2024

OPTIMIZACIÓN DE TARIFAS DE SEGUROS DE SALUD MEDIANTE MODELOS DE MACHINE LEARNING

ENTREGA FINAL

CRISTHIAN ALEJO LEON
JAVIER BURGOS
SUSANA BARRIENTOS

Contenido

Tabla de figuras

INTRODUCCION

El proyecto que se presenta busca implementar un modelo de machine learning con el fin de prever no solo la frecuencia de uso de los servicios de salud, sino también los costos asociados a estos servicios. Esta capacidad predictiva se utilizará para establecer tarifas de seguros de salud más precisas y ajustadas a las necesidades reales de los asegurados. El enfoque en la predicción de costos y utilizaciones se traduce en una mejora significativa en la asignación de recursos dentro del sistema de salud, garantizando una distribución más eficiente de los recursos disponibles y, en última instancia, elevando el estándar de atención médica ofrecida a los beneficiarios del seguro. Este modelo no solo tiene el potencial de optimizar los procesos internos de las compañías aseguradoras, sino que también puede contribuir a una atención médica más equitativa y accesible para la población en general.

1. DISEÑO DE SOLUCIÓN PROPUESTO

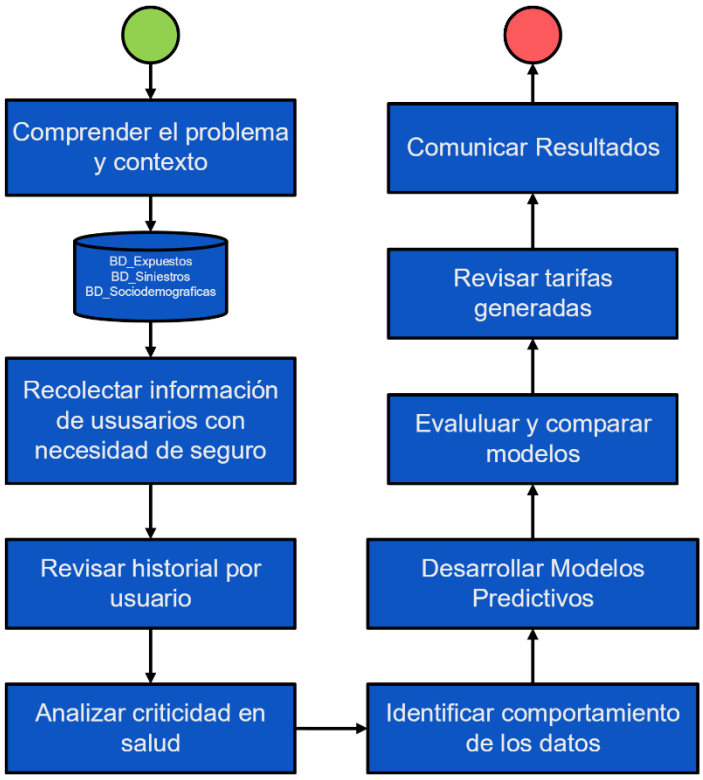

Figura 1 Diseño de solución propuesta

El diseño de la solución propuesta es un enfoque basado en datos para la fijación de precios de seguros que utiliza modelos predictivos para calcular primas para usuarios individuales. El proceso comienza con la **recopilación de información** de los usuarios que necesitan un seguro. Esta información puede incluir datos sociodemográficos, historial médico, historial de siniestros y otras variables relevantes para la evaluación del riesgo. Luego, la información se **analiza** para identificar

patrones y tendencias utilizando técnicas como el análisis estadístico, el aprendizaje automático y la minería de datos. Este análisis ayuda a identificar factores que afectan el riesgo de siniestros y a desarrollar modelos predictivos para calcular primas.

2. EXPLORACIÓN INICIAL

Durante el análisis exploratorio de las bases de datos de una aseguradora, se examinaron tres conjuntos de datos principales: **expuestos**, **sociodemográficos** y **siniestros**. La base de **expuestos**, con 300,900 registros y 5 variables, no presenta duplicados, aunque muestra 148,937 registros sin fecha de cancelación, y revela que los usuarios tienen en promedio 1.13 compras de seguros. En la base **sociodemográficos**, que incluye 267,312 registros y 9 variables, no se encontraron duplicados, pero se observó un pequeño porcentaje de registros con información de género faltante.

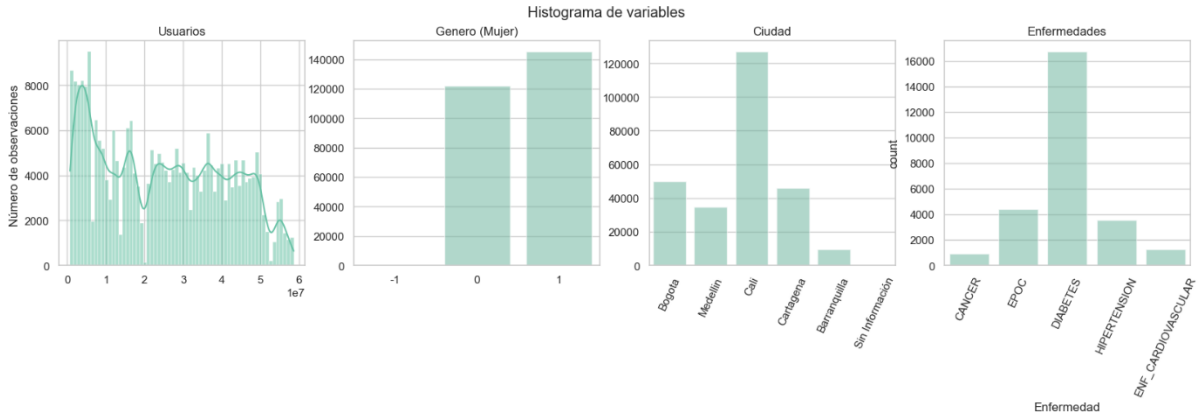

Figura 2 Histograma general de variables

Los usuarios tienen edades entre 1 y 110 años y provienen principalmente de 6 ciudades, siendo Cali la más representada.

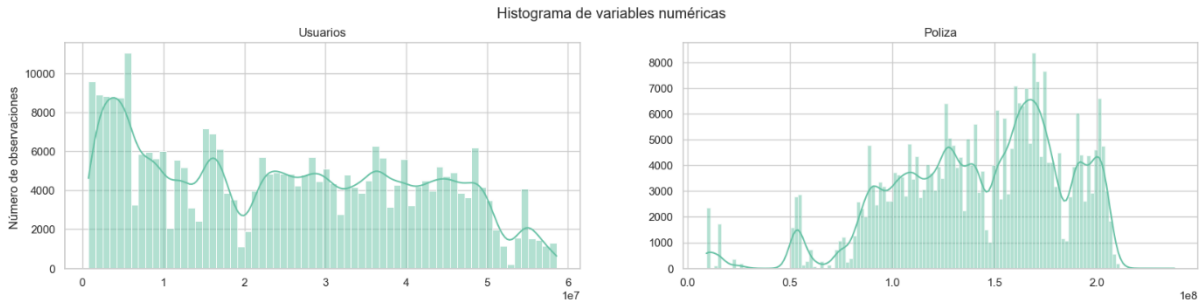

Figura 3 Histograma de variables numéricas

La base de **siniestros**, con 3,308,480 registros y 7 variables, tampoco contiene duplicados ni datos faltantes. En ella, se identificaron 41 tipos de reclamaciones y

5,830 códigos de diagnóstico, con la mayoría de los diagnósticos pendientes de especificación.

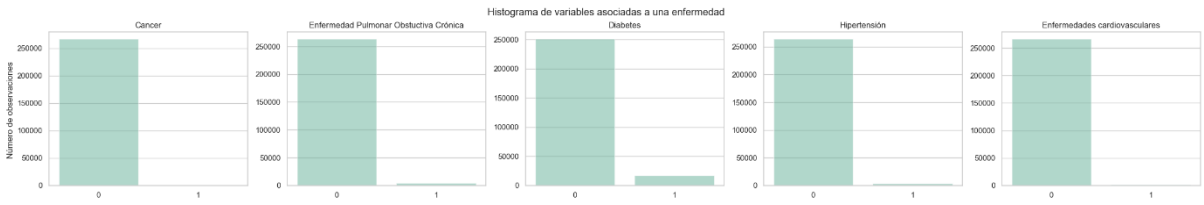

Figura 4 Grafica de variables asociadas a enfermedades

La integración de estos datos permitió una visión global de los asegurados y sus reclamaciones, destacando patrones significativos en las fechas de inicio y fin de las pólizas, así como en la prevalencia de enfermedades crónicas entre los usuarios.

3. PREPROCESAMIENTO

Se inició con una exploración básica para entender la estructura de las bases de datos **expuestos**, **sociodemográficas**, y **siniestros**, utilizando métodos como **.info()** y **.shape** para revisar la cantidad de entradas y el tipo de datos en cada columna. Esta etapa fue crucial para identificar columnas con datos faltantes o nulos e identificarlas con un nivel informativo bajo (no hay mucha variabilidad de datos).

- **BD_Expuestos**: Información detallada sobre las pólizas de los asegurados.
- **BD_Sociodemograficas**: Datos sociodemográficos de los asegurados.
- **DB_Siniestros**: Información sobre siniestros asociados a los asegurados.

3.1. Integración de Bases de Datos

Para un análisis cohesivo, fue necesario integrar los tres conjuntos de datos en uno solo. Esto se logró a través de las siguientes etapas:

- **Eliminación de duplicados**: Se eliminaron los duplicados en el conjunto de datos sociodemográficos, conservando solo la primera aparición de cada asegurado.
- **Fusión de datos**: Los tres conjuntos de datos se combinaron mediante uniones internas (inner join) en la columna común Asegurado_Id. Esto permitió crear un único DataFrame que contiene toda la información relevante sobre cada asegurado.

3.2. Transformaciones

Con los datos integrados, se realizaron varias transformaciones para limpiar y preparar los datos para su análisis:

- **Imputación y Eliminación de Información**

Eliminación de columnas irrelevantes: Se eliminaron las columnas Poliza_Asegurado_Id y Mes_Pago, ya que no aportaban información significativa al análisis.

Manejo de valores faltantes y categorías específicas:

Se eliminaron registros donde la Ciudad se reportaba como "Sin Información" o era nula.

Se excluyeron registros con el valor -1 en las variables Mujer y Diagnostico_Codigo, que indicaban datos faltantes o inválidos.

Creación y transformación de variables:

Se creó una nueva variable estado_poliza para clasificar el estado de la póliza como "Poliza activa" o "Poliza inactiva" basado en la columna FECHA_CANCELACION.

La variable FECHA_CANCELACION se imputó utilizando FECHA_FIN para los casos nulos, y luego se segmentó estado_poliza en las dos categorías mencionadas.

- **Reasignación de Tipo y Renombrado de Variables**

Tipos de datos: Las columnas de fechas (FECHA_INICIO, FECHA_CANCELACION, FECHA_FIN y FechaNacimiento) se convirtieron al formato datetime para facilitar el análisis temporal.

Renombrado de columnas:

Se cambiaron nombres de columnas para hacerlos más descriptivos y fáciles de manejar (Mujer a genero, FechaNacimiento a fecha_nacimiento, etc.).

Se adoptó la nomenclatura en Snake Case para los nombres de columnas, promoviendo la consistencia y claridad.

3.3. Segmentación

Segmentación por edad:

Se calculó la edad de los asegurados a partir de la diferencia entre la fecha de referencia (31 de diciembre de 2019) y la fecha_nacimiento.

Los asegurados se categorizaron en segmentos de edad específicos como "Primera infancia", "Infancia", "Adolescencia", etc., según su edad calculada.

Segmentación por diagnóstico:

Se creó una columna letra_CIE para extraer las primeras letras del código de diagnóstico (diagnostico_codigo).

Los diagnósticos se agruparon en 21 mega-categorías, simplificando así de 5821 diagnósticos individuales a grupos más manejables.

Se asignó la categoría "Diagnostico pendiente" a registros con diagnósticos no especificados.

3.4. Exportación de Datos

Finalmente, los datos procesados se guardaron en un nuevo archivo de texto para su uso futuro:

- **Exportación a archivo**: El DataFrame resultante se exportó a un archivo de texto (df.txt) en la carpeta data, asegurando un formato adecuado (sep='\t').

Este preprocesamiento ha dejado los datos listos para una exploración más profunda y análisis, proporcionando una base limpia y estructurada para el desarrollo de modelos y estudios posteriores.

4. SELECCIÓN DE MODELOS

4.1. Carga y Preprocesamiento de los Datos

El análisis comenzó con la carga del conjunto de datos preprocesado que contiene información de los asegurados y sus reclamaciones. Este dataset inicial contenía variables como el género, ciudad, diagnósticos, tipo de reclamación, y otras características médicas y de la póliza. Para optimizar el uso de los datos:

- **Cálculo del Valor Pagado Promedio**: Se calculó el valor pagado promedio por evento (valor_pagado_promedio), lo cual permitió normalizar los costos asociados a cada reclamación.
- **Eliminación de Variables Irrelevantes**: Se eliminaron las variables valor_pagado, eventos y asegurado_id que no eran necesarias para la predicción.
- **Conversión de Tipos de Datos**: Se aseguraron los tipos de datos apropiados para cada columna. Las variables binarias y numéricas se convirtieron a int64, mientras que las categóricas (ciudad, segmento_edad, diagnostico, reclamacion, y estado_poliza) se convirtieron a categorías.

4.2. Selección de Características y Variable Objetivo

Para construir el modelo, se definieron las variables predictoras X excluyendo valor_pagado_promedio, reclamacion, y diagnostico de la selección. La variable objetivo y fue el valor_pagado_promedio.

4.3. Preprocesamiento de las Características

Se identificaron las variables numéricas y categóricas:

- **Variables Numéricas**: Se incluyeron características como genero, cancer, epoc, diabetes, hipertension, enf_cardiovascular, y tiempo_poliza.
- **Variables Categóricas**: Se procesaron las variables categóricas ciudad, estado_poliza, segmento_edad, diagnostico y reclamacion mediante One-Hot Encoding.

Para asegurar que las características estuvieran en una escala adecuada, se utilizaron las siguientes transformaciones en el preprocesamiento:

- **Estandarización**: Se aplicó un StandardScaler a las características numéricas para normalizar su escala.
- **Codificación One-Hot**: Se aplicó a las variables categóricas para convertirlas en una representación numérica apropiada.

4.4. Transformación de la Variable Objetivo

Para facilitar el entrenamiento del modelo, se normalizó la variable objetivo valor_pagado_promedio usando MinMaxScaler.

4.5. Evaluación de Modelos de Regresión

Se evaluaron tres modelos de regresión para predecir el valor_pagado_promedio:

- **Decision Tree Regressor**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**

4.6. Validación y Entrenamiento de Modelos

Para cada modelo, se realizaron los siguientes pasos:

- **Pipeline de Preprocesamiento y Modelado**: Se creó un pipeline que incluía el preprocesador y el modelo de regresión.
- **Validación Cruzada**: Se utilizó validación cruzada con 5 pliegues para estimar la precisión del modelo. La métrica de error utilizada fue el Error Cuadrático Medio (MSE), transformado a la raíz cuadrada (RMSE) para facilitar la interpretación.
- **Entrenamiento del Modelo**: Cada modelo se entrenó usando el conjunto de entrenamiento.
- **Predicción y Evaluación en el Conjunto de Prueba**: Se evaluó el rendimiento de cada modelo en el conjunto de prueba calculando el RMSE y el coeficiente de determinación (R^2).

4.7. Resultados de los Modelos

Los resultados de la evaluación de los modelos fueron los siguientes:

- **Decision Tree Regressor**:
  - CV RMSE: 2,826,847.45 ± 410,651.56
  - Test RMSE: 2,815,930.20, R^2: 0.00
- **Random Forest Regressor**:
  - CV RMSE: 2,826,850.26 ± 410,651.83
  - Test RMSE: 2,816,097.64, R^2: 0.00
- **Gradient Boosting Regressor**:
  - CV RMSE: 2,826,850.90 ± 410,779.44
  - Test RMSE: 2,815,793.01, R^2: 0.00

Todos los modelos presentaron un rendimiento similar con un RMSE elevado y un coeficiente de determinación R^2 cercano a cero, indicando que los modelos no fueron capaces de explicar la variabilidad en el valor pagado promedio. Esto sugiere que el modelo necesita ajustes adicionales o que otros factores no incluidos en los datos actuales podrían estar influyendo significativamente en la predicción del valor pagado.

ENTRENAMIENTO Y EVALUACIÓN DEL MODELO CON RANDOM FOREST

Se entrenó un modelo de Random Forest Regressor para evaluar su capacidad predictiva en el conjunto de datos transformado:

- **Entrenamiento del Modelo**: Se utilizó el RandomForestRegressor de Scikit-learn con el conjunto de datos de entrenamiento estandarizado. Este modelo es robusto y puede manejar conjuntos de datos con gran cantidad de características y complejidad.
- **Predicciones**: Después del entrenamiento, se realizaron predicciones tanto en el conjunto de entrenamiento como en el de prueba para evaluar el rendimiento del modelo.

Evaluación del Rendimiento del Modelo

Se evaluó el rendimiento del modelo utilizando las siguientes métricas:

- **Error Cuadrático Medio (MSE)**: Se calculó el MSE para el conjunto de entrenamiento y prueba. El MSE mide el promedio de los cuadrados de los errores, es decir, las diferencias cuadradas entre los valores predichos y los valores observados.
    - MSE del Conjunto de Entrenamiento: 4,997,687,220,048.67
    - MSE del Conjunto de Prueba: 7,685,886,528,883.58
- **Coeficiente de Determinación (R^2)**: El R^2 mide la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. Un valor más cercano a 1 indica un mejor ajuste del modelo a los datos.
    - R^2 del Conjunto de Entrenamiento: 0.39
    - R^2 del Conjunto de Prueba: 0.04

Resultados y Conclusiones

El modelo de Random Forest Regressor mostró los siguientes resultados:

- **Desempeño en el Conjunto de Entrenamiento**: El modelo obtuvo un MSE de 4,997,687,220,048.67 y un R^2 de 0.39, lo que indica un ajuste moderado en el conjunto de entrenamiento.
- **Desempeño en el Conjunto de Prueba**: Sin embargo, en el conjunto de prueba, el MSE fue considerablemente más alto, 7,685,886,528,883.58, y el R^2 disminuyó a 0.04. Este resultado sugiere que el modelo no generaliza bien a nuevos datos y podría estar sobreajustado a los datos de entrenamiento.

Los resultados indican que, a pesar de los esfuerzos de preprocesamiento y la robustez del modelo de Random Forest, el rendimiento en el conjunto de prueba es bajo. Esto sugiere que podría ser necesario:

- **Explorar Características Adicionales**: Considerar otras variables que puedan influir en el valor_pagado_promedio.
- **Optimización del Modelo**: Ajustar hiperparámetros del modelo o explorar técnicas de reducción de dimensionalidad para mejorar la generalización.
- **Evaluar Modelos Alternativos**: Considerar otros tipos de modelos o combinaciones de modelos para mejorar la capacidad predictiva.

Los resultados subrayan la complejidad del problema y la necesidad de una mayor refinación en el enfoque de modelado