Reporte “El precio de los autos”

Módulo 1: Estadística para ciencia de datos - Inteligencia artificial avanzada para la ciencia de datos I

María Fernanda Torres Alcubilla A01285041

Resumen.

Una empresa de automóviles china quiere introducir una unidad de fabricación en EUA, se quieren obtener los factores principales que determinan el precio de los automóviles y qué tan bien lo describen, enfocado al mercado estadounidense. Por lo que se hace una análisis exploratorio de los datos y reducción de dimensionalidad para la creación de dos modelos de regresión lineal multivariadas, con y sin interacción y su validación de supuestos a través de métodos estadísticos. En el modelo con interacción se consideran 4 variables y 2 interacciones, obteniendo un coeficiente de determinación ajustado de 0.803 y solamente se cumple el supuesto de residuos con media 0. Por otro lado, en el modelo sin interacción, se consideran 3 variables, se obtuvo un coeficiente de determinación ajustado de 0.7234 y se cumplen los supuestos de residuos con media 0 e independencia.

Introducción

Los precios de los productos varían dependiendo del país de venta, es por esto que, cuando una empresa extranjera quiere introducir una unidad de fabricación en algún país para la producción local y competir con las demás empresas, se debe realizar un análisis. En esta problemática, se trabaja con el mercado de automóviles, donde una empresa china se desea introducir al mercado estadounidense y específicamente se quieren obtener los factores principales que determinan el precio de los automóviles y qué tan bien lo describen, enfocado al mercado estadounidense.

Estos análisis son necesarios por diversas razones, por ejemplo: la competencia en el mercado local, con esto se debe determinar si hay competencia en términos de precios o se debe realizar algún ajuste; el precio frontera que los clientes locales estén dispuestos a pagar y el tiempo de retorno de la inversión con las predicciones de venta y ganancias de los automóviles.

Análisis de resultados

Este análisis se divide en dos secciones, la primera, *Técnicas de procesamiento de datos para el análisis estadístico y construcción de modelos*, y la segunda, *Construcción de un modelo estadístico base.* En general, la primera sección se enfoca en la exploración de datos, donde se analizaron y visualizaron estos y se realizó una selección y preparación de 6 variables con mayor importancia para determinar el precio. La segunda sección se enfoca principalmente en la creación de un modelo de regresión lineal multivariada para la predicción del precio y el análisis de los supuestos del modelo implementado.

Técnicas de procesamiento de datos para el análisis estadístico y construcción de modelos

Para la exploración de los datos se calcularon medidas estadísticas como el promedio, desviación estándar, rango de las variables y percentiles, en el caso de variables numéricas, en el caso de las variables categóricas se obtuvo la moda y frecuencias. Esto permitió tener un primer acercamiento a los datos, donde se observó que ninguna característica tiene valores nulos, las variables *curbweight, enginesize, peakrpm* y *price* presentan un rango grande y variabilidad. En cuanto a las variables categóricas, la variable *CarName* es la que más presenta valores únicos, con un total de 147 modelos distintos, por lo que para futuros análisis se hizo caso omiso de esta variable debido a la gran diversidad.

Se analizó la distribución de los datos y se encontró que todas las variables cuentan con forma asimétrica y algunas con sesgos notorios. En caso de sesgos positivos, este se presenta en las variables *carheight, Stroke* y *Peakrpm* y el sesgo negativo se presenta en las variables *wheelbase, carlenght, carwidth, horsepower, enginesize, citympg, highwaympg, compressionratio, curbwaight* y *price.*

Con el objetivo de conocer la relación entre las variables numéricas se calculó la correlación entre ellas, la cual se puede observar en la *Figura 1*, con esto se obtuvo que hay una relación muy fuerte (0.97) entre las variables de *citympg* y *highwaympg,* además, las variables que presentaron una correlación fuerte positiva con *price* son *curbweight* y *enginesize* con valores de 0.84 y 0.87 respectivamente.

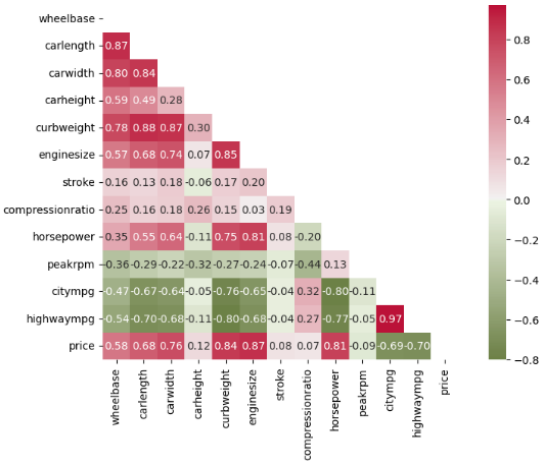

Figura 1. Correlación con variables numéricas

En cuanto a las variables categóricas se realizó un análisis de boxplots por precio donde se buscan separaciones de los distintos elementos de estas variables y el precio. Por ejemplo, en la primera gráfica de la *Figura 2* podemos observar el comportamiento de la variable *enginetype*, donde existen intersecciones entre los distintos elementos, esto puede presentarse como ruido al momento de realizar predicciones ya que el modelo no podrá realizar una separación definida. Por otra parte, las gráficas restantes de la *Figura 2 r*epresentan las variables *drivewheel* y *cylindernumber,* las cuales son las que mayor separación presentan y podemos observar que las intersecciones de los rango intercuartiles son mínimas.

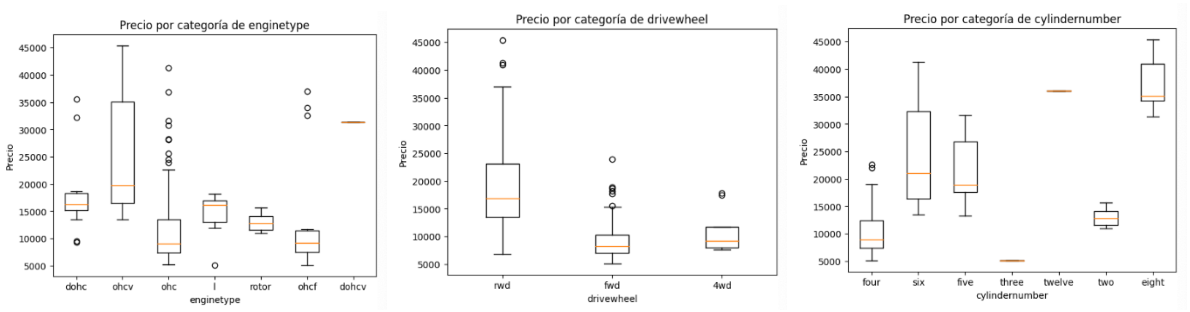

Figura 2. Precio por categoría de variables categóricas

Selección de variables

Por último, se realizó la selección y transformación de 6 variables que describen mejor el precio, los criterios de selección en caso de variables numéricas fue su relación con la variable objetivo y en caso de las categóricas, su separación por categoría. Primero, se consideraron todas las variables numéricas con correlación moderada o mayor con el precio, se obtuvieron un total de 7 variables, después, para evitar que una variable a considerar se pueda describir por medio de otra de este mismo apartado se observa la correlación entre ellas y se elimina la variable que tenga menor correlación con el precio, con este filtro se obtuvieron un total de 4 variables numéricas.

Con esto, las variables seleccionadas son 4 numéricas y 2 categóricas:

- enginesize (numérica)
- curbweight (numérica)
- horsepower (numérica)
- highwaympg (numérica)
- cylindernumber (categórica)
- drivewheel (categórica)

Transformación de datos

Para el tratamiento de los datos, la variable cylindernumber se cambió a numérica ya que son datos ordinales y se realizó una discretización en las variables numéricas. Por último, para el caso de escalas significativamente diferentes entre las variables y su distribución asimétrica, se realizó una transformación Yeo-Johnson para la normalización, obteniendo una media cercana a 0 y varianza a 1, en la *Figura 3* podemos observar esta transformación, donde las variables numéricas no presentan distribución asimétrica.

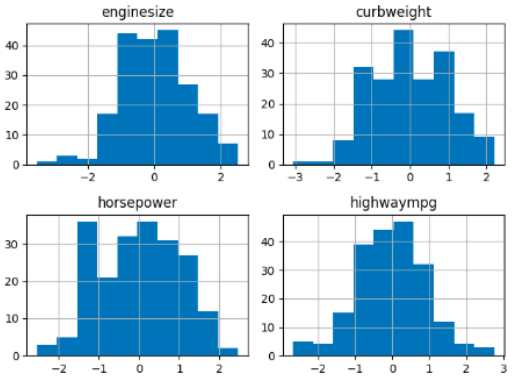

Figura 3. Distribución después de transformación Yeo-Johnson

Construcción de un modelo estadístico base

En la segunda sección se crearon dos modelos estadísticos de regresión lineal multivariada, con y sin interacción, solamente con las variables numéricas.

Creación de modelos

Modelo con interacción

Para la creación del modelo con interacción, se obtuvieron las variables de entrada que tuvieran una correlación fuerte entre ellas, las cuales fueron: enginesize y curbweight y highwaympg con horsepower, el procedimiento fue el siguiente:

H_o: Hay interacción entre horsepower y highwaympg

H_1: No hay interacción

Regla de decisión: Se rechaza H_o si su p-value es mayor a 0.03

p-value: 0.000891, por lo tanto, no hay pruebas suficientes para rechazar la interacción.

H_o: La variable cylindernumber es significativa

H_1: No es significativa

Regla de decisión: Se rechaza H_o si su p-value es mayor a 0.03

p-value: 0.093, por lo tanto, esta variable no es significativa

H_o: Hay interacción entre enginesize y curbweight

H_1: No hay interacción

Regla de decisión: Se rechaza H_o si su p-value es mayor a 0.03

p-value: $1.28e^{-5}$, por lo tanto, no hay pruebas suficientes para rechazar la interacción.

Por lo que el modelo final con interacción se ve de la siguiente manera:

$$price = 11065.31 + 1493.492\, enginesize + 2695.951\, curbweight + 1755.018\, horsepower - 1005.226\, highwaympg + 1348.248\, enginesize * curbweight - 1200.563\, horsepower * highwaympg$$

Con los siguientes resultados, lo que demuestra que nuestro modelo es significativo considerando el mismo nivel de significancia y con una varianza total explicada alta:

R^2	$R^2\ ajustada$	$p - value$
0.8088	0.803	$< 2.2e^{-16}$

Modelo sin interacción

Para el modelo sin interacción, las pruebas de hipótesis tienen la siguiente forma:

H_o: La variable es significativa

H_1: No es significativa

Regla de decisión: Se rechaza H_o si su p-value es mayor a 0.03

Donde se tiene el siguiente proceso de eliminación:

Variable	p-value	Decisión
enginesize	0.23002	Se rechaza H_o
highwaympg	0.0986	Se rechaza H_o

Por lo que nuestro modelo final se ve de la siguiente manera:

$$price = 13276.71 + 3431.807\ curbweight$$
$$+ 2202.562\ horsepower + 2221.619\ cylindernumber$$

Con los siguientes resultados, lo que demuestra que nuestro modelo es significativo considerando el mismo nivel de significancia y con una varianza total explicada moderada:

R^2	$R^2\ ajustada$	$p - value$
0.7275	0.7234	$< 2.2e^{-16}$

Conclusiones

Se obtuvieron dos modelos, con y sin interacción, con un R^2 ajustada de 0.80 y 0.72 respectivamente. El primer modelo toma en consideración las siguientes variables y sus interacciones: enginesize, curbweight, horsepower, highwaympg, enginesize con curbweight y horsepower con highwaympg. El segundo modelo considera las siguientes variables: curbweight, horsepower, cylindernumber.

Validación de modelos

El siguiente paso es la validez de ambos modelos de regresión lineal con un alfa de 0.03 donde se harán las siguientes pruebas con sus siguientes hipótesis y regla de decisión:

- Los residuos se distribuyen como una normal
  - H_0: $los\ datos\ provienen\ de\ una\ población\ normal$
  - H_1: $los\ datos\ no\ provienen\ de\ una\ población\ normal$
  - Regla de decisión: se rechaza H_0 si $p - value < 0.03$
  - Para esta prueba se aplicó el test de Anderson Darling, se realizó el gráfico Q-Q y el histograma de los residuos.
- Media cero de residuos
  - H_0: $los\ residuos\ tienen\ media\ 0$
  - H_1: $los\ residuos\ no\ tienen\ media\ 0$
  - Regla de decisión: se rechaza H_0 si $p - value < 0.03$
  - Se aplica una prueba de hipótesis t-Student para medias
- Los residuos tiene homocedasticidad
  - H_0: $los\ datos\ presentan\ homocedasticidad$

- H_1: $los\ datos\ no\ presentan\ homocedasticidad$
- Regla de decisión: se rechaza H_0 si $p - value < 0.03$
- Para esta prueba se realiza el ncvTest, que se basa en una chi-cuadrada, además se grafican los residuos y valores ajustados.

- Los residuos son independientes
  - Para esta prueba se observa la gráfica de residuos y valores ajustados para ver si hay una tendencia en estos puntos.

Modelo con interacción

Los resultados de este modelo son los siguientes:

H_0	p-value	Conclusión
Los datos provienen de una población normal	$4.185e^{-11}$	Se rechaza H_0
Los residuos tienen media 0	1	Se rechaza H_0
Los datos presentan homocedasticidad	$< 2.22e^{-16}$	Se rechaza H_0

En la *Figura 4* se observa el gráfico Q-Q y el histograma de los residuos, en la primera observamos nuestros residuos como los puntos y la línea recta como el comportamiento normal, se resalta que nuestros residuos no siguen la distribución normal en las colas. Por otro lado, en el histograma, donde la curva azul es la teórica de la normalidad y la roja la de nuestros residuos, observamos una gran diferencia en nuestros residuos con la normal y que los datos tienen una forma leptocúrtica, ya que su curva está más alargada que la distribución normal. Estas gráficas ayudan a visualizar el rechazo de la normalidad.

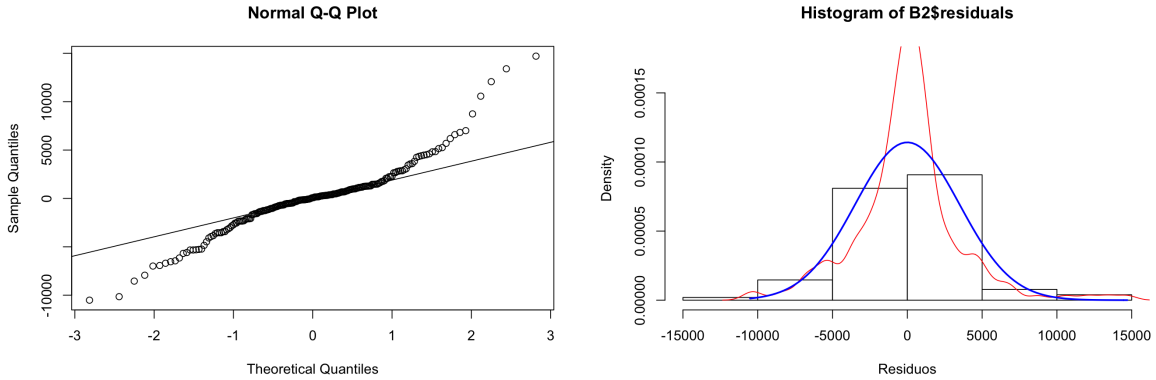

Figura 4. Gráficas de supuesto de normalidad, modelo con interacción

En el caso del supuesto de homocedasticidad, en la *Figura 5* se tiene la gráfica de residuos y valores ajustados, donde no se observa homocedasticidad ya que en el lado izquierdo se tiene una mayor densidad de puntos que del lado derecho, además, estos tienden a abrirse mientras mayor sea el valor ajustado, por lo que se presenta simetría. Por lo tanto, debido a la ausencia de sesgo decimos que nuestros datos no presentan independencia.

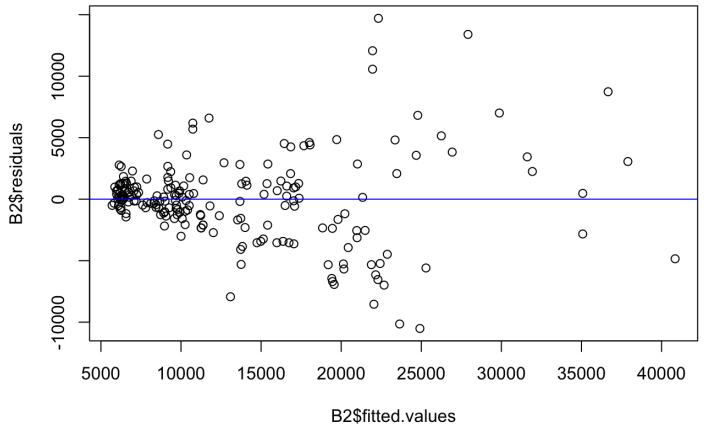

Figura 5. Gráfica de supuesto de homocedasticidad, modelo con interacción

Modelo sin interacción

Los resultados de este modelo son los siguientes:

H_0	p-value	Conclusión
Los datos provienen de una población normal	$3.963e^{-12}$	Se rechaza H_0
Los residuos tienen media 0	1	Se rechaza H_0
Los datos presentan homocedasticidad	$< 2.22e^{-16}$	Se rechaza H_0

En la *Figura 4* se observa el gráfico Q-Q y el histograma de los residuos, en la primera observamos de igual manera, se resalta que nuestros residuos no siguen la distribución normal en las colas. De la misma forma, en el histograma, observamos una gran diferencia en nuestros residuos con la normal, que los datos tienen una forma leptocúrtica, además, a diferencia del modelo anterior, se observa un sesgo negativo.

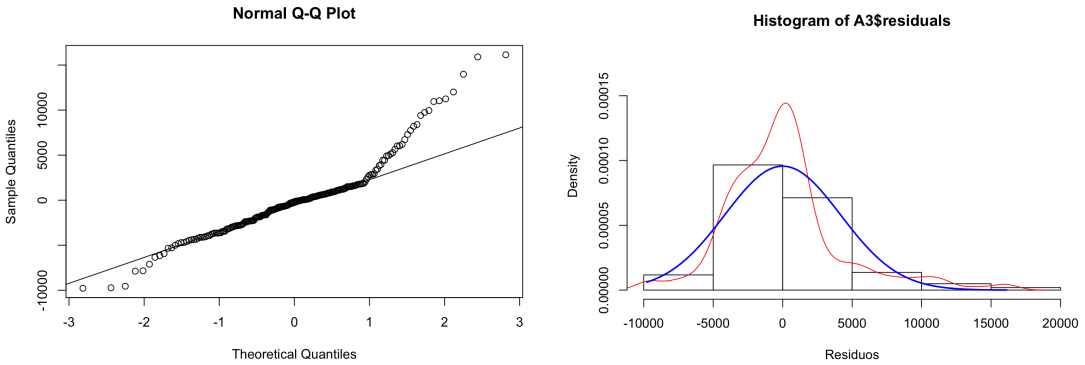

Figura 6. Gráficas de supuesto de normalidad, modelo sin interacción

En nuestra gráfica de residuos y valores ajustados, no se observa homocedasticidad ya que en el lado izquierdo hay una mayor densidad de puntos que del lado derecho, además,

estos tienden a la forma de parábola postiva, lo que indica un sesgo. Por lo tanto, debido al sesgo decimos que nuestros datos, además de heterocedasticidad, presentan independencia.

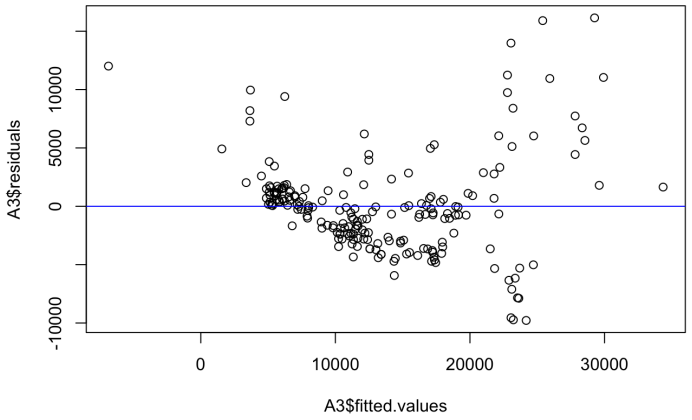

Figura 7. Gráfica de supuesto de homocedasticidad, modelo sin interacción

Conclusión

En resumen, la validación del modelo con interacción tuvo los siguientes resultados:

- Los datos no provienen de una distribución normal
- Los residuos tienen media 0
- Los datos tienen heterocedasticidad y dependencia

Y el modelo sin interacción:

- Los datos no provienen de una distribución normal
- Los residuos tienen media 0
- Los datos tienen heterocedasticidad e independencia

Ninguno de los dos modelos cumple con todos los supuestos para la regresión lineal, por lo que se concluye que no son modelos confiables. Sin embargo, se escogerá el modelo sin interacción ya que este en las pruebas de hipótesis cumplió con la independencia, teniendo más supuestos aceptados que el modelo con interacción. Aunque este tenga un R^2 un poco menor que al modelo con interacción.

Análisis de datos influyentes

Por último, se realiza un análisis de datos atípicos e influyentes del modelo sin interacción, esto con el objetivo de mejorar este ya que estos datos pueden perjudicar el desempeño. Para esto se utiliza la distancia de Cook, donde un dato se considera influyente si su distancia es mayor a 1. La tabla con los posibles datos influyentes mostró que la distancia máxima es de 0.26, por lo que se concluye que ningún dato es influyente y el modelo no puede ser mejorado en este aspecto.

Conclusión

Los dos modelos implementados, con y sin interacción, no cumplen con todos los supuestos, sin embargo, tienen un R^2 normal y ajustada aceptables por lo que en este aspecto se puede clasificar como un modelo bueno, sin embargo, no puede ser tan

confiable en sus predicciones debido a la falta de aceptación de la mayoría de los supuestos.
En cuanto al análisis del modelo considerando el contexto, se querían obtener los factores principales que determinen el precio de los automóviles y qué tan bien lo describen, por lo que se puede decir que, considerando ambos modelos:

- El precio base promedio de los automóviles, sin considerar alguna característica, es de 12,100 USD.
- La variable curbweight es la que más aumenta el precio de los automóviles, teniendo una relación positiva.
- La variable horsepower es la segunda variable que más aumenta el precio.

Además, considerando las interacciones, observamos que el producto de enginesize y curbweight tienen impacto positivo en el precio, por otro lado, la relación entre horsepower y highwaympg con el precio es negativa.

Anexos

Carpeta con documentos de análisis y base de datos

Portafolio de análisis

Portafolio de implementación