INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES MONTERREY

CAMPUS MONTERREY

Módulo 1: Estadística para ciencia de datos

Inteligencia Artificial Avanzada para la Ciencia de Datos I (Grupo 101)

Maestros:

Blanca Rosa Ruíz Hernández

Reporte Final de “El Precio de los Autos”

Ana Lucía Cárdenas Pérez A01284090

Monterrey, Nuevo León

12 de septiembre del 2023

Resumen

Una empresa automovilista busca los principales factores que pueden determinar el precio de un vehículo. Se busca llevar a cabo un análisis de las variables para encontrar a las que puedan predecir el precio de los autos, probarlo para ver si este es fiable si se necesita otro modelo u otros parámetros.

Se llevó a cabo una transformación de datos, análisis de estadísticas (media, desviación estándar, cuantiles), se crearon visualizaciones con histogramas para las variables cuantitativas, análisis de correlación para seleccionar las variables.

Se hicieron pruebas con modelos de regresión lineal y regresión lineal múltiple para encontrar que modelo nos daba mejores predicciones y se llegó a la conclusión de que en este caso la regresión lineal múltiple era el modelo para utilizar.

Introducción

¿Qué modelo y variables nos pueden ayudar a predecir el precio de un vehículo?

¿Qué variables son las más relevantes en el impacto al precio de un automóvil en Estados Unidos?

Con el rápido crecimiento del mercado de la industria automotriz, la competencia en la industria se ha vuelto cada vez más fuerte ya que en el caso de las nuevas empresas. Estas empresas buscan oportunidades para poder expandirse a otros mercados, en este caso, una empresa China busca expandirse al mercado estadounidense, el cual es uno de los más competitivos. La empresa tiene como objetivo abrir una unidad de fabricación en Estados Unidos para que estos puedan ser producidos localmente y así poder competir con los otros fabricantes que ya están establecidos en el país.

Para poder tomar las mejores decisiones, la empresa decidió contratar a una empresa consultora que los ayudará con el estudio de mercado en este nuevo

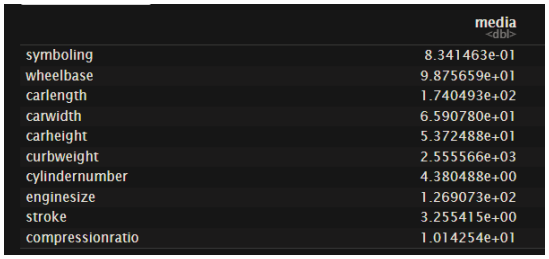
mercado. La consultoría tiene como objetivo identificar los factores que pueden ser factores influyentes en el precio de los automóviles para poder utilizarlos en un modelo de predicción para poder obtener precio estimado del automóvil.

La importancia de este problema radica en que el éxito de la empresa china en el mercado estadounidense depende de su capacidad para comprender y adaptarse a las dinámicas específicas de este mercado. Los clientes estadounidenses son conocidos tener preferencias únicas en cuanto a automóviles.

En este reporte vamos a explorar los modelos que utilizamos y comparamos para así poder proporcionar un modelo que puedan utilizar el cual tenga el menor porcentaje de error posible, para que la empresa pueda tomar las decisiones correctas que los puedan a llevar a obtener los resultados deseados con el modelo de predicción.

Análisis de los Resultados

1. *Configuración y Carga de Datos*
    a. Primero cargamos las librerías necesarias y cargamos el archivo a utilizar y lo guardamos en un dataframe, en este caso, guardamos el archivo “precios_autos.csv” en un df llamado “data”. También transformarnos datos como la columna de “cylindernumber” para que este tuviera valores numéricos en lugar de caracteres y así poder realizar un mejor análisis y predicción.
2. *Análisis de Variables*
    a. Para el análisis de las variables cuantitativas, calculamos las medidas estadísticas de las variables, en este caso mostramos la tabla de la media de las variables cuantitativas.



	media <dbl>
symboling	8.341463e-01
wheelbase	9.875659e+01
carlength	1.740493e+02
carwidth	6.590780e+01
carheight	5.372488e+01
curbweight	2.555566e+03
cylindernumber	4.380488e+00
enginesize	1.269073e+02
stroke	3.255415e+00
compressionratio	1.014254e+01

b. *Análisis de las variables cualitativas*

Creamos una visualización con tablas de frecuencia, las cuales se muestra con el siguiente ejemplo, creado con la variable “carbody”:

Var1 <fctr>	Freq <int>
convertible	6
hardtop	8
hatchback	70
sedan	96
wagon	25

Estas tablas nos ayudaron a entender los valores promedios de las características que los autos tienen y así poder analizar y entender que es lo que los clientes buscan en sus autos. La tabla de frecuencia también nos ayuda a ver que es lo que más se vende en cuestión de las características físicas. Como el ejemplo de “carbody”, donde podemos ver que tipos de autos son los que la gente está buscando/comprando más.

c. *Análisis de Colinealidad*

Buscamos los valores de correlación entre todas las variables cuantitativas en comparación al precio, donde obtuvimos los siguientes valores:

[1] "Coeficiente Correlación	[,1]
symboling	-0.07997822
wheelbase	0.57781560
carlength	0.68292002
carwidth	0.75932530
carheight	0.11933623
curbweight	0.83530488
cylindernumber	0.71830490
enginesize	0.87414480
stroke	0.07944308
compressionratio	0.06798351
horsepower	0.80813882
peakrpm	-0.08526715
citympg	-0.68575134
highwaympg	-0.69759909
price	1.00000000

Esta tabla fue la que nos ayudó a seleccionar las variables que vamos a utilizar en el modelo de predicción. Para la selección nos enfocamos en aquellas que tuvieran su valor más cercano a 1 o -1. Consideramos la variable de citympg ya que algo en lo que muchas personas se enfocan al elegir un automóvil, es el gasto de gasolina cuando manejan por la ciudad, y al ser un valor negativo, tajmbién nos puede ayudar a balancear los

resultados para una predicción más acertada.

3. *Calidad de Datos*
    a. Para entender un poco mejor la calidad de los datos del dataset llevamos a cabo dos procesos, el primero fue identificar si alguna de las variables cuantitativas tenía valores faltantes. Segundo, utilizando la función de “boxplot.stats”, pudimos obtener los outliers que se visualizan en un una gráfica estilo “boxplot”, y así poder tenerlos identificados. Este proceso se vió de esta manera:

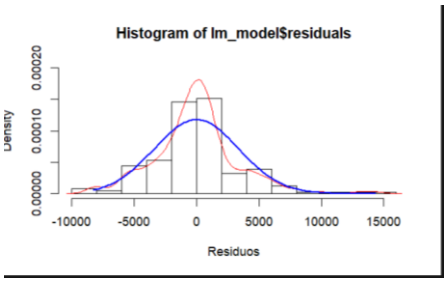
```
[1] "Valores Faltantes Price"
[1] 0
[1] "Valores Faltantes Engine Size"
[1] 0
[1] "Valores Faltantes Curb Weight"
[1] 0
[1] "Valores Faltantes Horse Power"
[1] 0
[1] "Valores Faltantes Car Width"
[1] 0
[1] "Valores Faltantes Cylinder Number"
[1] 0
[1] "Valores Faltantes Car Length"
[1] 0
[1] "Outliers Engine Size"
 [1] 209 209 209 258 258 326 234 234 308 304
[1] "Outliers Curb Weight"
integer(0)
[1] "Outliers Horse Power"
[1] 262 200 207 207 207 288
[1] "Outliers Car Width"
[1] 71.4 71.4 71.4 71.7 71.7 71.7 72.0 72.3
[1] "Oultiers Cylinder Number"
 [1]  6  5  5  5  5  5  5  6  6  6  6  6  6  3  6  6 12  2  2  2  2
[22]  5  5  5  5  8  8  8  8  6  6  6  6  6  6  6  6  6  6  8  6  6  6
[43]  6  5  6  6
[1] "Outliers City MPG"
[1] 47 49
```

4. *Pruebas de modelos*

    Los modelos por probar en este análisis fueron:

    a. *Regresión lineal simple*

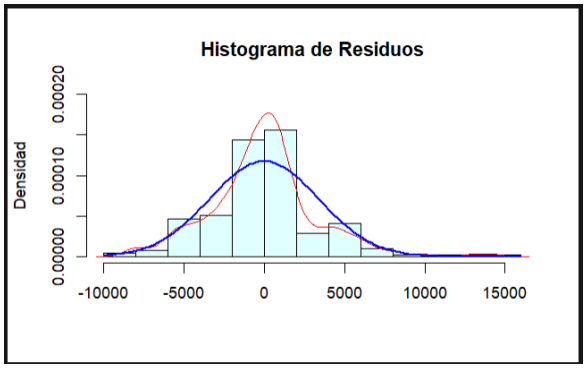
    Utilizamos histograma con el modelo de regresión, para visualizar la gráfica de residuales del modelo con las variables numéricas y se obtuvieron los siguientes resultados.



Podemos ver en este histograma, que los datos de la línea roja no siguen la forma exacta de la azul, por lo que la desviación sigue existiendo. La curtosis que se presenta en este histograma es una curtosis "leptocúrtica", es decir, que contienen datos atípicos. En el análisis anterior pudimos encontrar que hay algunas variables como cylinder number y engine size, que tienen muchos outliers en sus datos.

b. *Regresión Lineal Múltiple*

También llevamos a cabo el modelo de regresión multiple con el cual también trabajamos con las variables cuantitativas elegidas anteriormente. Utilizando una fórmula básica de las variables independientes numéricas, pudimos obtener un valor de r^2 = 0.8145 (mismo valor obtenido en el primer proceso de regresión lineal simple), pero una ventaja de este modelo es que podemos obtener que variables podrían darnos un mejor resultado, y así poder llevar a cabo una nueva prueba. En la segunda prueba con regresión lineal múltiple, utilizamos 4 de las 6 variables que usamos inicialmente y con la misma prueba para obtener r^2, ahora obtuvimos un valor de 0.8162, el cual fue una diferencia de 0.0017, y el siguiente histograma:



5. *Conclusión*

Se llevaron a cabo pruebas con dos modelos de regresión, uno lineal y otro lineal múltiple utilizando las mismas variables en ambas. El modelo de regresión lineal utilizó todas las variables cuantitativas que elegimos para darnos el valor

de R squared, el cual fue de 0.8145, mientras que al realizar el proceso de regresión lineal múltiple pudimos obtener cuales columnas podrían darnos un mejor valor en r squared. Ambos análisis se iniciaron con las columnas de "carwidth", "curbweight", "cylindernumber", "enginesize", "horsepower", "citympg" y "price". En el de regresión lineal simple todo el proceso se llevó a cabo con esas columnas, mientras que en regresión multiple iniciamos con esas columnas y después obtuvimos una nueva formula que nos ayudó a obtener un mejor valor de r squared, la nueva formula utilizó las columnas "price", "curbweight", "enginesize" y "horsepower", lo cual nos dió un valor de r squared de 0.8162, una diferencia de 0.0017.

Por lo que llegamos a la conclusión de que el modelo de regresión lineal múltiple es una mejor opción ya que al tener muchas variables que pueden mover nuestros resultados, este nos puede dar aquellas que nos van a dar los mejores resultados, o mejor posible predicción.

6. *Referencias*

Autocosmos. (2022, April 29). El consumo de gasolina es el factor principal de compra de un automóvil. Retrieved September 12, 2023, from Autocosmos website: https://noticias.autocosmos.com.mx/2022/04/29/el-consumo-de-gasolina-es-el-factor-principal-de-compra-de-un-automovil

7. *Link a GitHub con los archivos con los que trabajamos durante el bloque*

https://github.com/analucia2107/Evidencia_Modulo1_A01284090