Proyecto Claims Severity Prediction

ENTREGA 2

INTEGRANTES:

YOHEL OSVALDO PEREZ GARCIA

TATIANA ELIZABETH SÁNCHEZ SANIN

DANIELA ANDREA PAVAS BEDOYA

MATERIA:

Introducción a la Inteligencia Artificial para las Ciencias e Ingenierías

Profesor:

RAUL RAMOS POLLAN

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

2023

Exploración del dataset

Iniciamos la exploración de los datos del dataset, donde hay un total de 132 columnas o predictoras distintas que excluyen la variable respuesta y una columna adicional correspondiente a la variable respuesta. Estas características contienen tipos de datos categóricos y numéricos. De 131 variables, 116 son categóricas y 14 numéricas. Hay además una columna adicional correspondiente a los id de cada registro.

```
1 train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 188318 entries, 0 to 188317
Columns: 132 entries, id to loss
dtypes: float64(15), int64(1), object(116)
memory usage: 189.7+ MB
```

Se observa que la base de datos de entrenamiento consta de 188,318 observaciones y 131 variables que incluyen 72 variables categóricas binarias, 43 variables no binarias, 14 variables continuas y la variable de resultado, "Loss"

```
1 test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125546 entries, 0 to 125545
Columns: 131 entries, id to cont14
dtypes: float64(14), int64(1), object(116)
memory usage: 125.5+ MB
```

La base de datos test está compuesta de manera similar a training, sin embargo, tiene 125546 observaciones y no tiene la columna "Loss" pues es la variable respuesta.

```
1 #Evaluacion de valores nulos
2 print(train.isnull().values.any())
3 print(test.isnull().values.any())
```

```
False
False
```

```
print("datos duplicados en train:", train.duplicated().sum())
print("datos duplicados en train:", test.duplicated().sum())
```

```
datos duplicados en train: 0
datos duplicados en train: 0
```

Se encontró también que ni la base de datos de training o test tienen datos nulos o duplicados, lo cual evita que sea necesario realizar un preprocesamiento o estandarización.

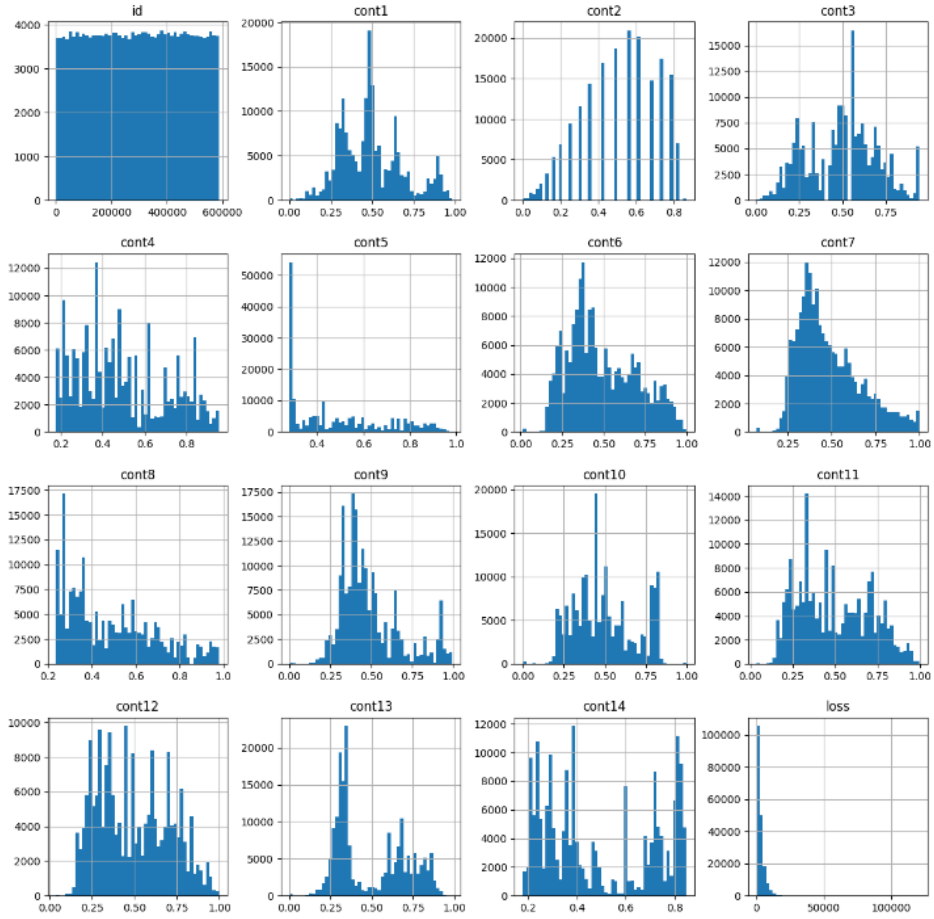

Fig. 1. Histograma

Se realiza un histograma para evaluar similitudes entre variables, como es el ajuste de cada una, si tiene valores atipicos y como es la distribución. Hay muchos picos en el continuo que muestran que las características no se distribuyen normalmente. Intentamos transformar estas características para que su distribución sea más gaussiana, pero es posible que no mejore el rendimiento del modelo.

Fig. 2. Matriz de correlación

Se creó una matriz de correlación para evaluar variables continuas que pudieran estar relacionadas linealmente. En este caso se puede ver cont11 y cont12 dan un patrón casi lineal (0.99). Por lo tanto, uno debe eliminarse

cont1 y cont9 también están altamente correlacionados (0.93), cualquiera de ellos podría eliminarse con seguridad

cont6 y cont10 también muestran una muy buena correlación (0.88)

Vemos una alta correlación en las características mencionadas anteriormente. Esto puede ser el resultado de una multicolinealidad basada en datos, ya que dos o más predictores están altamente correlacionados. Causa muchos problemas, por lo que debemos tener mucho cuidado al implementar modelos de regresión lineal en el conjunto de datos actual.

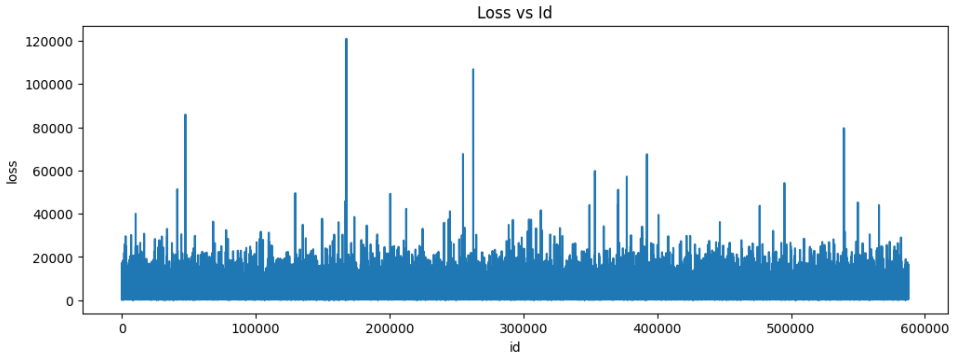

Fig. 3. Analisis variable respuesta

Se analizó la variable respuesta donde se presentan algunos picos con valores muy altos correspondientes a perdidas debido a accidentes graves.

Tal distribución de datos hace que esta función sea muy sesgada o asimétrica y puede dar como resultado un rendimiento subóptimo del regresor.

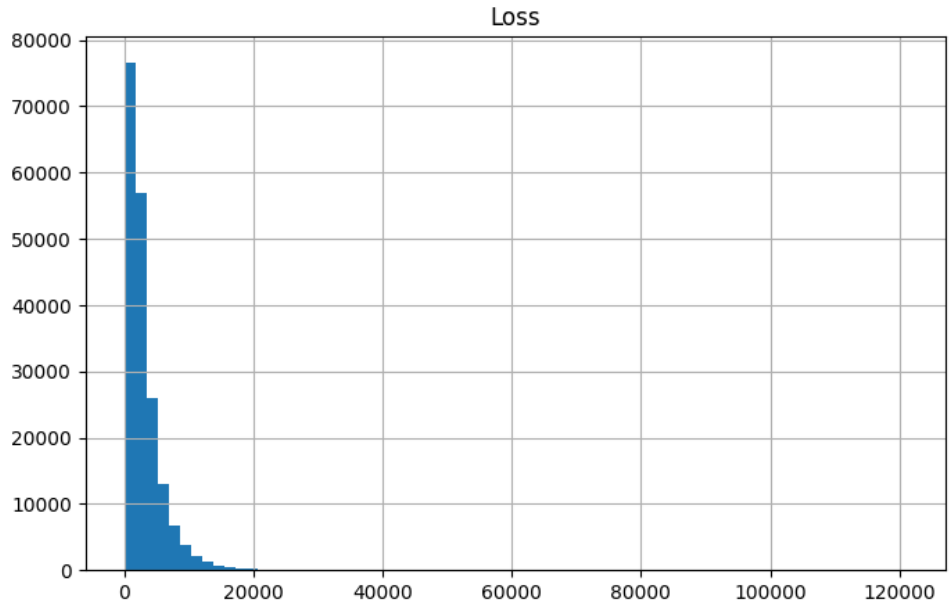

Fig. 4. Evaluación distribución de la variable respuesta

A partir de los resultados, se observa que la variable loss está sesgada y es asimétrica hacia la derecha.

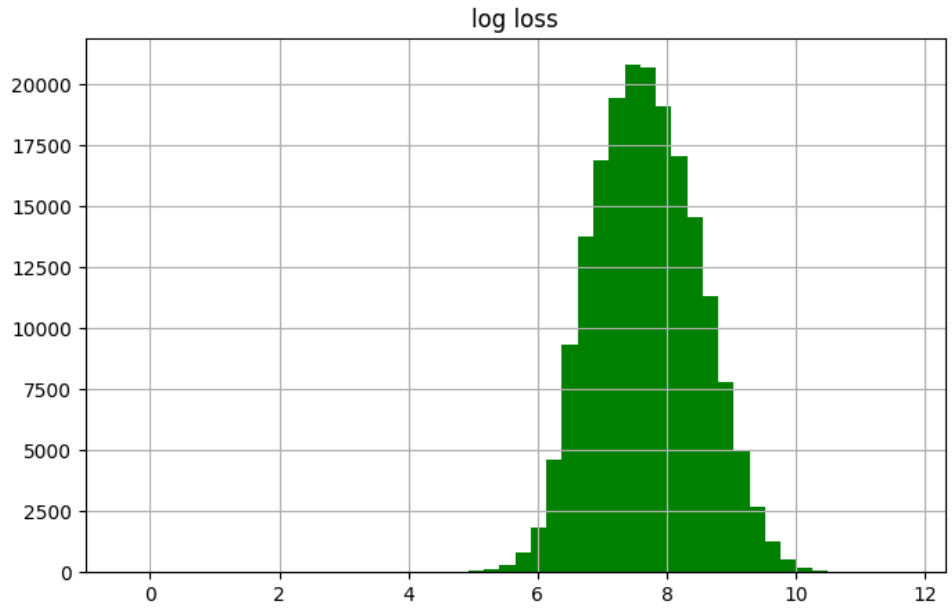

Fig. 4. Análisis de la distribución después de aplicar la transformación logarítmica

De la figura anterior, se puede confirmar que después de aplicar la transformación, ahora los datos parecen estar distribuidos normalmente. Esto permitirá que los análisis estadísticos sean confiables