DATA SCIENCE – CODER HOUSE

Manuel Muñoz – Venta de vehículos en USA

Introducción:

I.I.- Descripción Temática

El presente análisis tiene como objetivo investigar y comprender los factores que influyen en los precios de los automóviles en el mercado. Para ello, utilizaremos un conjunto de datos que contienen información sobre diferentes atributos de los vehículos como marca, modelo, año de fabricación, kilometraje, color, entre otros. A través de este estudio, buscaremos identificar las relaciones entre estas características y los precios de venta de los automóviles.

El análisis se llevará a cabo mediante técnicas estadísticas y de visualización de datos, con el fin de obtener insights valiosos que puedan ser útiles para compradores, vendedores y otros actores del mercado automotriz.

El objetivo principal del proyecto será analizar cuál podría llegar a ser el precio de venta de un vehículo que no pertenezca al dataset. La intención es brindarle al modelo los datos del vehículo como la marca, kilometraje, estado, etc. para que pueda predecir un precio de venta potencial.

Considerando las variables identificadas en el data set y con el objetivo de obtener un valor predecible continuo, utilizaremos el modelo de regresión lineal con múltiples variables.

I.II.- Hipótesis:

La hipótesis nula para este análisis es que no hay diferencias significativas en los precios de los automóviles entre las diferentes categorías de las variables estudiadas. En otras palabras, los precios de venta de los automóviles son iguales independientemente de la marca, el modelo, el año de fabricación u otras características consideradas en el análisis.

Esta hipótesis nula establece que no hay efectos discernibles de las variables consideradas en los precios de venta de los automóviles y proporciona una base para comparar y evaluar cualquier relación que se encuentre en el análisis.

Durante el transcurso del análisis, buscaremos evidencia estadística para rechazar la hipótesis nula (H0), que establece que no hay diferencias significativas en los precios de los automóviles entre las diferentes categorías de las variables estudiadas. En lugar de eso, nuestro objetivo será encontrar evidencia que sugiera la presencia de diferencias significativas en los precios de venta de los automóviles, lo que respaldaría la hipótesis alternativa (H1) de que existen relaciones o efectos reales entre las variables consideradas.

I.III.- Objetivo:

El objetivo principal del proyecto generar un modelo predictivo, que ayude a identificar cual podría llegar a ser el valor de venta de un vehículo en el mercado.

Por otro lado, se realiza un análisis exploratorio de los datos para poder conocer el comportamiento de los mismos.

- ¿Existen valores nulos dentro del dataset que dificulten el análisis?
- ¿En qué porcentaje del total inciden?
- ¿Existe alguna columna de código único que afecte negativamente al modelo?
- ¿Dentro de las columnas con variables clasificatorias, existe algún valor que sea repetitivo o que tenga mayor incidencia sobre los datos? Como por ejemplo que predominen 10 marcas de autos. Lo mismo podría suceder con los modelos.
- ¿Existen valores atípicos?
- ¿En qué porcentaje del total inciden?
- ¿Se deben eliminar? ¿Por qué?
- ¿Hay un mayor número de ventas en algún estado en particular?
- ¿Cuáles son los precios de venta más frecuentes? ¿Hay algún rango de precios predominante?
- ¿Existe algún tipo de relación entre las marcas de los vehículos y el precio de venta?
- ¿Qué variables están relacionadas entre sí?
- ¿En caso de que haya alguna relación, es lineal?
- ¿Hay alguna tendencia sobre la transmisión utilizada en Estados Unidos?
- Se evidencia en el dataset que haya marcas de vehículos que son más premium y/o caras que otras?
- Si bien contamos con la fecha de venta del vehículo, no se especifica la antigüedad que tiene el vehículo cuando se vende. ¿Es un dato relevante?
- Que variables ayudan en la predicción de un potencial valor de venta?

I.V.- Fuente del dataset:

https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

I.VI.- Metadata:

El dataset seleccionado cuenta con las siguientes características:

- Número de filas: 558.837
- Número de columnas: 16
- Detalle de columnas:
    1. Year: Año de fabricación del vehículo (por ejemplo, 2015)
    2. Make: Marca o fabricante del vehículo (por ejemplo, Kia, BMW, Volvo)
    3. Model: Modelo específico del vehículo (por ejemplo, Sorento, Serie 3, S60, serie 6 Gran Coupé)
    4. Trim: Designación adicional para una versión o paquete de opciones particular del modelo (por ejemplo, LX, 328i SULEV, T5, 650i)
    5. Body: Tipo de carrocería del vehículo (por ejemplo, SUV, Sedán)
    6. Transmission: Tipo de transmisión en el vehículo (por ejemplo, automática)

7. VIN: Número de Identificación del Vehículo, un código único utilizado para identificar vehículos individuales
8. State: Estado en el que se encuentra o está registrado el vehículo (por ejemplo, CA para California)
9. Condition: Representación numérica de la condición del vehículo (por ejemplo, 5.0)
10. Odometer: Millaje o distancia recorrida por el vehículo
11. Color: Color exterior del vehículo
12. Interior: Color interior del vehículo
13. Seller: Entidad o empresa que vende el vehículo (por ejemplo, Kia Motors America Inc, Financial Services Remarketing)
14. MMR: Manheim Market Report, una herramienta de precios utilizada en la industria automotriz norteamericana. Es el equivalente a Acara, en Argentina. Se usa para tener precios de referencia del mercado de los automóviles.
15. Selling Price: Precio al que se vendió el vehículo
16. Sale Date: Fecha y hora en que se vendió el vehículo

- Tipos de variables:
  - Categóricas: Make, Model, trim, body, transmission, state, color, interior, seller,
  - Numéricas: Año, condition (flotante), odometro, mmr, sellingprice.
  - Fecha: Saledate

VI.- : Colab

https://colab.research.google.com/drive/1d50IiaDbX1jJ_UCGN_zd2VxOpY4p_ZfX?usp=sharing

VII.- :

VII.- Resumen final :