**Proyecto Final Data Science II**

***"****Análisis de precios de autos usados****"***

**Profesor:** Gustavo Benítez

**Tutores:** Hugo Mon, Ignacio Manuel Fernández

**Alumno:** Rodríguez Damián

**Comisión:** 61220

**Dataset:** [**https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data**](https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data)

**Abstract**

Este dataset contiene información sobre autos usados, con el objetivo de predecir el precio de los vehículos. La base de datos incluye diversas columnas que describen las características de los autos, como el año de fabricación, el fabricante, el modelo, el estado, el tipo de combustible, el kilometraje (odómetro), y más. Sin embargo, se observa una cantidad significativa de valores faltantes en varias de las columnas. Las columnas más afectadas por datos ausentes son "condition" (condición del auto), "cylinders" (número de cilindros), "manufacturer" (fabricante), "model" (modelo), y "VIN" (número de identificación del vehículo). Además, hay columnas relacionadas con la localización, como la región, el estado y las coordenadas geográficas, algunas de las cuales también presentan valores faltantes. La columna objetivo, "price", no tiene valores ausentes, lo que es fundamental para el objetivo de predicción.

La tarea de predicción se centra en predecir el "price" utilizando las características disponibles, aunque es necesario abordar el tratamiento de datos faltantes y la posible transformación de algunas variables categóricas para mejorar el desempeño de los modelos predictivos. También es importante señalar que algunas de las columnas tienen una alta cantidad de valores nulos, lo que podría requerir una limpieza y transformación detallada antes de proceder con el modelado predictivo.

**Preguntas de Investigación:**

1. ¿Cómo influye el año de fabricación de un auto en su precio de venta?
2. ¿El tipo de combustible (gasolina, diésel, eléctrico, etc.) afecta significativamente el precio de los autos usados?
3. ¿Cómo impacta el kilometraje (odómetro) de un auto en su precio de venta?

**Hipótesis:**

1. El año de fabricación tiene una relación positiva con el precio de venta. Los autos más nuevos tienen precios más altos debido a la demanda y al menor desgaste. Los autos de modelos más antiguos tienen precios más bajos debido al deterioro y la menor demanda.
2. Los autos con motor eléctrico o híbrido tienen un precio de venta más alto en comparación con los autos de gasolina o diésel debido a la creciente demanda de vehículos ecológicos y las políticas de incentivos gubernamentales.
3. Un mayor kilometraje está negativamente correlacionado con el precio de venta. Los autos con mayor kilometraje tienden a tener un precio más bajo debido a la percepción de mayor desgaste y menor vida útil.

### Contexto Analítico y Problema Comercial

### **Desde una perspectiva comercial, este tipo de análisis tiene implicaciones importantes para empresas en el mercado de autos usados, como concesionarios, plataformas de compra-venta de autos, y otros actores en la industria automotriz. Algunas de las preguntas y problemas comerciales que podrían abordarse son:**

### **Determinación de Precios Óptimos: Para un concesionario o plataforma de venta de autos usados, tener la capacidad de predecir de manera precisa el precio de los vehículos les permitiría establecer precios más competitivos y alineados con las expectativas del mercado. Esto optimiza tanto las ventas como la rentabilidad, al evitar la subvaloración o sobrevaloración de los vehículos.**

### **Optimización del Inventario: Un análisis del precio en función de las características del auto permitiría a las empresas identificar qué tipos de vehículos tienen una mayor demanda en ciertas regiones o qué características están relacionadas con precios más altos o más bajos. Esto podría ayudar a los concesionarios a tomar decisiones sobre qué autos comprar o mantener en inventario.**

### **Estimación de Valor de Reventa: Los usuarios o concesionarios que planean vender vehículos usados pueden beneficiarse de estimaciones de precios precisas para obtener una mejor valoración de su vehículo al momento de venderlo. Esto también es relevante para plataformas de reventa que permiten a los usuarios obtener estimaciones rápidas de su vehículo.**

### **Tendencias del Mercado: La predicción del precio también podría identificar tendencias o patrones emergentes en el mercado, como el impacto de ciertos factores (por ejemplo, el tipo de combustible o el año de fabricación) en la valoración de los vehículos. Esto es valioso para la toma de decisiones estratégicas en las operaciones comerciales.**

### 

### *Storytelling*

#### Paso 1: Lectura de datos

Comenzamos con el **data wrangling**, un proceso crucial que aseguraría que los datos estuvieran en el mejor estado posible para los modelos predictivos.

A screen shot of a computer

Description automatically generated

#### A screenshot of a computer Description automatically generated

**Descripción de variables:**

id:

* Tipo: int64
* Descripción: Un identificador único para cada entrada en el dataset.

url:

* Tipo: object
* Descripción: URL del anuncio del vehículo en la plataforma donde se publicó.

region:

* Tipo: object
* Descripción: Región geográfica donde el vehículo está listado.

region\_url:

* Tipo: object
* Descripción: URL específica de la región o localización del anuncio.

price:

* Tipo: int64
* Descripción: Precio del vehículo en la moneda local.

year:

* Tipo: float64
* Descripción: Año de fabricación del vehículo.

manufacturer:

* Tipo: object
* Descripción: Fabricante o marca del vehículo, como Toyota, Ford, etc.

model:

* Tipo: object
* Descripción: Modelo del vehículo, como Corolla, F-150, etc.

condition:

* Tipo: object
* Descripción: Condición del vehículo (ej. nuevo, usado, como nuevo).

cylinders:

* Tipo: object
* Descripción: Número de cilindros del motor del vehículo (ej. 4, 6, 8).

fuel:

* Tipo: object
* Descripción: Tipo de combustible que utiliza el vehículo (ej. gasolina, diésel, eléctrico).

odometer:

* Tipo: float64
* Descripción: Kilometraje del vehículo (distancia recorrida).

title\_status:

* Tipo: object
* Descripción: Estado del título del vehículo (ej. limpio, reconstruido, etc.).

transmission:

* Tipo: object
* Descripción: Tipo de transmisión del vehículo (ej. automática, manual).

VIN:

* Tipo: object
* Descripción: Número de identificación del vehículo (VIN por sus siglas en inglés).

drive:

* Tipo: object
* Descripción: Tipo de tracción del vehículo (ej. tracción delantera, tracción trasera, tracción en las 4 ruedas).

size:

* Tipo: object
* Descripción: Tamaño del vehículo (ej. compacto, mediano, SUV).

type:

* Tipo: object
* Descripción: Tipo de vehículo (ej. sedán, camioneta, SUV).

paint\_color:

* Tipo: object
* Descripción: Color de la pintura del vehículo.

image\_url:

* Tipo: object
* Descripción: URL de la imagen del vehículo en el anuncio.

description:

* Tipo: object
* Descripción: Descripción textual del vehículo en el anuncio, proporcionada por el vendedor.

county:

* Tipo: float64
* Descripción: Este campo parece no contener datos.

state:

* Tipo: object
* Descripción: Estado en el que se encuentra el vehículo listado (en Estados Unidos, esto podría ser el nombre del estado como California, Texas, etc.).

lat:

* Tipo: float64
* Descripción: Latitud de la ubicación geográfica donde está listado el vehículo.

long:

* Tipo: float64
* Descripción: Longitud de la ubicación geográfica donde está listado el vehículo.

posting\_date:

* Tipo: object
* Descripción: Fecha en la que el anuncio del vehículo fue publicado en la plataforma

#### Paso 2: Data Wrangling - Limpieza y Preparación de los Datos

**Tratamiento de Valores Faltantes:** Revisamos los datos faltantes, eliminamos las columnas sin datos y rellenamos valores con la moda, media y mediana.

A screenshot of a computer

Description automatically generated

**A screenshot of a computer program

Description automatically generated**

**Revisión de Outliers:** Otro paso importante en la fase de wrangling fue la detección de outliers (valores atípicos) en las variables numéricas como price y odometer procedimos a acomodarlos, a fin de lograr una mejor calidad de datos.

A screen shot of a computer program

Description automatically generated

#### Paso 3: Análisis Univariado, Bivariado y Multivariado

Una vez que el conjunto de datos estaba limpio, era importante realizar un análisis exploratorio para entender mejor las relaciones entre las diferentes características.

**Análisis Univariado:** El análisis univariado nos permitió estudiar cada variable por separado. Utilizamos gráficos como histogramas y boxplots para visualizar la distribución de las variables individuales, como la edad, altura y peso. Esto nos ayudó a detectar la dispersión de los datos, identificar posibles sesgos y detectar si las variables seguían una distribución normal.

A graph with red lines

Description automatically generatedA graph of a graph of a car

Description automatically generated with medium confidence

A graph of a graph showing a number of vehicles

Description automatically generated with medium confidenceA green rectangular box with numbers

Description automatically generated with medium confidence

**Análisis Bivariado:** en este análisis realizamos un un diagrama de dispersión para ver cómo varían los precios con el año de fabricación, también usamos un boxplot para ver cómo varía el precio según la condición del vehículo.

A blue and white dotted chart

Description automatically generatedA chart of different colored squares

Description automatically generated

**Análisis Multivariado:** Finalmente, tomamos dos casos: Primero la relación entre el precio, el año y el kilometraje (donde se usaron gráficos de heatmap, pairplot y gráfico de dispersión 3D). En segundo lugar, se utilizó un gráfico de violín para ver la distribución de precios por tipo de vehículos.

A graph of a number of blue dots

Description automatically generated with medium confidenceA graph of a colorful cube

Description automatically generated with medium confidence

A graph of different shapes

Description automatically generated with medium confidence

**Conclusiónes**

**1.Distribución de la variable objetivo (Precio)**

Precio de los vehículos: La variable price sigue una distribución sesgada hacia la derecha, con muchos vehículos a precios más bajos y unos pocos a precios altos. Esto es común en mercados de vehículos usados, donde la mayoría de los vehículos son más baratos, pero algunos vehículos de lujo o nuevos pueden tener precios elevados.

**2. Valores atípicos (Outliers)**

Precios atípicos: En el caso de que haya vehículos con precios muy bajos o extremadamente altos, estos podrían ser errores en los datos o vehículos excepcionales (por ejemplo, vehículos de lujo o clásicos).

**3. Relación entre variables numéricas**

Precio vs. Año: Es probable que veas que los vehículos más nuevos tienden a tener precios más altos, aunque esto no es una regla estricta debido a factores como la marca, el estado y la demanda del modelo.

Precio vs. Kilometraje: Los vehículos con menor kilometraje generalmente tienen precios más altos, lo cual tiene sentido porque los vehículos más nuevos y menos usados suelen ser más caros.

Año vs. Kilometraje: En general, los vehículos más nuevos tienden a tener menos kilometraje, aunque puede haber excepciones (por ejemplo, un vehículo usado mucho en viajes largos).

**4. Relación entre variables categóricas y numéricas**

Precio vs. Condición del vehículo: Los vehículos en mejores condiciones suelen tener precios más altos. Si hay categorías como "Excelente", "Bueno", "Regular" o "Para piezas", probablemente verás que los vehículos en la mejor condición tienden a ser más caros.

### Precio vs. Tipo de combustible: Dependiendo del tipo de combustible, podrías encontrar que los vehículos eléctricos o híbridos tienen precios más altos que los de gasolina o diésel, lo cual refleja las tendencias de mercado actuales.

### Precio vs. Fabricante: Algunas marcas tienen precios más altos en promedio que otras. Por ejemplo, marcas de lujo o premium tienden a tener precios más altos que marcas más asequibles.

### 5. Datos faltantes

### Valores nulos: Varias columnas como condition, cylinders, manufacturer, model, etc., contienen valores nulos. Esto puede ser una señal de que los datos fueron incompletos al momento de ser cargados o recopilados.

### Columnas sin valores: La columna county no tiene ningún dato, por lo que fue eliminada sin afectar el análisis.

### 6. Posibles áreas de mejora en la limpieza de datos

### Columnas irrelevantes: Columnas como url, region\_url o image\_url pueden no ser necesarias para análisis posteriores, ya que no aportan información significativa al modelo predictivo.

### Errores en las variables: Si se encuentran valores incorrectos o fuera de rango (por ejemplo, un precio de 0 o un kilometraje muy alto), estas filas fueron tratadas.

### Los precios están fuertemente influenciados por el año de fabricación, el kilometraje y la condición del vehículo.

### Los vehículos más nuevos y con menor kilometraje tienen precios más altos.

### El tipo de combustible, la marca y la condición del vehículo también son factores importantes que afectan el precio.

### Las variables categóricas, como la marca del fabricante o el tipo de combustible, tienen un impacto significativo en el precio.

### El EDA es una etapa crucial para comprender los patrones y relaciones en los datos. Estas conclusiones servirán como base para un modelado predictivo.

### Modelado Predictivo

**Preprocesamiento de datos**

#### Hacemos una inspección y limpieza de datos:

#### Convertimos la columna posting\_date a datetime para facilitar la manipulación

#### Nos fijamos que las columnas sean el dato adecuado para facilitar en análisis

#### **3.** Eliminamos valores outliers fuera del rango

A screenshot of a computer program

Description automatically generated

**Feature Selection**

Usamos la técnica de Selección univariante de características con la función SelectKBest de scikit-learn para obtener las características más relevantes.

A screen shot of a computer program

Description automatically generated

**Modelado**

Entrenamos los modelos y los gráficamos, en este caso usaremos 3, regresión líneal, random forest y XGBoost.

A graph with blue dots and a red line

Description automatically generatedA graph with green dots and red line

Description automatically generated

A graph with orange dots

Description automatically generated

**Conclusión**

Después de entrenar tres modelos de regresión (Regresión Lineal, Random Forest y XGBoost) para predecir el precio de los vehículos, hemos evaluado su desempeño utilizando tres métricas clave: Error Absoluto Medio (MAE), Error Cuadrático Medio (MSE) y Coeficiente de Determinación (R²).

1. Regresión Lineal MAE: 0.67 MSE: 0.79 R²: 0.21 La Regresión Lineal presenta un rendimiento relativamente bajo con un R² de 0.21, lo que indica que solo el 21% de la variabilidad en el precio puede ser explicada por las características del modelo. Aunque tiene un error absoluto medio (MAE) de 0.67 y un MSE de 0.79, estos valores sugieren que el modelo no ha capturado bien la relación entre las características del vehículo y su precio.
2. Random Forest MAE: 0.45 MSE: 0.41 R²: 0.59 El Random Forest ha mostrado un desempeño significativamente mejor que la regresión lineal, con un R² de 0.59. Esto indica que aproximadamente el 59% de la variabilidad en el precio de los vehículos puede ser explicada por las características. Además, el MAE de 0.45 y el MSE de 0.41 son considerablemente más bajos, lo que sugiere que este modelo es más preciso y robusto en comparación con la regresión lineal.
3. XGBoost MAE: 0.42 MSE: 0.36 R²: 0.64 El modelo XGBoost ha demostrado el mejor rendimiento de todos los modelos evaluados, con el R² más alto (0.64), lo que indica que es capaz de explicar el 64% de la variabilidad en los precios de los vehículos. El MAE de 0.42 y el MSE de 0.36 son los más bajos, lo que sugiere que XGBoost tiene un excelente ajuste para los datos y genera predicciones más precisas.

Comparación de Modelos Rendimiento: XGBoost ha sido el modelo más preciso, seguido por Random Forest y luego Regresión Lineal. XGBoost presenta un R² de 0.64, significativamente más alto que los otros modelos, lo que lo convierte en el más efectivo para este problema de predicción de precios. Precisión de predicción: El MAE y el MSE más bajos de XGBoost (0.42 y 0.36, respectivamente) sugieren que este modelo es el que ofrece las predicciones más cercanas a los valores reales. Flexibilidad: Random Forest y XGBoost son modelos más complejos y no lineales que la Regresión Lineal, lo que explica su mejor rendimiento. Ambos modelos pueden manejar interacciones no lineales y relaciones complejas entre las variables, lo que los hace más adecuados para este tipo de problemas. Recomendación Basado en los resultados, el modelo XGBoost es claramente el mejor para este problema de predicción de precios de vehículos, ya que ofrece el mejor rendimiento en términos de precisión y capacidad explicativa. Random Forest también ofrece buenos resultados y podría ser una opción viable si se busca una alternativa menos compleja. Por otro lado, la Regresión Lineal no es la mejor opción debido a su rendimiento relativamente bajo en este caso.

**Optimización de modelo:**

En este caso quisimos optimizar el random forest con hiperparámetros.

A graph with blue lines

Description automatically generated

**Conclusión:**Eficiencia del modelo Random Forest: Los resultados obtenidos, tanto antes como después de la optimización de hiperparámetros, indican que el modelo Random Forest ya está bastante bien ajustado para el problema, ya que el MAE y el MSE son bastante bajos y el R² es moderado. Un R² de 0.59 significa que aproximadamente el 59% de la variabilidad en el precio de los autos se explica por las variables del modelo.

La optimización de hiperparámetros en el modelo Random Forest ha mostrado que el modelo ya está funcionando bastante bien, pero no ha generado una mejora significativa en las métricas. Esto sugiere que el modelo está cerca de su mejor rendimiento con los parámetros actuales. Para continuar mejorando, su puede probar otras técnicas de ingeniería de características, probar diferentes modelos o revisar la calidad y cantidad de los datos.