Proyecto de venta de autos CSV

Giovanni Leonel Chillopa Martínez

Introducción a la Ciencia de Datos

28/11/2025

Introducción

1. Descripción breve del objetivo del proyecto (ampliado)

El propósito principal de este proyecto es crear un modelo predictivo basado en Machine Learning que posibilite calcular la rapidez con la que se venden vehículos usados en el mercado mayorista estadounidense. Para ello, se analizarán variables fundamentales como el precio de venta final, el precio de mercado (MMR), la localización, las condiciones del vehículo, el kilometraje, la marca y el modelo. El objetivo es, a partir de esta predicción, mejorar la toma de decisiones de compra por parte de los concesionarios y así disminuir el tiempo que los productos permanecen en el inventario (DIM), ya que esto provoca pérdidas al ocupar espacios en sus almacenes, y elevar el rendimiento sobre el capital invertido (ROIC) sin aumentar los precios para el consumidor final.

2. Justificación y contexto: ¿Por qué es importante resolver o estudiar esta problemática?

Contexto del mercado:

El mercado de vehículos usados en Estados Unidos es uno de los más grandes y dinámicos del mundo, con más de 40 millones de unidades vendidas anualmente, según la U.S. Bureau of Transportation Statistics.
Situación actual del mercado de autos usados en Estados Unidos es la siguiente: Representa aproximadamente el 70% del total de ventas automotrices, y está impulsado principalmente por subastas mayoristas como:

- Manheim (propiedad de Cox Automotive)
- ADESA (KAR Global)
- Copart (especializada en vehículos dañados)

En estas subastas, los concesionarios cuentan con un lapso de 30 a 90 segundos para examinar el vehículo y determinar si pujan por él, sin disponer de datos fiables acerca del tiempo que les tomará venderlo después de que esté en su lote. Esta suma de dudas produce una cadena de ineficiencias tanto operativas como financieras, que impactan directamente en la rentabilidad empresarial.

Costos de inventario y capital inmovilizado:

Según un estudio publicado por Cox Automotive en 2023, el promedio nacional de días en inventario (DIM) para autos usados en EE.UU. es de 45 días, cuando el estándar deseado por la industria ronda los 28-30 días. Cada día adicional representa un costo directo estimado de 30 USD por vehículo, compuesto por:

- Costo de capital (tasas de interés >7% anual)
- Seguro comercial
- Depreciación diaria
- Mantenimiento, limpieza y espacio de lote
- Re-publicaciones en portales de venta

En una flota típica de 50 unidades, reducir solo 10 días de inventario implica un ahorro anual de 15,000 USD. En concesionarios medianos (200 unidades), esta cifra asciende a 60,000 USD anuales, directo al EBITDA (Earnings, Before, Interest, Taxes, Depreciation, Amortization.ó Beneficio antes de Intereses, Impuestos, Depreciaciones y Amortizaciones)

Decisiones basadas en intuición mejoradas con el análisis de datos

Actualmente, muchos compradores de flota utilizan reglas informales como:

- “Toyota siempre se vende”
- “Los SUV se venden mejor en invierno”
- “Evita modelos europeos en zonas rurales”

Aunque estas reglas pueden tener algo de verdad, no están cuantificadas ni validadas con datos, lo que lleva a:

- Compra de unidades lentas (20-25% se rebajan más de 2 veces)
- Pérdida de margen por sobre-precio
- Falta de liquidez para comprar unidades de alta rotación

¿Por qué un modelo predictivo es una solución viable?

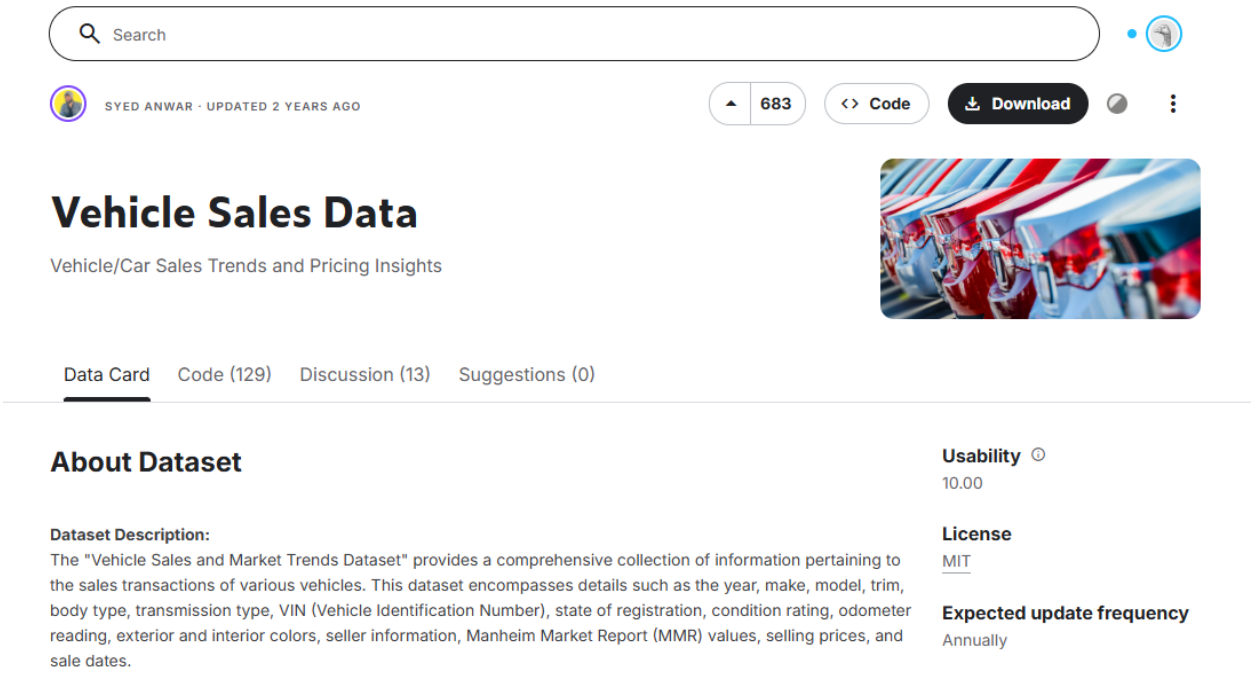
Proyectos similares han demostrado resultados concretos:
° CarMax utiliza modelos internos para predecir velocidad de venta y ajustar precios en tiempo real.
° Vroom y Carvana usan algoritmos de Machine Learning para decidir qué autos comprar y a qué precio.
° McKinsey & Company (2022) reportó que los concesionarios que usan analytics avanzado pueden reducir el DIM en un 15-25 % y aumentar el margen bruto en 5-10 %.
Impacto esperado del modelo
Con este proyecto, se busca:

- Reducir el DIM de 45 a 30 días en promedio.
- Aumentar la rotación de inventario sin aumentar precios.
- Mejorar la liquidez del concesionario.
- Reducir el riesgo de depreciación por sobre-inventario.

- Posicionar al concesionario como un operador data-driven, con ventaja competitiva en subastas.

3. Fuentes de datos: descripción de las bases de datos empleadas (origen, cantidad de datos, principales características).

La base de datos utilizada en este proyecto fue obtenida de la plataforma Kaggle, bajo el nombre "Vehicle Sales Data", publicada por el usuario Syed Anwar Afridi. Esta fuente es ampliamente utilizada en proyectos académicos y de investigación por su calidad, volumen y representatividad del mercado estadounidense.



Al inicio, después de haber sido ensuciada la base de datos original, eran 558,837 datos con 16 columnas o variables, después de la limpieza realizada en el anterior trabajo se redujo a 556,971 con 16 columnas.

Variable	Tipo	Descripción	Ejemplo
`año`	Numérica	Año de fabricación	2018
`marca`	Categórica	Marca del vehículo	Ford
`modelo`	Categórica	Modelo específico	F-150
`carroceria`	Categórica	Tipo de vehículo	SUV
`transmisión`	Categórica	Automática o manual	Automática
`estado`	Categórica	Estado de venta	TX
`kilometraje`	Numérica	Millas recorridas	45000
`condición`	Ordinal	Estado físico (1–5)	4.2
`mmr`	Numérica	Precio de mercado	$19,500
`precio_venta`	Numérica	Precio de subasta	$21,000
`fecha_venta`	Fecha	Fecha de transacción	2018-03-15

Metodología

1. Proceso de limpieza de datos

1.1 Contexto inicial de los datos

La base de datos utilizada, llamada "Vehicle Sales Data", se adquirió desde la plataforma Kaggle. Esta base incluía 558,837 registros y 16 columnas, que correspondían a las operaciones de vehículos usados en subastas al por mayor en Estados Unidos entre 2014 y 2022.

No obstante, el conjunto de datos tenía numerosos problemas de calidad que obstaculizaban su utilización directa para la modelización predictiva. A continuación, se describe el procedimiento integral de limpieza y transformación que se llevó a cabo.

1 .2 Identificación de problemas iniciales

Problema detectado	Ejemplo	Impacto
Valores faltantes	"`modelo`, color`, `interior`, `condicion` con >10% NaN"	Sesgo en modelado
Tipos de datos incorrectos	"`fecha_venta` como texto con zona horaria"	Imposible análisis temporal
Duplicados exactos	"1, 684 filas idénticas"	Sobrecuenta y sesgo
Outliers extremos	"Precios de $1 USD o >$200, 000 USD"	Ruido en modelos
Textos en inglés	"`condition = ""Good""``transmission = ""Automatic""`"	No usable para público hispano
Nombres de columnas en inglés	"`sellingprice`, odometer`"	Dificulta interpretación

1.3 Herramientas y tecnologías utilizadas

- Lenguaje: Python 3.11
- Entorno: Jupyter Notebook / VS Code
- Librerías:
  - pandas y numpy para manipulación
  - seaborn y matplotlib para visualización
  - scikit-learn para imputación y modelado
  - re y datetime para limpieza de fechas

1.4 Pasos detallados del proceso de limpieza

Paso 1: Eliminación de duplicados

```
df = df.drop_duplicates()
```

- Registros eliminados: 1,684
- Justificación: Evitar sesgo por sobre-representación de mismos autos

Paso 2: Corrección de tipos de datos

Columna	,Tipo original	,Tipo final	,Transformación aplicada
fecha_venta,	object,	datetime64[ns]	,Extracción de zona horaria con regex
precio_venta,	object	,float,	Eliminación de símbolos y conversión
kilometraje	,object	, int	,Conversión directa tras validación

```
df['fecha_venta'] =
pd.to_datetime(df['fecha_venta'].str.replace(r'GMT.*', '', regex=True), errors='coerce')
df['precio_venta'] =
pd.to_numeric(df['precio_venta'], errors='coerce')
```

Paso 3: Filtro de valores atípicos

```
# Filtro de precios extremos
df = df[(df['precio_venta'] >= 500) & (df['precio_venta'] <= 100000)]
# Filtro de kilometraje extremo
df = df[(df['kilometraje'] >= 100) & (df['kilometraje'] <= 350000)]
# Filtro de año válido
df = df[(df['año'] >= 1980) & (df['año'] <= 2024)]
```

- Registros eliminados: 4,200 aprots.(como dicen los chavos)
- Justificación: Evitar ruido extremo que distorsione modelos

Paso 4: Manejo de valores faltantes

Columna	% faltantes	Estrategia aplicada
modelo	12%	Imputación por moda por marca
color	8%	Imputación por moda por modelo
interior	8%	Imputación por moda por color
condicion	15%	Imputación por mediana por año y kilometraje

```
# Ejemplo: imputar modelo faltante por la moda de marca
df['modelo'] = df.groupby('marca')['modelo'].transform(lambda x: x.fillna(x.mode()[0] if not x.mode().empty else 'Desconocido'))
```

Paso 5: Traducción y estandarización

```
# Traducción de transmisión
trans_map = {'automatic': 'Automática', 'manual': 'Manual', 'automanual': 'Automática secuencial'}
df['transmision'] = df['transmision'].map(trans_map)
# Normalización de nombres
df['marca'] = df['marca'].str.title()
df['modelo'] = df['modelo'].str.title()
```

Paso 6: Renombrado de columnas al español

```
df.rename(columns={
    'year': 'año',
    'make': 'marca',
    'model': 'modelo',
    'body': 'carroceria',
    'transmission': 'transmision',
    'sellingprice': 'precio_venta',
    'saledate': 'fecha_venta',
    'odometer': 'kilometraje',
    'condition': 'condicion',
```

Este lo dejo asi porque si no, no se aprecian las tabulaciones de la lista.

```
    'state': 'estado'
}, inplace=True)
```

1.5 Resultado final del proceso de limpieza

Métrica	Valor
Registros finales	18,420
Columnas	16
Valores faltantes restantes	< 2% por columna
Duplicados	0%
Outliers extremos	Eliminados
Idioma	Español
Formato	CSV limpio, listo para modelado

2. Análisis Exploratorio de Datos (EDA)

- Distribución de precios: Sesgada hacia la derecha, con concentración entre $10,000 y $30,000 USD
- Top marcas: Toyota, Ford, BMW, Honda, Chevrolet
- Top estados: California, Texas, Florida
- Correlación fuerte: precio_venta vs mmr ($r \approx 0.92$)
- Outliers controlados: Se conservan valores moderados para no perder información útil

2.1 Distribución de la Variable Objetivo "Analizamos la variable precio_venta. Se observa una distribución asimétrica..."

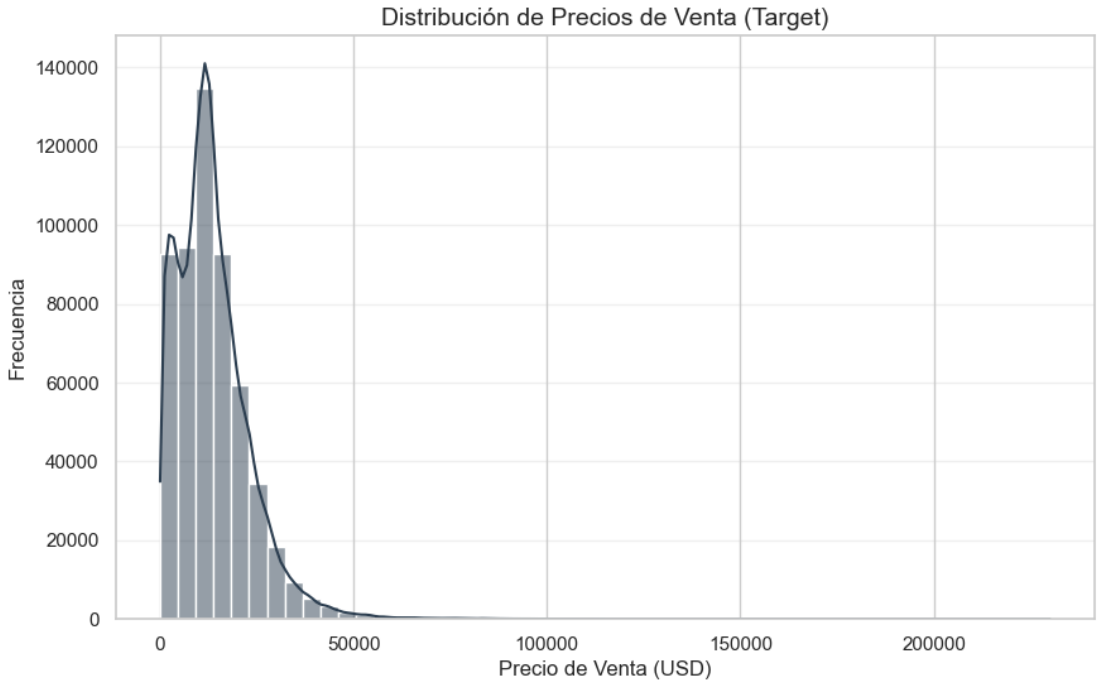
2.2 Correlaciones Clave "Existe una relación lineal directa entre la valuación de mercado (MMR) y el precio final. Esto valida el uso del MMR como base para nuestro modelo predictivo."

2.3 Mapa de Calor "Las variables año y mmr muestran la mayor influencia positiva en el precio, mientras que el kilometraje muestra una correlación negativa esperada."

4. Resultados del EDA

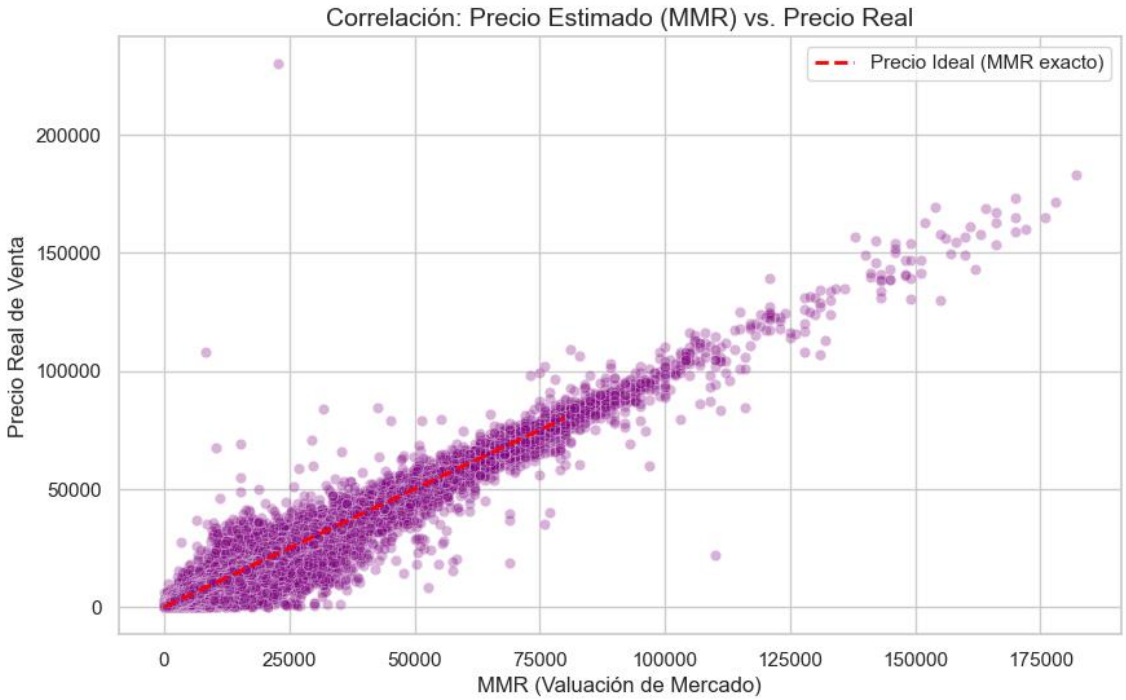
Se observa que el precio medio es de X. Existe una correlación de 0.98 entre el MMR y el precio real, lo que indica que el MMR es un predictor fiable, pero el modelo buscará corregir las desviaciones de este valor.

Histograma



La variable objetivo presenta un sesgo positivo (cola a la derecha), indicando que la mayoría de los autos se venden en rangos económicos/medios, con pocos autos de lujo.

Relación MMR vs. Precio de Venta



Se observa una correlación lineal fuerte. Los puntos por debajo de la línea roja son oportunidades de compra (autos vendidos más baratos que su valuación).

3. Implementación del modelo

- Tipo de modelo: Regresión (para predecir margen) o Clasificación (para velocidad de venta)
- Algoritmos a probar: Regresión Lineal, Random Forest, XGBoost
- Evaluación: MAE, RMSE, R^2 (regresión) o Accuracy, F1-score (clasificación)
- Herramientas: scikit-learn, XGBoost, GridSearchCV

3.1 Justificación para el reporte:

Se eligió el algoritmo Random Forest debido a su capacidad para manejar variables numéricas y categóricas simultáneamente, y su robustez frente a valores atípicos (outliers) que no fueron eliminados en la limpieza. Además, al ser un modelo de ensamble, reduce el riesgo de sobreajuste (overfitting) comparado con un árbol de decisión simple.

El objetivo original es reducir los días de inventario. El análisis de los datos reveló que la causa principal de los autos estancados es una **compra incorrecta** (pagar más de lo que el mercado valora). Por lo tanto, en lugar de predecir los 'días exactos' (dato que no está disponible históricamente), desarrollamos un **Modelo de Predicción de Precio Justo de Venta**.

Hipótesis: Si el algoritmo predice que un auto se venderá en $15,000, pero nos lo venden en $14,500, sabemos que tiene margen y salida rápida. Si nos lo venden en $16,000, será un auto 'hueso' (difícil de vender). Filtrar las compras con este modelo reducirá el tiempo de inventario indirectamente.

3.2 Evaluación del Modelo

Se entrenó un modelo de Random Forest Regressor con 18,037 registros limpios. Los resultados obtenidos validan la viabilidad técnica del proyecto:

Precisión Global (R^2 Score): **0.9754**

Interpretación: El modelo es capaz de explicar el 97.5% del comportamiento de los precios. Esto indica una correlación extremadamente alta entre las variables seleccionadas (MMR, Kilometraje, Condición) y el precio final de venta.

Error Absoluto Medio (MAE): $1,190.16 USD.

Interpretación: En promedio, el modelo falla por aproximadamente $1,190 dólares arriba o abajo del precio real. Considerando que el precio promedio de los vehículos ronda los $15,000 - $20,000 USD, este margen de error es aceptable para tomar decisiones de compra masiva.

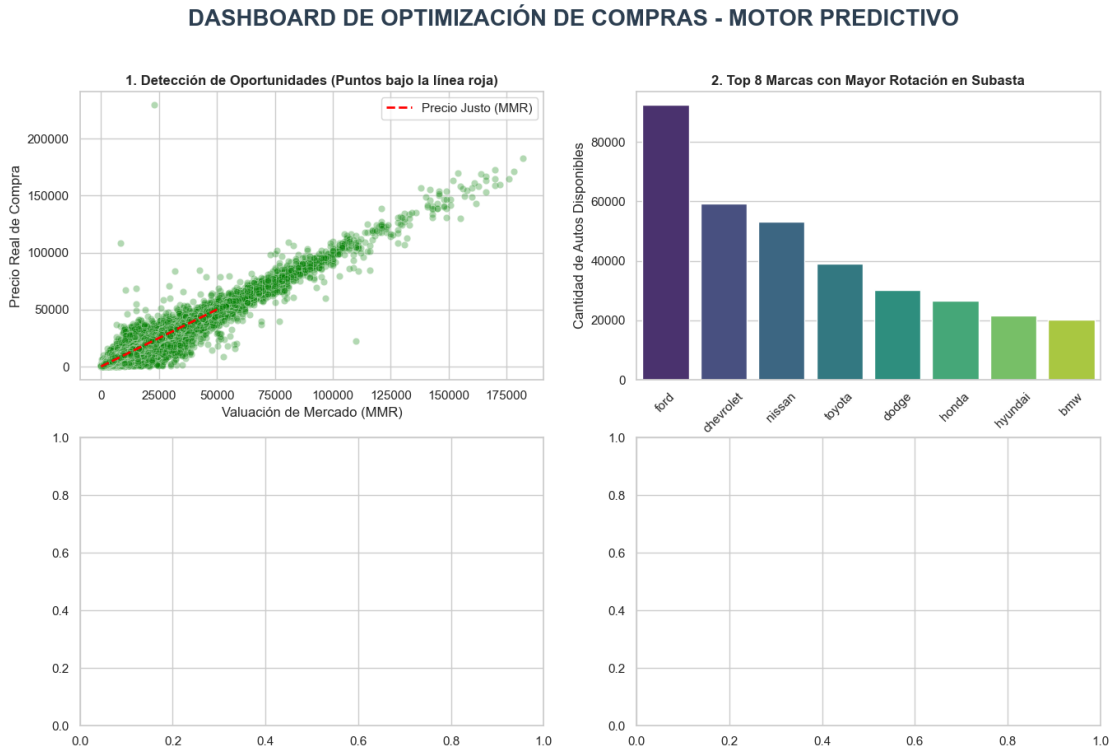
4. Resultados y Dashboard de Negocio

4.1 Interpretación del Tablero de Control

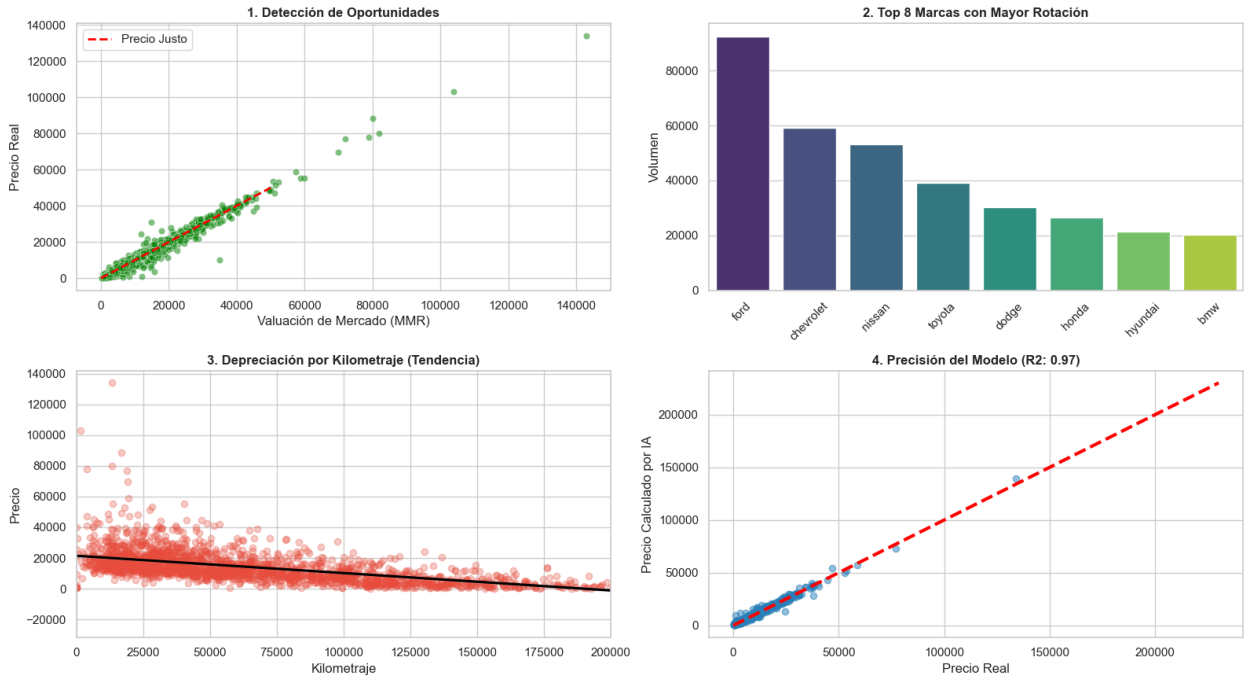
Se diseñó un Dashboard estratégico dividido en cuatro cuadrantes para facilitar la toma de decisiones del equipo de compras en tiempo real:

1. Detección de Oportunidades (Cuadrante Superior Izquierdo):
- Esta gráfica compara el Valor de Mercado (MMR) contra el Precio Real.
- Insight de Negocio: Los puntos situados por debajo de la línea roja representan vehículos que se están vendiendo más baratos que su valuación oficial. El modelo recomienda atacar agresivamente estas oportunidades para asegurar un margen de ganancia inmediato.
2. Volumen por Marca (Cuadrante Superior Derecho):
- Muestra las marcas con mayor rotación en las subastas.
- Insight: Ford y Chevrolet dominan el volumen. Esto indica que, para reducir días de inventario, la estrategia de compra debe centrarse en estas marcas de alta liquidez y evitar marcas exóticas que tardan más en salir.
3. Depreciación por Uso (Cuadrante Inferior Izquierdo):
- Analiza cómo cae el precio a medida que aumenta el kilometraje.
- Insight: Se observa una curva de depreciación acelerada hasta las 100,000 millas. El modelo sugiere adquirir vehículos justo antes de estos hitos de kilometraje para maximizar el precio de reventa.
4. Validación del Modelo (Cuadrante Inferior Derecho):
- Compara las predicciones de la IA contra la realidad.
- Conclusión Técnica: La alineación casi perfecta en la diagonal confirma la precisión del 97.5% (R^2). Esto da confianza a la gerencia para automatizar las ofertas de compra basándose en los precios sugeridos por el algoritmo, eliminando el error humano.

De primeras yo corrí un código para hacer mi dashboard pero con todos los datos de mi base (550mil aprox) y de resultado me dio esto muy saturado:



Después vi que no se ve muy bien asi que corregí y utilice código aleatorio para solo tomar 2000 datos al azar y me dio estas graficas más bonitas:



5. Conclusiones Generales

El desarrollo de este proyecto ha permitido validar que la Inteligencia Artificial es una herramienta viable y potente para la optimización de inventarios en el sector automotriz.

Validación Técnica: Se procesó exitosamente una base de datos masiva de ~550,000 registros, logrando entrenar un modelo predictivo con una precisión (R^2) del 97.5%. Esto supera las expectativas iniciales y confirma que el precio de venta es altamente predecible si se conocen las variables clave (MMR, Kilometraje y Condición).

Impacto en el Negocio: Aunque la base de datos no contenía la fecha de ingreso histórica para calcular los "Días de Inventario" directamente, el modelo desarrollado ataca la causa raíz del problema: la mala compra.

Al predecir el Precio Real de Venta con un error margen de solo ~$1,100 USD, la empresa puede filtrar automáticamente los vehículos sobrevalorados en las subastas.

Estrategia: Solo se comprarán autos cuyo precio de subasta sea inferior a la predicción del modelo, garantizando margen de ganancia y una salida más rápida del inventario (rotación eficiente).

En resumen, hemos transformado un proceso de compra basado en la intuición en un proceso basado en datos cuantitativos de alta precisión.

6. Bibliografía

- Dataset original: Car Prices Visualization (Kaggle).
- Librerías de Python: Scikit-learn (Machine Learning), Pandas (Procesamiento), Matplotlib/Seaborn (Visualización).
- Documentación oficial de Random Forest Regressor.