta_6\beta_7 + 82.0\beta_6\beta_8 + 130.0\beta_6\beta_9 \\
& +78.0\beta_6\beta_{10} + 102.0\beta_6\beta_{11} + 804.0\beta_7^2 + 406.0\beta_7\beta_8 \\
& +426.0\beta_7\beta_9 + 384.0\beta_7\beta_{10} + 392.0\beta_7\beta_{11} \\
& +244.0\beta_8^2 + 0.0\beta_8\beta_9 + 0.0\beta_8\beta_{10} + 0.0\beta_8\beta_{11} \\
& +278.0\beta_9^2 + 0.0\beta_9\beta_{10} + 0.0\beta_9\beta_{11} + 231.0\beta_{10}^2 \\
& +0.0\beta_{10}\beta_{11} + 247.0\beta_{11}^2
\end{aligned} \tag{7}$$

Una vez generado el polinomio "S" (7) el algoritmo prosiguió por realizar la derivadas parciales. En las siguientes ecuaciones se plasman los resultados de las derivadas parciales de cada variable independiente.

$$\begin{aligned}
\frac{\delta S}{\beta_0} = {} & -26151511.7666 + 2000\beta_0 + 79230.0\beta_1 \\
& +990.0\beta_2 + 1010.0\beta_3 + 61726.76\beta_4 \\
& +2160.0\beta_5 + 392.0\beta_6 + 1608.0\beta_7 \\
& +488.0\beta_8 + 556.0\beta_9 + 462.0\beta_{10} + 494.0\beta_{11}
\end{aligned} \tag{8}$$

$$\begin{aligned}
\frac{\delta S}{\beta_1} = {} & -1148066958.4452 + 79230.0\beta_0 \\
& +3538962.0\beta_1 + 39914.0\beta_2 + 39316.0\beta_3 \\
& +2465190.33\beta_4 + 86628.0\beta_5 + 15478.0\beta_6 \\
& +63752.0\beta_7 + 19574.0\beta_8 + 21706.0\beta_9 \\
& +18264.0\beta_{10} + 19686.0\beta_{11}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\frac{\delta S}{\beta_2} = {} & -12471080.4387 + 990.0\beta_0 + 39914.0\beta_1 \\
& +990.0\beta_2 + 0.0\beta_3 + 30496.45\beta_4 + 1044.0\beta_5 \\
& +162.0\beta_6 + 828.0\beta_7 + 240.0\beta_8 + 286.0\beta_9 \\
& +220.0\beta_{10} + 244.0\beta_{11}
\end{aligned} \tag{10}$$

$$\begin{aligned}
\frac{\delta S}{\beta_3} = {} & -13680431.3279 + 1010.0\beta_0 + 39316.0\beta_1 \\
& +0.0\beta_2 + 1010.0\beta_3 + 31230.31\beta_4 + 1116.0\beta_5 \\
& +230.0\beta_6 + 780.0\beta_7 + 248.0\beta_8 + 270.0\beta_9 \\
& +242.0\beta_{10} + 250.0\beta_{11}
\end{aligned} \tag{11}$$

$$\begin{aligned}
\frac{\delta S}{\beta_4} = {} & -834418962.1425 + 61726.76\beta_0 \\
& +2465190.33\beta_1 + 30496.45\beta_2 + 31230.31\beta_3 \\
& +1978166.3798\beta_4 + 67038.02\beta_5 + 12035.8\beta_6 \\
& +49690.96\beta_7 + 15025.4\beta_8 + 18555.9\beta_9 \\
& +13538.45\beta_{10} + 14607.01\beta_{11}
\end{aligned} \tag{12}$$

$$\begin{aligned}
\frac{\delta S}{\beta_5} = {} & -29754521.3646 + 2160.0\beta_0 + 86628.0\beta_1 \\
& +1044.0\beta_2 + 1116.0\beta_3 + 67038.02\beta_4 \\
& +5204.0\beta_5 + 420.0\beta_6 + 1740.0\beta_7 \\
& +552.0\beta_8\hat{}1 + 570.0\beta_9 + 486.0\beta_{10} + 552.0\beta_{11}
\end{aligned} \tag{13}$$

$$\begin{aligned}
\frac{\delta S}{\beta_6} = {} & -12587097.4118 + 392.0\beta_0 + 15478.0\beta_1 \\
& +162.0\beta_2 + 230.0\beta_3 + 12035.8\beta_4 + 420.0\beta_5 \\
& +392.0\beta_6 + 0.0\beta_7 + 82.0\beta_8 + 130.0\beta_9 \\
& +78.0\beta_{10} + 102.0\beta_{11}
\end{aligned} \tag{14}$$

$$\begin{aligned}
\frac{\delta S}{\beta_7} = {} & -13564414.3548 + 1608.0\beta_0 + 63752.0\beta_1 \\
& +828.0\beta_2 + 780.0\beta_3 + 49690.96\beta_4 + 1740.0\beta_5 \\
& +0.0\beta_6 + 1608.0\beta_7 + 406.0\beta_8 + 426.0\beta_9 \\
& +384.0\beta_{10} + 392.0\beta_{11}
\end{aligned} \tag{15}$$

$$\begin{aligned}
\frac{\delta S}{\beta_8} = {} & -5875959.6244 + 488.0\beta_0 + 19574.0\beta_1 \\
& +240.0\beta_2 + 248.0\beta_3 + 15025.4\beta_4 + 552.0\beta_5 \\
& +82.0\beta_6 + 406.0\beta_7 + 488.0\beta_8 + 0.0\beta_9 \\
& +0.0\beta_{10} + 0.0\beta_{11}
\end{aligned} \tag{16}$$

$$\frac{\delta S}{\beta_9} = -7861194.0634 + 556.0\beta_0 + 21706.0\beta_1 + 286.0\beta_2 + 270.0\beta_3 + 18555.9\beta_4 + 570.0\beta_5 + 130.0\beta_6 + 426.0\beta_7 + 0.0\beta_8 + 556.0\beta_9 + 0.0\beta_{10} + 0.0\beta_{11} \quad (17)$$

$$\frac{\delta S}{\beta_{10}} = -5631028.771 + 462.0\beta_0 + 18264.0\beta_1 + 220.0\beta_2 + 242.0\beta_3 + 13538.45\beta_4 + 486.0\beta_5 + 78.0\beta_6 + 384.0\beta_7 + 0.0\beta_8 + 0.0\beta_9 + 462.0\beta_{10} + 0.0\beta_{11} \quad (18)$$

$$\frac{\delta S}{\beta_{11}} = -6783329.3079 + 494.0\beta_0 + 19686.0\beta_1 + 244.0\beta_2 + 250.0\beta_3 + 14607.01\beta_4 + 552.0\beta_5 + 102.0\beta_6 + 392.0\beta_7 + 0.0\beta_8 + 0.0\beta_9 + 0.0\beta_{10} + 494.0\beta_{11} \quad (19)$$

Una vez teniendo todas las derivadas parciales el algoritmo crea un sistema de ecuaciones con los resultados de las derivadas parciales, que en este caso de estudio se creó un sistema de 12 ecuaciones. El algoritmo procede a resolver el sistema de ecuaciones con el fin de calcular los valores de los coeficientes del modelo de regresión lineal. La ecuación número 20 representa el modelo de regresión lineal obtenido para predecir los gastos médicos de un asegurado.

$$y = 4557.14706126 + 255.94419238x_0 + 1922.19810188x_1 - 1922.19810188x_2 + 409.17076529x_3 + 444.37506114x_4 + 24077.7869702x_5 - 24077.7869702x_6 + 241.40225208x_7 - 2256.00714796x_8 + 1396.1515506x_9 + 709.8584987x_{10} \quad (20)$$

Dónde:

- x_0 Representa la edad de la persona asegurada.
- x_1 Representa si es mujer la persona asegurada.
- x_2 Representa si es hombre la persona asegurada.
- x_3 Representa el índice de masa corporal de la persona asegurada.
- x_4 Representa el número de hijos de la persona asegurada.
- x_5 Representa si es fumador la persona asegurada.
- x_6 Representa si no es fumador la persona asegurada.
- x_7 Representa si la persona asegurada vive en la región de "southwest".
- x_8 Representa si la persona asegurada vive en la región de "southeast".
- x_9 Representa si la persona asegurada vive en la región de "northwest".
- x_{10} Representa si la persona asegurada vive en la región de "northeast".

VII. CONCLUSIÓN

De acuerdo a los resultados obtenidos se puede concluir que las variables más significativas para predecir los gastos médicos de un asegurado según el modelo de regresión linear obtenido son: la edad, el genero, el índice de masa corporal, el número de hijos, si es fumador y si el asegurado vive en las regiones de southeast y northwest.

Finalmente podemos concluir que los modelos de regresión lineal son fundamentales para determinar las relaciones de dependencia lineal entre variables, dónde la variable dependiente se encuentra en función de las variables independientes. El objetivo es determina una ecuación de regresión linear que nos permita realizar pronósticos o proyecciones de los valores de la variable dependiente ante posibles cambios de las variables independientes.

REFERENCES

[1] Zach Stednick y Brett Lantz. (2018). Medical Cost Personal Datasets, Version 1. Retrieved 19 Marzo, 2020. From https://www.kaggle.com/mirichoi0218/insurance.

[2] Douglas C Montgomery; Elizabeth A Peck; G Geoffrey Vining. (2012). Introduction to linear regressxion analysis. Hoboken, NJ: Wiley