Trabajo final

Lucio Enrique Cornejo Ramírez

2025-06-16

Table of contents

1 Introducción

1.1 Marco del problema

Entre los costos que más desastibilizan económicamente a las personas, se encuentra el pago por procedimientos médicos. Estos precios pueden variar en gran medida dependiendo de características del paciente, como detallaremos más adelante.

En ese sentido, resulta de gran valor predecir adecuadamente el costo que un seguro médico cubrirá, respecto a un procedimiento médico. Para un paciente, aquella predicción puede servir para que planifique qué tanto sería desestabilizado económicamente debido a algún tipo de procedimiento particular. Por otro lado, también para las aseguradoras resulta útil aquellas predicciones, ya que pueden anticipar qué tanto dinero estarían perdiendo por el monto a cubrir de la operación; además, con ese conocimiento pueden monitorear mejor qué pedidos de cobertura resultan anómalos, potencialmente fraudulentos.

1.2 Plan de modelamiento

Anteriormente, hemos planteado como variable por predecir a la cantidad monetaria S que la aseguradora de un paciente cubrirá debido a un procedimiento médico. Note que, fijando el precio P de un procedimiento médico, tal predicción equivale a predecir la cantidad $P-S$ que llegaría a pagar el paciente, habiendo descontado lo que cubre su aseguradora.

En ese sentido, el **carácter por modelar** es aquella variable aleatoria $P-S$, monto que paga un paciente por un tratamiento médico, tras aplicarse el descuento de su aseguradora.

Para el modelamiento, se considerará los gastos del hospital debido al procedimiento médico, además de las siguientes características del paciente:

- Sexo.
- Si fuma o no.
- Región de la que provee.
- Edad.
- Índice de masa corporal.
- Cantidad de hijos e hijas.
- Costo médico que pagaría en caso no se aplicase seguro médico.
- Número de procedimientos pasados.
- Número de pasos que realizó en cierto día.
- Número de veces que ha sido hospitalizado.
- Salario anual.

1.3 Datos/Observaciones

Las observaciones que consideraremos para este proyecto fueron descargadas de este sitio web.

Estos datos son de distintos pacientes que recibieron algún tipo de tratamiento médico, de los cuales se tienen variables recopiladas, como edad, sexo, si fuma o no, etc. Así, descartamos que los datos consistan de una serie de tiempo.

No obstante, aquella página web no provee información más específica sobre el origen de los datos. Por ejemplo, si han sido recopilados en un único hospital, o en diversos hospitales, pero de qué país, etc.

Aún así, en esta investigación, no solo se considera la predicción de la variable mencionada, sino también cómo es que influyen las variables que emplearemos como regresores, en la predicción final. Por ejemplo, si su relación es directa o inversamente proporcional.

A continuación justificamos el posible uso de los caracteres presentes en los datos, como covariables:

- **Sexo**: Debido al riesgo y costos distintos entre hombres y mujeres, para ciertos tipos de operaciones; por ejemplo, parto.
- **Si fuma o no**: Pues fumar aumenta la probabilidad de desarrollar complicaciones médicas
- **Región de la que provee**: Ya que el costo de un procedimiento médico puede variar mucho por región, así que también varía cuánto cubriría una aseguradora.
- **Edad**: Puesto que pacientes mayores suelen requerir más cuidados.
- **Índice de masa corporal**: En base a que un IMC elevado está asociado a mayores riesgos durante cirugías.
- **Cantidad de hijos e hijas**: Esto puede influir en el tipo de cobertura familiar (de seguro) que tiene el paciente.
- **Costo médico que pagaría en caso no se aplicase seguro médico**: Importante incluirlo, pues incluso se espera que presente una fuerte correlación positiva con la variable por predecir.
- **Número de procedimientos pasados**: Puede resultar útil en base a que pacientes con muchos procedimientos suelen tener enfermedades crónicas, por lo que se esperaría una mayor cobertura.
- **Número de pasos que realizó en cierto día**: Esta variable tampoco se explica en la fuente, pero la podemos considerar como una medida de la condición física de una persona, qué tan activa es.
- **Número de veces que ha sido hospitalizado**: Pues más hospitalizaciones implican mayor riesgo en la operación, aumentando posiblemente así los costos que cubre la aseguradora.

- **Salario anual**: Como indicador de nivel socioeconómico, se espera que pacientes con ingresos altos cuenten con aseguradoras que cubren mayor parte el costo por intervención médica.

1.4 Itinerario metodológico de la modelización

A continuación, describrimos los pasos a seguir para la construcción de diferentes modelos de predicción:

1. hola
2. hola

2 Materiales y métodos

2.1 Descripción genérica de los datos

3 Estructura de los datos

A continuación, mostramos los datos descargados del sitio web mencionado en la sección previa.

3.1 Variables iniciales

- Variables **cualitativas**
  - `sex`: Sexo.
  - `smoker`: Si el paciente fuma o no.
  - `region`: Región de la que provee el paciente.
- Variables **cuantitativas**
  - `age`: Edad.
  - `bmi`: Índice de masa corporal.
  - `children`: Cantidad de hijos e hijas.
  - `Claim_Amount`: Costo médico que el paciente pagaría en caso no se aplicase seguro médico.
  - `past_consultations`: Número de procedimientos pasados del paciente.
  - `num_of_steps`: Número de pasos que realizó el paciente en cierto día.
  - `Hospital_expenditure`: Gastos del hospital debido al procedimiento médico.
  - `Number_of_past_hospitalizations`: Número de veces que el paciente ha sido hospitalizado.

- `Anual_Salary`: Salario anual.
- `charges`: **Pago final** que el paciente realizó por el procedimiento médico, tras haberse descontado el monto que cubre el seguro médico.

Carácter respuesta: `charges`

Debido a la limitación, para esta investigación, de máximo 10 variables, además de solo una o dos variables cualitativas, ignoraremos algunas variables para este trabajo.

3.2 Filtro de variables

Para el filtro de variables categóricas, descartaremos aquella para la cual las distribuciones de la variable respuesta, respecto a los valores de aquella variable categórica sean relativamente similares.

3.2.1 Variable cualitativa `region`

Inspeccionamos la distribución de la variable respuesta, respecto a los valores de la variable categórica `region`.

En base a que aquellas funciones densidad no presentan una difencia resaltante, descartaremos la variable `region`. De esa manera, las variables cualitativas que emplearemos para esta investigación son solo `sex` y `smoker`.

Por otro lado, en el caso de las variables cuantitativas, primero inspeccionamos la correlación entre aquellas. Esto para descartar alguna de las variables que presente (de ser el caso) alta correlación lineal con otra variable cuantitativa.

A continuación, presentamos un heatmap de aquella matriz de correlaciones:

Observamos una alta correlación entre `Anual_Salary` y `Hospital_expenditure`, con un valor de 0.9692177. Asimismo, como la variable de salario anual es más sencilla de recopilar (por ejemplo, en una encuesta) que la de gasto de hospital, descartamos la variable cuantitativa `Hospital_expenditure`.

3.2.2 Variable cuantitativa `num_of_steps`

Inicialmente se consideró descartar la variable referente al número de pasos que realizó el paciente en cierto día. Esto pues, a primera vista, no se esperaría que tal información resulte relevante para el costo final por el procedimiento médico.

Graficamos tal posible regreso contra la variable respuesta:

En base a que la relación parece asemejarse a una exponencial, graficamos la variable `num_of_steps` contra el logaritmo de la variable respuesta:

En base a que aquella relación parece ser *aproximadamente* lineal, optamos por no descartar la variable cuantitativa `num_of_steps`.

3.2.3 Variable cuantitativa `num_of_steps`

3.2.4 Variable cuantitativa `age`

3.2.5 Variable cuantitativa `bmi`

3.2.6 Variable cuantitativa `children`

3.2.7 Variable cuantitativa `past_consultations`

Descartamos la variable cuantitativa `children`, pues, en base a este simple análisis inicial, no parece indicar algún de tipo de relación lineal con la variable por predecir. Es más, su gráfico de dispersión parece sugerir que consideremos a la variable `children` como cualitativa.

3.3 Variables finales

- Variables cualitativas:
  - `sex`
  - `smoker`
- Variables cuantitativas:
  - `age`
  - `bmi`
  - `Claim_Amount`
  - `past_consultations`
  - `num_of_steps`
  - `Number_of_past_hospitalizations`
  - `Anual_Salary`
  - `charges` (**variable respuesta**)

4 Base de datos

La base de datos consiste de 1338 observaciones. Si eliminamos filas que posean algún dato vacío, se tienen 1287 observaciones.

Para limitarnos a 500 filas, realizaremos un muestreo: