Caso: Predicción de Reclamaciones de Seguros Médicos

Ejercicio de aplicación

ID Bootcamps - Junio 2024

1 Descripción del Problema

Una importante compañía de seguros médicos busca predecir sus reclamaciones futuras e identificar los factores que provocan un aumento de los costes sanitarios. Para ello, nos proporciona una base de datos con información sobre 1,338 asegurados con 7 atributos que describen su salud e información demográfica, tales como la edad, el sexo, el IMC y el hábito de fumar. La empresa busca estimar el monto total de reclamaciones (*charges*) que se facturarán a la compañía de seguros.

2 Descripción del Dataset

- **age**: Edad del beneficiario principal
- **sex**: Género del asegurado (mujer, hombre)
- **bmi**: Índice de masa corporal. Muestra los pesos que son relativamente altos o bajos en relación con la altura. Idealmente de 18.5 a 24.9
- **children**: Número de hijos cubiertos por el seguro de enfermedad / Número de personas a cargo
- **smoker**: Fumador (yes/no)
- **region**: Zona residencial del beneficiario (northeast, southeast, southwest, northwest)
- **charges**: Gastos médicos individuales facturados por el seguro de enfermedad

3 Guía para la Realización del Modelo

3.1 Explorar el Problema

- **Análisis Exploratorio de los Datos**

  1. **Entender el dataset**: Significado de las columnas, comprender sus características.
  2. **Analizar los tipos de datos**: ¿Hay variables categóricas? ¿Hay numéricas?
  3. **Analizar estadísticamente el dataset**: Separar el análisis de las variables numéricas y categóricas.
  4. **Determinar el porcentaje de nulos**
  5. **¿Hay outliers?**
  6. **Análisis Univariado**
     - Distribución de cada variable (¿está centrada? ¿se asemeja a una normal? ¿cuál es su simetría? ¿cuál su curtosis?)
     - ¿Qué nos dice la estadística descriptiva de cada una de las variables?
     - ¿Qué nos dice la variable *sex*, *smoker* y *region*? ¿Cuántas pólizas contiene cada una? ¿Cuál es el promedio total de los reclamos de cada tipo?
  7. **Análisis Bivariado**
     - Analizar la distribución de reclamos (*claim amount $*) por región, por tipo de fumador, por edad, por BMI, por sexo, por región+sexo, por región+sexo+fumador. Puedes utilizar también cualquier gráfico que consideres relevante. ¿Qué conclusiones obtenemos?
     - Analizar la relación de cada una de las variables vs la variable pronóstico. ¿Qué podemos inferir? ¿Qué relaciones observamos? ¿Qué variables tienen más relación con el pronóstico?
  8. **Análisis Multivariado**
     - Análisis bivariado (Una variable vs otra variable)
     - Analizar las correlaciones contra la variable pronóstico
     - Correlaciones entre todas las variables. ¿Qué variables tienen más correlación?
  9. **Preseleccionar cuáles son las variables más importantes del dataset**: Las más relacionadas con el pronóstico.

3.2 Limpieza de los Datos

- Eliminar filas duplicadas
- Completar los valores faltantes si los hay (nulos reemplazar por la media)

3.3 Ingeniería de Atributos (Feature Engineering)

- Crear atributos (columnas) nuevas. Por ejemplo, discretizar el BMI en diferentes secciones.
- Convertir las categóricas en discretas: Sexo (0,1), Smoker (0,1).
- Utilizar *OneHotEncoder* para transformar la región y la columna nueva del punto a (clase de BMI). Así te quedarían una columna binaria por tipo de región, y otras columnas binarias por clase de BMI.
- Age, BMI, Children. ¿Conviene escalarlas? Si las escalas, puedes usar *MinMaxScaler*. Puedes combinar algunas variables con otras con la operación de la multiplicación (o concatenación de strings). Por ejemplo el BMI discretizado con el smoker.

3.4 Selección de Modelos

- Entrenar muchos modelos rápidos usando parámetros estándar: *Linear Regression* y *K-Nearest Neighbors*, *Random Forest*, *Gradient Boosting*, *XGBoost*. Puedes hacer uno o varios modelos.
- Medir los resultados (puedes utilizar *RMSE*, y *R2* ajustado)
- Hacer una lista de los 2 o 3 modelos más "prometedores".

3.5 Afinar los Modelos

- Ajustar hiperparámetros (*Grid Search/CV*) de los modelos prometedores (opcional)
- Seleccionar el mejor modelo y medir su rendimiento (medir en el test)

3.6 Interpretación del Modelo

- ¿Cuáles son las características más importantes? Puedes utilizar si lo deseas el *model.feature_importances_*. Haz un gráfico de barras con ellas, ordenado de mayor a menor.
- Observar el error. ¿Por qué se puede dar? ¿Outliers? Corregir outliers si los hay (opcional)
- Realizar los ajustes necesarios y volver a entrenar el modelo
- Analizar los resultados.

3.7 Opcional

- Puedes repetir el análisis realizando un modelo para cada una de las regiones. Un modelo por región ofrecerá una predicción de costes más específica, ya que permite analizar por separado el comportamiento de cada uno de los asegurados para cada una de las cuatro regiones: suroeste, sureste, noroeste y noreste.

3.8 Conclusiones Generales

- ¿Cuáles son las conclusiones más importantes?
- ¿Qué le recomendamos a la empresa?