PROYECTO INTEGRADOR MODULO 6

Proyecto de estudio de mercado

Cliente: Automotriz China

Descripción del Proyecto:

Objetivo General:

El objetivo general de este proyecto de ciencia de datos es analizar y comprender las diferencias entre los vehículos de gama alta y baja en el mercado automotor actual, con el fin de proporcionar a nuestro cliente, una automotriz china, que busca ingresar al mercado de automóviles, la información necesaria para tomar decisiones estratégicas en cuanto a la cotización de vehículos.

Para lograr este objetivo general, se desglosará en los siguientes objetivos específicos:

Objetivos Específicos:

1.- Clasificación de Vehículos: Desarrollar un modelo de clasificación con aprendizaje supervisado que permita clasificar los vehículos en dos categorías: "baratos" y "caros". Esta clasificación se basará en la mediana de los precios como punto de corte.

2.- Predicción de Precios: Implementar un modelo de regresión con aprendizaje supervisado para predecir el precio final de los vehículos. Esto ayudará a nuestro cliente a determinar los precios adecuados para sus vehículos en función de sus características.

3.- Análisis Exploratorio de Datos (EDA): Realizar un análisis exploratorio de datos exhaustivo para comprender las características clave de los vehículos presentes en el mercado actual. Esto incluye el análisis de variables como tipo de combustible, tipo de motor, número de puertas, etc.

4.- Reprocesamiento de Datos: Preparar y limpiar los datos adecuadamente, lo que incluye la codificación de variables categóricas, manejo de valores faltantes y escalados de características numéricas.

5.- Modelado y Evaluación del Modelo de Clasificación: Seleccionar y entrenar un modelo de clasificación que pueda distinguir entre vehículos baratos y caros. Evaluar el rendimiento del modelo utilizando métricas como precisión, recall y F1-score.

6.- Modelado y Evaluación del Modelo de Regresión: Seleccionar y entrenar un modelo de regresión que pueda predecir el precio de los vehículos. Evaluar el rendimiento del modelo utilizando métricas como el error cuadrático medio (MSE) y el coeficiente de determinación (R^2).

7.- Generación de Resultados: Generar archivos de texto plano con las predicciones de los modelos de clasificación y regresión para su entrega al cliente.

8.- Informe Final: Preparar un informe final que resuma los hallazgos clave, detalles sobre los modelos utilizados y sus métricas de rendimiento, y proporcionar recomendaciones para la cotización de vehículos.

9.- Presentación al Cliente: Presentar los resultados y conclusiones al cliente de manera clara y comprensible, destacando las ventajas y aplicaciones prácticas de los modelos desarrollados.

Información general del DataFrame

RangeIndex: Indica que el índice de las filas del DataFrame es un rango que va desde 0 hasta 204. Esto significa que el DataFrame tiene 205 filas en total.

Data columns: Muestra la información sobre las columnas del DataFrame, incluyendo el nombre de la columna, la cantidad de valores no nulos y el tipo de datos (Dtype) de cada columna. Esto significa que el DataFrame tiene 26 Columnas.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **NOMBRE DE LAS VARIABLES** | **DESCRIPCION** | **Columna1** | **Columna2** | **Columna3** | **Columna4** | **Columna5** |
| car\_1D | Número de Identificación del vehículo en la base de datos | | | |  |  |
| symboling | Calificación de riesgo asociada al vehículo, +3 es riesgoso poco seguro, -3 es poco riesgoso muy seguro | | | | | |
| CarName | Nombre de fantasía del vehículo | | |  |  |  |
| fueltype | Tipo de combustible | |  |  |  |  |
| aspiration | Tipo de aspiración del motor | |  |  |  |  |
| doornumber | Número de puertas | |  |  |  |  |
| carbody | Tipo de carrocería del vehículo | |  |  |  |  |
| drivewheel | Ubicación del volante del conductor | | |  |  |  |
| enginelocation | Ubicación del motor en el vehículo | | |  |  |  |
| Wwheelbase | Distancia entre ejes | |  |  |  |  |
| carlength | Longitud del vehículo | |  |  |  |  |
| carwidth | Ancho del vehículo | |  |  |  |  |
| carheight | Altura del vehículo | |  |  |  |  |
| curbweight | Peso del vehículo sin carga ni ocupantes | | |  |  |  |
| enginetype | Tipo de motor |  |  |  |  |  |
| cylindernumber | Número de cilindros del motor | |  |  |  |  |
| enginesize | Tamaño del motor | |  |  |  |  |
| fuelsystem | Sistema de administración de combustible del motor | | | |  |  |
| boreratio | Relación diámetro/carrera de los pistones del motor | | | |  |  |
| stroke | Volumen de cilindrada | |  |  |  |  |
| compressionratio | Relación de compresión del aire dentro del motor | | | |  |  |
| horsepower | Potencia del vehículo, en caballos de fuerza (HP) | | | |  |  |
| peakrpm | Revoluciones máximas que soporta el motor | | |  |  |  |
| citympg | Consumo en ciudad, en millas por galón de combustible | | | |  |  |
| highwaympg | Consumo en ruta, en millas por galón de combustible | | | |  |  |
| price | Precio del vehículo | |  |  |  |  |

(Dtype) indica que hay columnas con valores de tipo int64 (números enteros) y float64 (números decimales), así como columnas con valores de tipo object (cadenas de texto u otros tipos no numéricos).Esta información ayuda a comprender la estructura y las características del DataFrame, lo que es esencial para realizar análisis y modelado de datos.

"Non-Null Count" con un valor de 205 en una columna, significa que esa columna tiene 205 valores válidos en todas sus filas, lo que indica que no hay valores faltantes en esa columna en el DataFrame. En otras palabras, no hay celdas vacías o valores no registrados.

Los valores faltantes al realizar el Análisis de datos y Reprocesamiento, pueden afectar el rendimiento de los modelos de Machine Learning y los resultados del análisis.

# Estadísticas descriptivas

df.describe() muestran estadísticas descriptivas para las columnas numéricas de tu conjunto de datos.

Interpretación de lo que significan estos resultados:

count: Este valor indica la cantidad de observaciones (filas) que tienen valores válidos en la columna. Por ejemplo, en la columna "car\_ID", (Número de Identificación del vehículo) en la base de datos hay 205 observaciones válidas.

mean: La media (promedio) de la columna. Por ejemplo, la media de la columna "wheelbase" (Distancia entre ejes) es aproximadamente 98.76.

std: La desviación estándar, que mide la dispersión de los datos alrededor de la media. Valores más altos de desviación estándar indican una mayor variabilidad en los datos.

min: El valor mínimo en la columna. Por ejemplo, el valor mínimo en la columna "wheelbase" (Distancia entre ejes) es 86.6.

25%: El percentil 25, que es el valor por debajo del cual se encuentra el 25% de los datos. Por ejemplo, el percentil 25 en la columna "wheelbase" (Distancia entre ejes)es 94.5.

50%: El percentil 50, que es la mediana de la columna. Es el valor que divide los datos en dos mitades iguales. Por ejemplo, la mediana en la columna "wheelbase" (Distancia entre ejes) es 97.0.

75%: El percentil 75, que es el valor por debajo del cual se encuentra el 75% de los datos. Por ejemplo, el percentil 75 en la columna "wheelbase" (Distancia entre ejes) es 102.4.

max: El valor máximo en la columna. Por ejemplo, el valor máximo en la columna "wheelbase" (Distancia entre ejes) es 120.9.

Estas estadísticas descriptivas proporcionan una visión general de la distribución de los datos en cada columna numérica. Por ejemplo, puedes observar que la media de "price" (precio) es de aproximadamente 13276.71, lo que da una idea del valor promedio de los vehículos en el conjunto de datos. Del mismo modo, se puede analizar la variabilidad en las dimensiones de los vehículos, el tamaño del motor, la potencia, el consumo de combustible, entre otros.

Estas estadísticas son útiles para comprender la naturaleza de los datos y pueden ayudar a tomar decisiones sobre cómo preprocesar y modelar los datos en pasos posteriores a este proyecto de Machine Learning.

Visualizaciónes

Matriz de Correlación: con 4 variables: precio, caballos de fuerza, peso en vacío y tamaño del motor. El gráfico es una cuadrícula cuadrada con 4 filas y 4 columnas. El gráfico está codificado por colores con rojo que representa una alta correlación y azul que representa una baja correlación. La diagonal de arriba a la izquierda a la derecha es toda roja, lo que representa una correlación perfecta de 1.00 entre cada variable y sí misma. La correlación más alta es entre el precio y el tamaño del motor en 0.87, representada por un cuadrado beige en la esquina inferior derecha. La correlación más baja es entre los caballos de fuerza y el peso en vacío en 0.75, representada por un cuadrado azul en el centro del gráfico. El gráfico se titula “Matriz de Correlación”

Relevancia para la Automotriz China: Esta matriz de correlación es relevante para la automotriz china, porque les ayuda a comprender qué características de sus vehículos tienen una fuerte influencia en el precio. Como se observa, existe una correlación positiva fuerte entre la potencia del motor y el precio, esto sugiere que los vehículos con motores más potentes tienden a tener precios más altos. Esto puede influir en las decisiones de diseño y marketing de la automotriz, ya que podrían enfocarse en promocionar vehículos con motores más potentes (mercado norteamericano, por ejemplo, valora mucho los motores y vehículos muy grandes), y si eso es lo que los consumidores valoran en su mercado objetivo.

Tabla Comparación de Precios Promedio por Modelo:

Esta diferencia promedio de precios entre los modelos más caros y los modelos más económicos, nos indica que en nuestro mercado de automóviles, los modelos más caros tienen un precio promedio aproximadamente de 26247,56 $, más alto que los modelos más económicos. Esto refleja la diversidad de opciones de precios que ofrecemos a los consumidores y puede influir en las estrategias de marketing y desarrollo de productos, es especial, para el mercado latinoamericano, que prefieren los precios finales bajos.

Gráfico de Barras: Relación Precio Promedio y tipo de motor

En este gráfico, estamos mostrando el precio promedio de los automóviles según el tipo de motor que tienen. Cada barra representa un tipo de motor diferente. Este gráfico, le hemos agregado etiquetas de precio en la parte superior de cada barra para ver claramente cuánto cuesta, en promedio, un automóvil con ese tipo de motor.

Por ejemplo, si miramos la barra correspondiente al tipo de motor 'dohc', podemos ver que el precio promedio es de $31400.50. De manera similar, puedes observar los precios promedio de otros tipos de motores.

Esto nos proporciona una visión rápida de cómo los precios varían según el tipo de motor. Esta información nos permite comprender mejor el mercado y tomar decisiones estratégicas sobre los vehículos que deseas ofrecer. Por ejemplo, si estás buscando ingresar al mercado con automóviles de precio competitivo, podrías considerar tipos de motores que tienden a tener precios promedio más bajos.

Gráfico de Barras Apiladas.

Cada barra representa un tipo de combustible, y la altura de la barra indica el promedio de millas por galón (MPG) que los vehículos de ese tipo de combustible pueden recorrer en condiciones de conducción en la ciudad.

Por ejemplo, si observamos la barra correspondiente al tipo de combustible 'Gasolina', podemos ver que los vehículos que utilizan gasolina tienen un promedio de aproximadamente 30.3 millas por galón en ciudad. De manera similar, los vehículos diésel tienen un promedio de alrededor de 24,7 millas por galón en ciudad.

Este gráfico es útil para comprender cómo varía la eficiencia en el consumo de combustible entre los diferentes tipos de combustibles, lo que puede influir en las decisiones estratégicas sobre qué tipo de vehículos ofrecer en el mercado. Si estamos buscando vehículos más eficientes en ciudad, este análisis nos proporciona información valiosa sobre cuál podría ser la mejor opción en términos de tipo de combustible, los cuales serían los preferidos en el mercado europeo, que prefieren el bajo consumo.

Vehículos considerados de Gama Alta y Gama Baja en función a su precio

Este gráfico muestra la asignación de los vehículos en dos categorías principales: "Gama Alta" y "Gama Baja" en función de su precio. La mediana de los precios de todos los vehículos se utilizó como punto de corte para determinar a qué categoría pertenece cada vehículo.

"Gama Alta": Esta categoría incluye los vehículos cuyo precio está por encima de la mediana. Estos vehículos se consideran de alta gama en términos de precio en comparación con otros vehículos del mercado.

"Gama Baja": Por otro lado, esta categoría incluye los vehículos cuyo precio está por debajo de la mediana. Estos vehículos se consideran de gama baja en términos de precio en comparación con otros vehículos del mercado.

La tabla presenta una lista de nombres de vehículos en cada una de estas categorías para que se pueda identificar rápidamente a qué categoría pertenece cada modelo. Esta información puede ser valiosa para tomar decisiones estratégicas en el mercado automotriz, ya que ayuda a entender cómo se distribuyen los vehículos en función de su precio y cómo se pueden adaptar las estrategias de marketing y ventas en consecuencia.

Mercado automotor de Norteamérica:

Se toma en cuenta dos factores claves: longitud y peso, características esenciales que a menudo son apreciadas por los consumidores en este mercado. Los vehículos que cumplen con estos criterios pueden tener un mayor potencial de éxito en el mercado de Norteamérica, ya que se alinean con las preferencias y las tendencias de los consumidores en esta región.