Campus Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Reporte final de Estadistica

Juan Pablo Castañeda Serrano A01752030

Resumen

Una compañía de autos de origen chino está intentando posicionarse en el mercado de Estados Unidos, enfrentándose a fabricantes nacionales y de Europa. Para ello, contrató servicios de consultoría para discernir qué factores influencian de manera significativa el costo de los autos en EE. UU. y la precisión con la que estas variables pueden predecir las fluctuaciones de precio. En esencia, el objetivo es comprender qué elementos son cruciales para estimar el valor de los autos en el contexto estadounidense y cuán acertadamente lo logran.
Se llevaron a cabo estudios detallados de cada factor en la base de datos usando herramientas de análisis visual. Luego, se implementaron pruebas estadísticas que resultaron en la selección de 6 factores clave para desarrollar un modelo de regresión lineal multivariante. Adicionalmente, se examinó la normalidad de los residuos del modelo.

Introducción

Todas las gráficas y matrices se pueden concorde al código para seguir este reporte, de manera que aparecen las gráficas y matrices en el orden que se mencionan en este reporte
Una firma automovilística de origen chino está decidida a hacer su debut en el mercado estadounidense. Con miras a fortalecer su posición, planea instalar una unidad de producción en el país y fabricar vehículos in situ para desafiar a las marcas autóctonas y europeas. Reconociendo las disparidades entre los mercados chino y estadounidense, la empresa ha recurrido a especialistas; contrataron una consultora especializada en la industria automotriz con el objetivo de descifrar los factores determinantes en la fijación de precios de los vehículos en EE. UU. En el corazón de su inquisición están dos preguntas clave:

¿Cuáles son las variables que tienen un impacto significativo en la determinación del precio de un vehículo?
¿Qué grado de precisión tienen estas variables al reflejar el precio de un automóvil?

Para arrojar luz sobre estas interrogantes, la consultora llevó a cabo encuestas exhaustivas y, como resultado, compiló un vasto conjunto de datos de diversos vehículos disponibles en el mercado estadounidense. Esta base de datos, junto con un diccionario explicativo que detalla cada variable y el tipo de información que contiene (ya sea categórica o numérica continua), ha sido compartida con los interesados en este estudio

Análisis de los resultados

Análisis de la base de datos

La base de datos proporcionada para el análisis, proveniente de la consultora, consta de 205 registros y 21 columnas, excluyendo encabezados. Estas columnas representan variables específicas que se categorizan en numéricas y categóricas. Específicamente, 7 de estas columnas son de tipo categórico y las 13 restantes son numéricas, incluyendo tanto números enteros como valores continuos. Las variables categóricas incluyen: symboling, fueltype, carbody, drivewheel, enginelocation, enginetype y cylindernumber. Por otro lado, las numéricas comprenden: wheelbase, carlength, carwidth, carheight, curbweight, enginesize, stroke, compressionratio, horsepower, peakrpm, citympg, highwaympg y price. Se efectuó

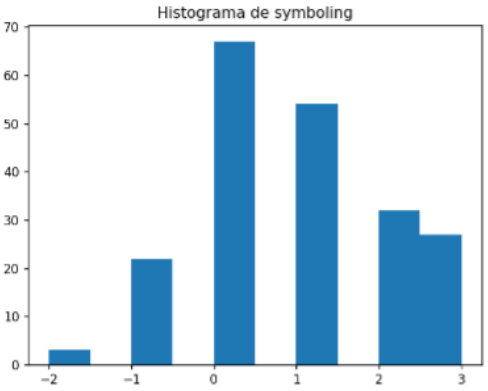
una revisión en busca de valores faltantes en la base de datos utilizando la función colSums(is.na(df)) y, gratamente, no se detectó ningún valor ausente.

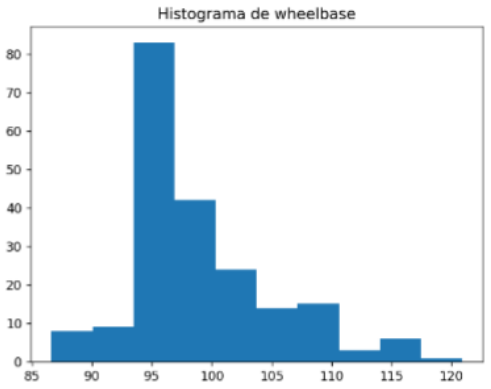
Además, en el análisis de las variables numéricas se llevaron a cabo histogramas para cotejar la distribución de cada una respecto a una distribución normal y se emplearon boxplots con el propósito de identificar posibles valores atípicos en el conjunto de datos.

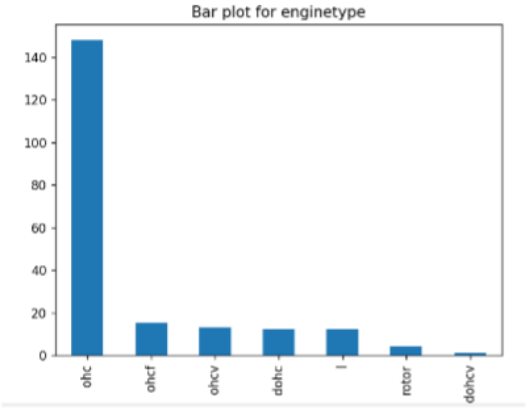
El código utiliza bibliotecas comunes en Python para el análisis y visualización de datos, como pandas, numpy, matplotlib, seaborn y scipy. También hay otras bibliotecas en Python que proporcionan funcionalidades específicas que podrías requerir, así que asegúrate de tener todas las bibliotecas instaladas o instalarlas si es necesario.

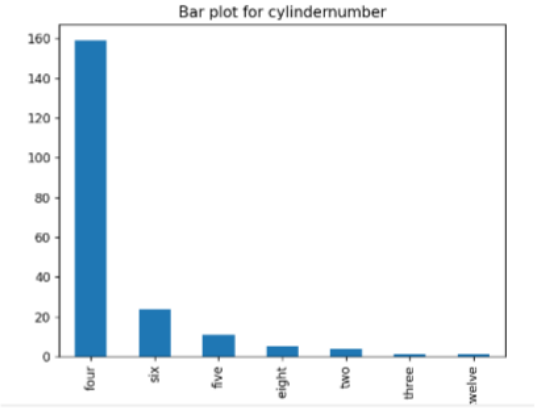
Selección de datos

Gráficas:









Así sucesivamente para cada uno de los datos

Dado estas gráficas se decidió, eliminar los datos de los bordes al eliminar los más alejados del primer y cuarto cuartil, y eliminar los datos nulos de esta manera:

Con esto se decidió utilizar los datos de

'carlength', 'carwidth', 'curbweight', 'enginesize' y 'horsepower'

Y con esto tendremos nuestros datos ya limpios

Después de llevar a cabo un detallado análisis de los histogramas de las variables numéricas, pudimos observar que todas ellas se ajustaban satisfactoriamente a lo que esperaríamos de

una distribución normal, y por lo tanto, pueden ser categorizadas como normales. Sin embargo, es importante señalar que no todas presentaban un comportamiento perfectamente simétrico. En efecto, detectamos sesgos de variada magnitud en algunas de estas variables. Un caso particularmente destacable es el de la variable "price" (como se puede observar en la figura 2). Esta variable reviste una importancia especial en nuestro estudio, dado que nuestro objetivo es descifrar cómo esta se ve influenciada o modulada por otras variables que, a su vez, podrían tener distribuciones diferentes o comportamientos atípicos.

Al analizar el conjunto de datos para identificar las variables más correlacionadas con "price", se empleó una matriz de correlación de coeficientes de Pearson de las variables numéricas presentes. Dado que se generó una matriz de dimensiones 13x13, se optó por crear un heatmap para facilitar la identificación visual de las correlaciones más fuertes. En este heatmap, se observa que aquellos recuadros con colores rojos y azules intensos representan las correlaciones más significativas.
Al observar el heatmap de correlación entre las variables numéricas del conjunto de datos, destacando que los cuadros con tonalidades más intensas de rojo y azul indican una mayor correlación.
A la hora de seleccionar variables predictoras, es vital considerar aquellas con una alta correlación con "price". Sin embargo, es crucial garantizar que estas variables no estén altamente correlacionadas entre sí para evitar problemas de colinealidad. Suponiendo que dos variables estén altamente correlacionadas con "price" y entre sí, la variabilidad que explican en un modelo multilineal para "price" podría ser similar a la que explicarían por separado, ya que ambas tendrían comportamientos similares ante cambios en "price". Así, es esencial seleccionar variables con alta correlación con "price", evitando al mismo tiempo la colinealidad.
De la matriz de correlación, se destacaron las cuatro variables con la correlación más significativa con "price": enginesize, curbweight, carwidth, y horsepower. Además, se eligieron cuatro variables categóricas como posibles candidatos para la regresión lineal. Esta elección se basó en la observación de boxplots por categoría, seleccionando aquellos que mostraban una diferencia notable en la media de "price" entre los grupos. Esta selección inicial cualitativa fue luego validada estadísticamente mediante pruebas de regresión lineal.

Análisis de regresión

Al examinar los modelos lineales generados para cada variable numérica en relación con la variable "price", se nota que el promedio de los residuos está próximo a 0 y se distribuyen de manera equilibrada. El valor F es notablemente elevado y el p-value en todos los modelos es inferior a 0.05. La varianza explicada por estos modelos varía entre el 2% para la variable "carheight" y el 81% para "curbweight". Se llevarán a cabo regresiones múltiples para identificar el modelo que mejor describa la variabilidad de "price".

Para determinar el modelo múltiple óptimo que explique la variabilidad de "price" en Python, se hará uso de la biblioteca statsmodels y su método OLS (Ordinary Least Squares). Con el propósito de hacer una selección automática de características, se puede emplear técnicas como el uso de criterios AIC o BIC para seleccionar el mejor subconjunto de predictores.

Después de ajustar el modelo utilizando la función OLS() de statsmodels, se concluyó que las variables "curbweight" y "horsepower" eran las más adecuadas como variables independientes para predecir "price". Al examinar este modelo, se observó un promedio de residuos casi nulo y distribuidos equilibradamente alrededor de ese valor. El valor F es alto, lo cual es lo que se busca, y el p-value es menor a 0.05, indicando que el modelo es significativo para describir la variabilidad de "price". Este modelo explica el 85.5% de la variabilidad, lo que representa un valor considerablemente alto para nuestro estudio.

```
Variable: wheelbase
count    205.000000
mean      98.756585
std        6.021776
min       86.600000
25%       94.500000
50%       97.000000
75%      102.400000
max      120.900000
Name: wheelbase, dtype: float64


Variable: carlength
count    205.000000
mean     174.049268
std       12.337289
min      141.100000
25%      166.300000
50%      173.200000
75%      183.100000
max      208.100000
Name: carlength, dtype: float64


Variable: carwidth
count    205.000000
mean      65.907805
std        2.145204
min       60.300000
25%       64.100000
50%       65.500000
75%       66.900000
max       72.300000
Name: carwidth, dtype: float64
```

```
Variable: curbweight
count     205.000000
mean     2555.565854
std       520.680204
min      1488.000000
25%      2145.000000
50%      2414.000000
75%      2935.000000
max      4066.000000
Name: curbweight, dtype: float64


Variable: enginesize
count    205.000000
mean     126.907317
std       41.642693
min       61.000000
25%       97.000000
50%      120.000000
75%      141.000000
max      326.000000
Name: enginesize, dtype: float64


Variable: stroke
count    205.000000
mean       3.255415
std        0.313597
min        2.070000
25%        3.110000
50%        3.290000
75%        3.410000
max        4.170000
Name: stroke, dtype: float64
```

```
Variable: price
count      205.000000
mean     13276.710571
std       7988.852332
min       5118.000000
25%       7788.000000
50%      10295.000000
75%      16503.000000
max      45400.000000
Name: price, dtype: float64
```

Conclusión

Tras el análisis estadístico llevado a cabo, hemos arribado a las siguientes conclusiones esenciales:
Variables Significativas: A través de un meticuloso proceso de selección, determinamos que "curbweight" y "horsepower" son las variables más influyentes al predecir el precio de los vehículos en el mercado estadounidense. Ambas variables juegan un rol crucial en la determinación de la variación del precio.

Relevancia de Symboling: La variable categórica "Symboling" ha demostrado tener importancia. La distinción notable entre sus categorías implica que el nivel de seguridad percibido de un vehículo, representado por esta variable, tiene un impacto en su precio en EE.UU.
Modelo Optimo: La combinación de "curbweight" y "horsepower" compone el modelo más adecuado, abarcando un 85% de la varianza en el precio. Esta amalgama de variables es altamente predictiva y puede servir de guía para las estrategias de establecimiento de precios y decisiones productivas en el sector automovilístico.
Análisis de Residuos y Variables: A pesar de que el modelo y las variables seleccionadas describen adecuadamente la variación de precio, los residuos no siguen una distribución normal. Esta anomalía podría deberse a un dato específico que, aunque no presenta una influencia lo suficientemente fuerte como para ser eliminado, debería ser investigado más detalladamente. Esta revisión podría ofrecer un panorama más claro sobre su posible exclusión en futuros análisis. Además, se sugiere reconsiderar la inclusión de "curbweight" en el modelo, ya que su impacto no parece ser tan significativo como el de otras variables. Lo mismo se sugiere para una de las variables dummy de "drivewheel".
En esencia, este análisis otorga perspectivas valiosas a cualquier compañía automovilística, particularmente a una empresa china interesada en incursionar en el mercado estadounidense. Entender las variables que más afectan el precio de los automóviles facilitará decisiones estratégicas bien fundamentadas, maximizando las chances de éxito en un ámbito tan competitivo.

Anexos

Todo el código se puede encontrar en la siguiente liga:
https://github.com/a01752030/PortafolioAnalisis