Instituto Tecnológico y de Estudios Superiores de Monterrey

Escuela de Ingeniería y Ciencias

Ingeniería en Ciencia de Datos y Matemáticas

El precio de los autos

Inteligencia artificial avanzada para la ciencia de datos I

Morales Ramón Michelle Yareni A01552627

supervisado por:
Dra. Blanca Rosa Ruiz Hernández

Monterrey, Nuevo León. Fecha, 11 de septiembre de 2023

1. Resumen

En este proyecto, se llevó a cabo un análisis de una base de datos que contiene información sobre automóviles, con el objetivo de predecir sus precios. El proceso se desarrolló en varias etapas. Primero se hizo un análisis exploratorio y se acondicionó del dataset. Posteriormente se buscaron seis variables para usarse en un modelo de regresión. Estas fueron seleccionadas mediante análisis de correlación y en el caso de las variables categóricas se usó ANOVA. Para hacer la regresión se intentó normalizar las variables pero no fue posible. Aún así se implementó un modelo pero como era de esperar, los residuos no siguieron una distribución normal, lo que puede tener implicaciones importantes en la calidad de las predicciones.

2. Introducción

El análisis de precios de automóviles basado en sus características se ha convertido en un campo importante dentro del análisis de datos y la inteligencia artificial. La capacidad de predecir el precio de un automovil permite a los fabricantes y compradores tomar decisiones informadas. Además, esta información puede ser de utilidad en la evaluación de estrategias de precios y la comprensión de las preferencias del mercado. En este contexto, surge la pregunta clave: ¿Cómo podemos utilizar las características de un automóvil para predecir su precio de manera efectiva?

Actualmente, con el acceso a muchos tipos de datos y el avance de técnicas de Machine Learning, se han logrado avances significativos en la predicción de precios de automóviles. La resolución de este problema requiere la aplicación de técnicas de regresión, donde las características del automóvil actúan como variables predictoras y el precio del automóvil es la variable de salida. Pero para lograr un buen modelo predictivo hay todo un proceso que seguir que va desde la preparación de los datos y la elección de las variables hasta la implementación y los ajustes al modelo.

En este estudio, se abordará el desafío de predecir los precios de los automóviles utilizando un modelo de regresión múltiple y se explorará cómo las diferentes características del vehículo influyen en la predicción de su precio.

2.1. Exploración de la base de datos

Lo primero que se hizo fue conocer el dataset, el cual tiene un total de 21 variables, 14 numéricas y 7 categóricas. Hay 205 registros y no hay datos duplicados ni datos faltantes.

2.2. Variables cuantitativas

En un primer análisis, se utilizaron diagramas de caja (boxplots) para examinar los datos. Este proceso reveló que la mayoría de las variables presentan valores atípicos, aunque solo algunas exhiben valores extremos. Además, se observó que algunas variables tienen escalas de magnitudes muy diferentes entre sí. Figura 1

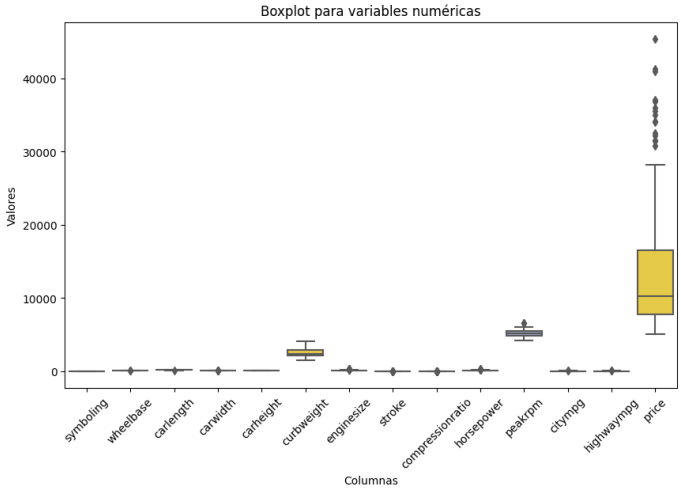

Figura 1: Boxplots.

También se realizaron histogramas, donde se pudo observar que todas las distribuciones de datos son asimétricas y hay algunas variables en particular que tienen un sesgo muy notorio como es el caso de horsepower, wheelbase, enginesize, compressionratio y, especialmente, el precio, que es la variable de interés para la predicción.

Durante el análisis de correlación, se identificaron variables altamente correlacionadas con el precio, destacándose las siguientes: carwidth, curbweight, horsepower, enginesize, citympg, high-

waympg y wheelbase. Figura 2

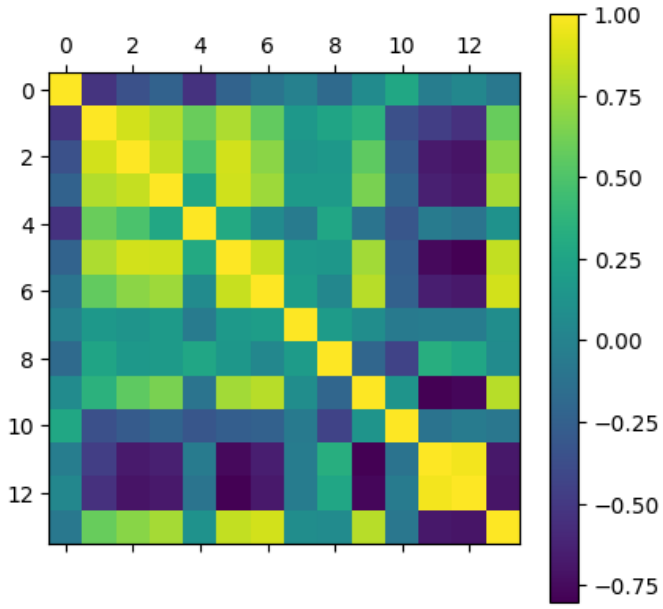

Figura 2: Mapa de calor.

2.3. Variables cualitativas

En el análisis de las variables cualitativas, se llevó a cabo una visualización gráfica para examinar la frecuencia de cada categoría dentro de estas variables. Lo primero que se observó fue que la variable 'carname' tiene más de 30 categorías por lo que no se tomó en cuenta en análisis posteriores.

También se hicieron diagramas de pastel para apreciar de manera mas clara la proporcion con la que se presenta una u otra categoria dentro de una variable y se hicieron los siguientes hallazgos:

- Fueltype: La mayoría de los autos utilizan gas y solo un pequeño porcentaje utiliza diesel.
- Carbody: Aproximadamente la mitad de los autos son del tipo sedan, y otro gran porcentaje corresponde a hatchback.
- Drivewheel: Más del 50 % de los autos tienen tracción delantera (fwd).

- Enginelocation: El 98.5 % de los autos tienen el motor ubicado en la parte delantera (front).
- Enginetype: Un gran porcentaje de los autos tienen un motor ohc.
- Cylindernumber: La mayoría de los autos son de 4 cilindros.

Se realizó un análisis adicional utilizando diagramas de caja (boxplot), esta vez considerando la variable 'price' en conjunto con las categorías de las variables cualitativas. Durante este análisis, se identificó que las categorías de la variable 'enginetype' fueron las únicas en las que se encontraron una significativa cantidad de valores atípicos y datos extremos.

2.4. Elección de variables

La selección de variables cuantitativas se basó en un análisis de la matriz de correlación. Como se mencionó previamente, algunas variables presentaban una alta correlación con el precio, pero también se observó que varias de ellas estaban fuertemente correlacionadas entre sí. Por esta razón se descartaron algunas de estas variables, con el objetivo de evitar la multicolinealidad ya que eso puede afectar la eficacia del modelo al hacer las predicciones.

Por otro lado, se realizaron pruebas de ANOVA para evaluar si las variables cualitativas tenían un efecto significativo sobre el precio y aunque todas las variables arrojaron p-values pequeños, se destacaron tres en particular que mostraron una mayor significancia en relación al precio. Figura 3

```
                Df    Sum Sq   Mean Sq F value  Pr(>F)
fueltype         1 1.454e+08 1.454e+08  10.704 0.00128 **
carbody          4 1.716e+09 4.290e+08  31.580 < 2e-16 ***
drivewheel       2 4.068e+09 2.034e+09 149.721 < 2e-16 ***
cylindernumber   6 4.081e+09 6.801e+08  50.069 < 2e-16 ***
enginetype       5 3.605e+08 7.210e+07   5.307 0.00014 ***
enginelocation   1 1.365e+08 1.365e+08  10.050 0.00178 **
Residuals      185 2.513e+09 1.358e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3: ANOVA.

Finalmente se escogieron seis variables predictoras para usar en la implementación de la regresión: 'carbody', 'drivewheel', 'cylindernumber', 'wheelbase', 'enginesize' y 'citympg'.

3. Preparación de los datos

Una vez elegidas las variables se hizo un nuevo dataframe que incluyera solo las columnas de interés. Se cambió la variable 'cylindernumber' a valores numéricos y debido a que ninguna de las variables seguía una distribución normal se intentó aplicar una transformación de boxcox.

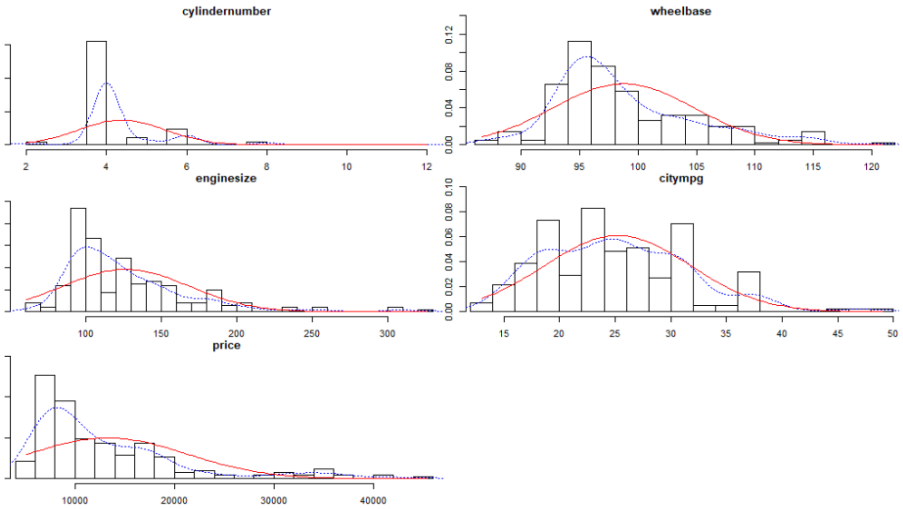

Figura 4: Distribuciones de las variables elegidas.

En un primer intento, al hacer la prueba de Anderson Darling se obtuvo que ninguna de las variables tenían normalidad, por lo que se quitaron los valores atípicos para repetir la transformación.

Al repetir la transformación sin los datos atípicos los histogramas mejoraron considerablemente pero aún así no pasaron la prueba de normalidad. Sin embargo se decidió usar los datos obtenidos con la transformación para implementar la regresión. Figuras 5 6 7

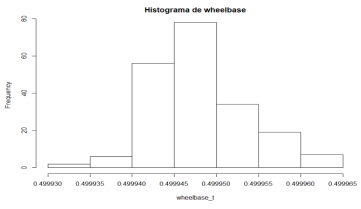

Figura 5: Wheelbase.

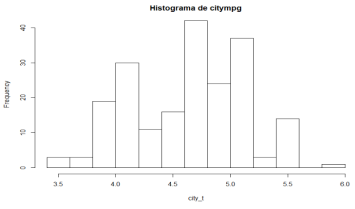

Figura 6: Citympg.

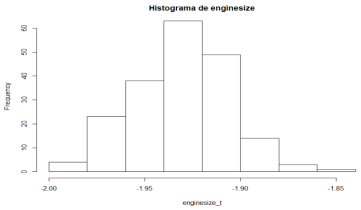

Figura 7: Enginesize.

3.1. Implementación de la Regresión

Con los datos preparados se procedió a implementar una regresión multivariada con las seis variables elegidas. En el summary del modelo se obtuvo un p-value significativo así que se hizo un step() para ver si el modelo mejoraba al quitar o poner variables. El resultado fue un modelo con solo cinco variables predictoras: 'carbody', 'cylindernumber', 'wheelbase', 'enginesize' y 'citympg'. Figura 8

```
Call:
lm(formula = M$price ~ M$wheelbase + M$carbody + M$citympg +
    M$drivewheel + M$cylindernumber)

Coefficients:
       (Intercept)         M$wheelbase    M$carbodyhardtop
         -1.040e+08           2.082e+08          -2.082e+03
M$carbodyhatchback      M$carbodysedan      M$carbodywagon
         -5.831e+03          -5.230e+03          -6.447e+03
          M$citympg     M$drivewheelfwd     M$drivewheelrwd
         -4.231e+03          -5.196e+01           3.075e+03
  M$cylindernumber
          2.702e+03
```

Figura 8: Modelo final.

Después se hicieron las pruebas de normalidad y homocedasticidad para validar el modelo. En la QQplot se observó que los residuos no se ajustaban del todo a la linea y el shapiro test arrojó un p-value muy pequeño, por lo que se concluye que los residuos no son normales. Por otro lado, en el gráfico de los residuos se observó que hay independencia pero tambien hay una ligera heterocedasticidad. Figuras 9 10

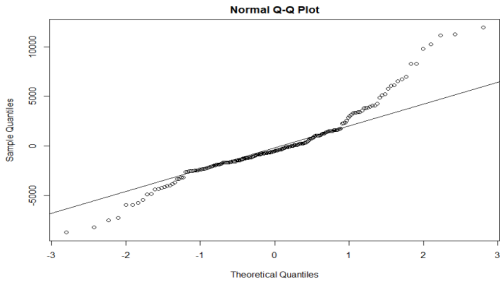

Figura 9: QQplot.

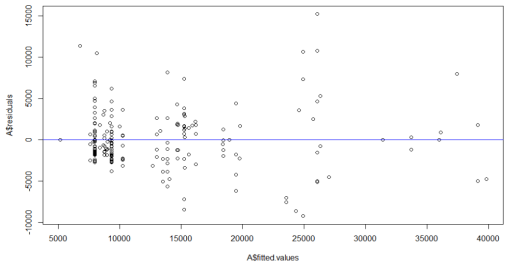

Figura 10: Residuals.

4. Conclusiones

Tras realizar una evaluación del modelo, podemos concluir que su capacidad para realizar predicciones no es eficiente debido a que no se cumplen los supuestos de normalidad y homocedasticidad en los residuos. No obstante, es importante destacar que se trabajó con pocos datos, ya que el dataset cuenta con tan solo 205 registros. Es posible que con una expansión del conjunto de datos se logren mejoras significativas en el desempeño del modelo.

Este hallazgo subraya la importancia de la calidad y cantidad de los datos disponibles en el proceso de construcción y validación de los modelos de regresión. A medida que se aumente la cantidad de datos y se aborde la adecuación de los supuestos, es posible que se obtenga un modelo más sólido y confiable en futuras iteraciones del análisis.

4.1. Anexos

Se trabajó tanto en Pyhton como en R.

Liga a la carpeta de Drive:`https://drive.google.com/drive/folders/1zh1KQ7woDmO-e-XydCl22XWuFQv24mjP?usp=drive_link`

Portafolio de Análisis: `https://colab.research.google.com/drive/1OCOZIJJz4O-K_7nGSe5AXRDjOZnB8uMU?usp=drive_link`

Portafolio de Implementación: `https://drive.google.com/file/d/1OPYrPuNAOIKrmZZO_ZKEmt6ItplUx1im/view?usp=drive_link`