Data Science II: Machine Learning para la Ciencia de Datos

Módulo II

Maximiliano Domínguez - Comisión 61605 - Profesor: Jorge Ruíz

Inicio

El proyecto se basó en un dataset que contiene información de ventas de autos de todo tipo y marcas variadas desde los más caros del mercado hasta los más accesibles, cuenta con información sobre la venta, modelo, color, estado, etc.

Contexto Comercial

Actualmente se está trabajando en la solicitud de una automotora que cuenta con una plataforma donde postea sus autos usados y en base a datos históricos desea mejorar la fijación de precios de los nuevos autos que ingresen.

Obtener un buen resultado en la fijación de precios los ayudará tanto en la rentabilidad como en la rotación de vehículos.

Problema Comercial

La automotora en cuestión está teniendo ciertos problemas actualmente con la rotación (venta) de sus vehículos a través de la plataforma y consideran que el problema sea la precisión de precios al estimar el valor de los autos que ingresan.

Esto desengloba en dos problemas:

- Precios Subestimados: Menos ganancia porque los autos se venden por menos de su valor real.

- Precios Sobreestimados: Los autos no se venden rápidamente, aumentando los costos de almacenamiento y disminuyendo la rotación del inventario.

Descripción de columnas del dataset

- year: Año del modelo del auto (int64) [1982 al 2015]
- make: Marca del auto (object) Ejemplos: ['Aston Martin', 'Audi', 'BMW', 'Cadillac', 'Chevrolet', 'Dodge', 'FIAT','Ferrari','Ford', 'Honda']
- model: Modelo del auto (object) Ejemplos: ['tucson', 'tt', 'thunderbird','taurus', 'tahoe', 'tC', 'sx4', 'swift', 'subrbn','sprinter', 'sportage', 'sonoma']
- trim: Versión del auto (object) Ejemplos: ['Wagon XLT','Wagon Titanium LWB', 'Wagon', 'WS', 'WRX TR', 'WRX STi','WRX STI Limited', 'WRX STI', 'WRX Limited']
- body: Tipo de carrocería (object) Ejemplos: ['wagon', 'van', 'tsx sport wagon', 'transit van', 'suv','supercrew', 'supercab', 'sedan','quad cab', 'q60 coupe']
- transmission: Tipo de transmisión (object) Ejemplos: ['manual', 'automatic']
- vin: Número de identificación del vehículo (object)
- state: Estado (object) Ejemplos: ['ny', 'nv', 'ns', 'nm', 'nj', 'ne', 'nc', 'ms', 'ma', 'la', 'in', 'il', 'hi', 'ga', 'fl','co', 'ca']
- condition: Condición del auto (float64) [1 al 50]
- odometer: Kilometraje del auto (float64)
- color: Color exterior (object) Ejemplos: ['yellow', 'white', 'turquoise', 'silver', 'red', 'lime', 'green', 'gray', 'gold','charcoal','blue', 'black']
- interior: Color interior (object) Ejemplos: ['yellow', 'white', 'tan', 'silver', 'red', 'purple', 'orange','burgundy','black', 'beige']
- seller: Vendedor (object) Ejemplos: ['zygi auto corp', 'zumbrota ford sales llc', 'zuma autoboat','101motors', '1 for all auto sales', '1 cochran of monroeville']
- mmr: Precio promedio de mercado (float64) Ejemplo: 21500.0
- sellingprice: Precio de venta (float64) Ejemplo: 21500.0
- saledate: Fecha de venta (object) Ejemplo: [Wed May 27 2015 17:00:00 GMT-0700 (PDT)]

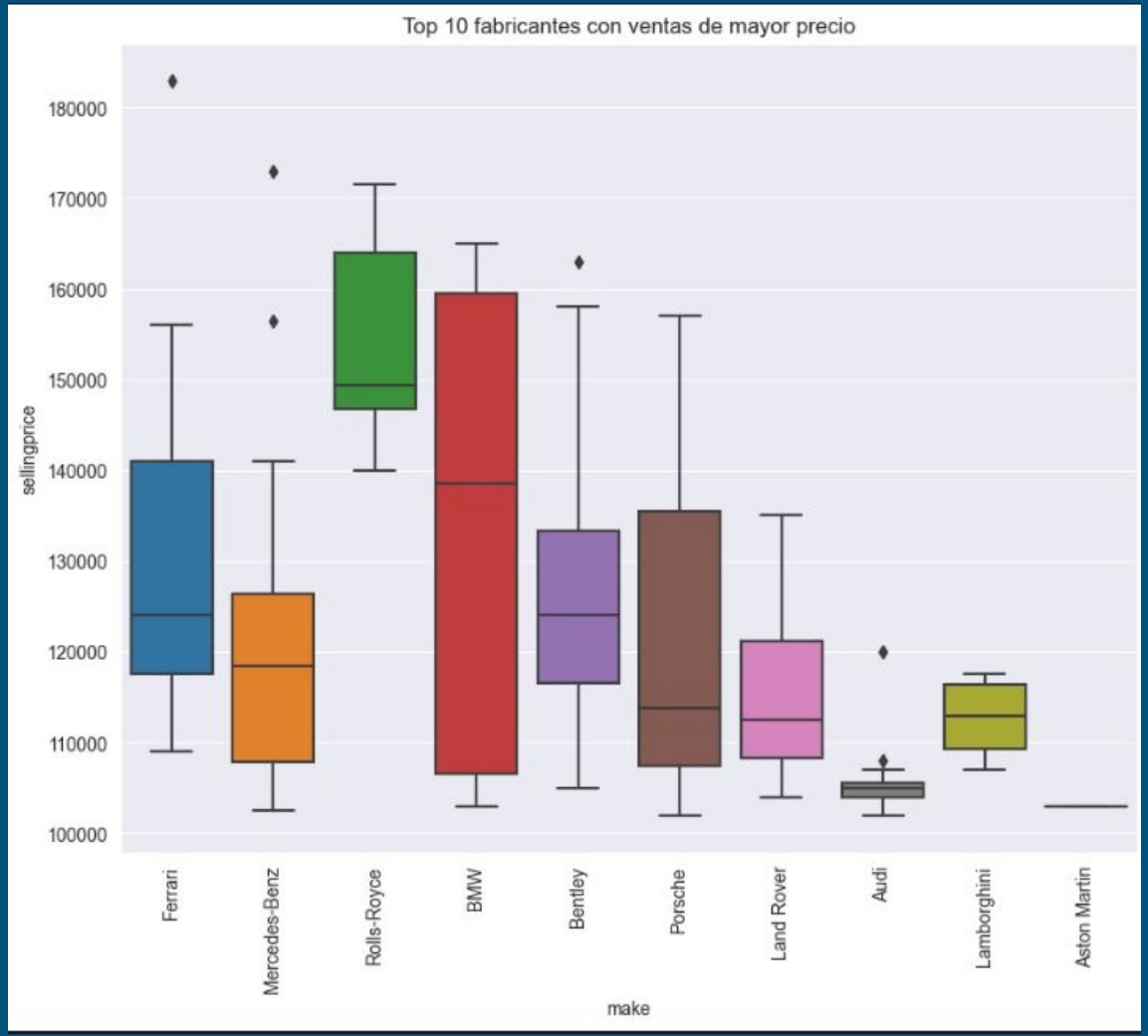
Información general del dataset

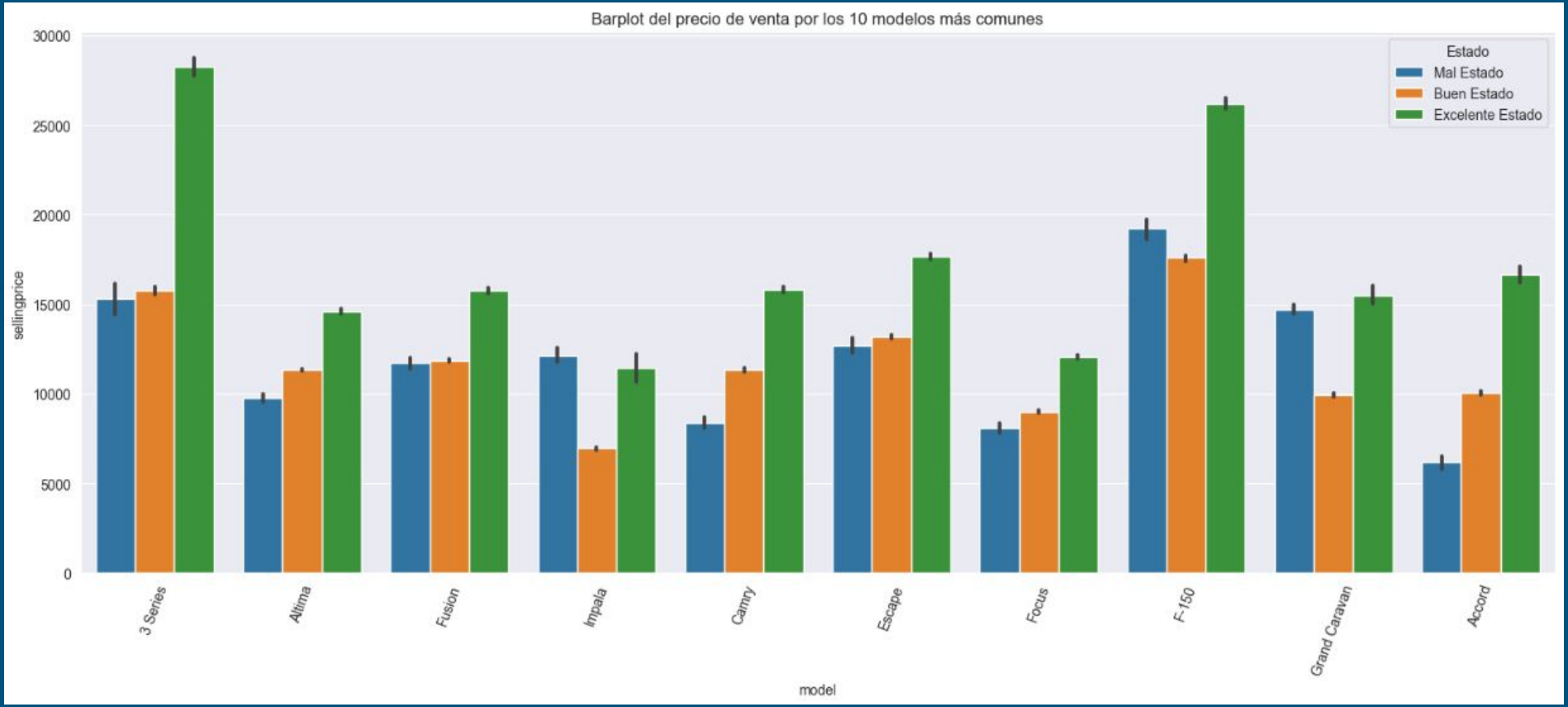
- Contiene: 558837 filas y 16 columnas, de las cuales 10 (diez) son variables independientes a analizar y el target es ‘sellingprice’
- Columnas con valores nulos: ‘make’, ‘model’, ‘trim’, ‘transmission’ y ‘condition’
  - Para resolver esto el enfoque estuvo en la columna body donde se realizó la búsqueda de autos similares a los que no tuviera el body definido y se rellenó ese valor vacío
  - Lo mismo se hizo con la transmisión y al realizar esta solución, conjuntamente borrando 13 mil filas que no tuvieron éxito las demás columnas con esta problemática fueron solventadas.
- Al modificar el dataset su contenido se redujo a 533 mil filas.

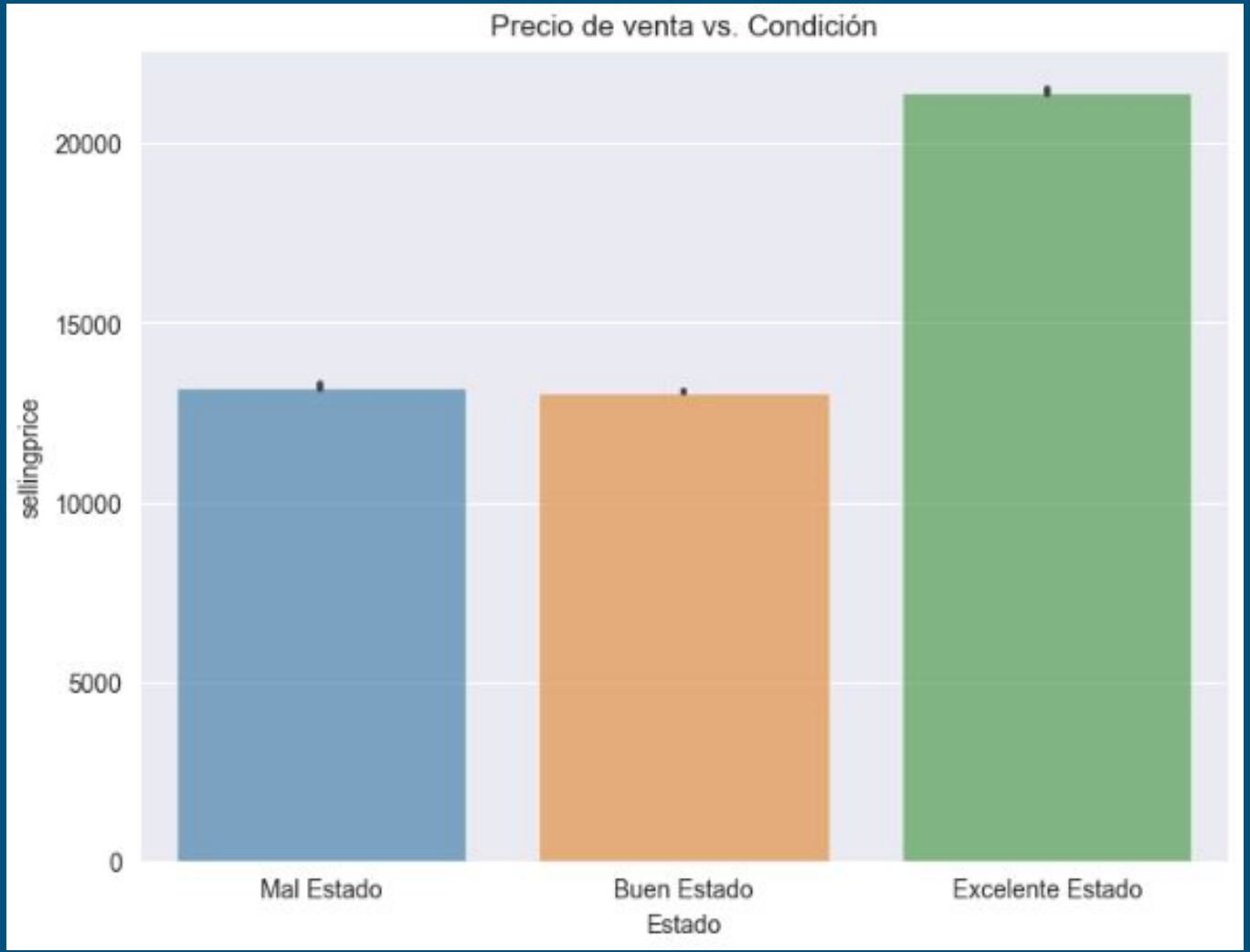
Gráficos

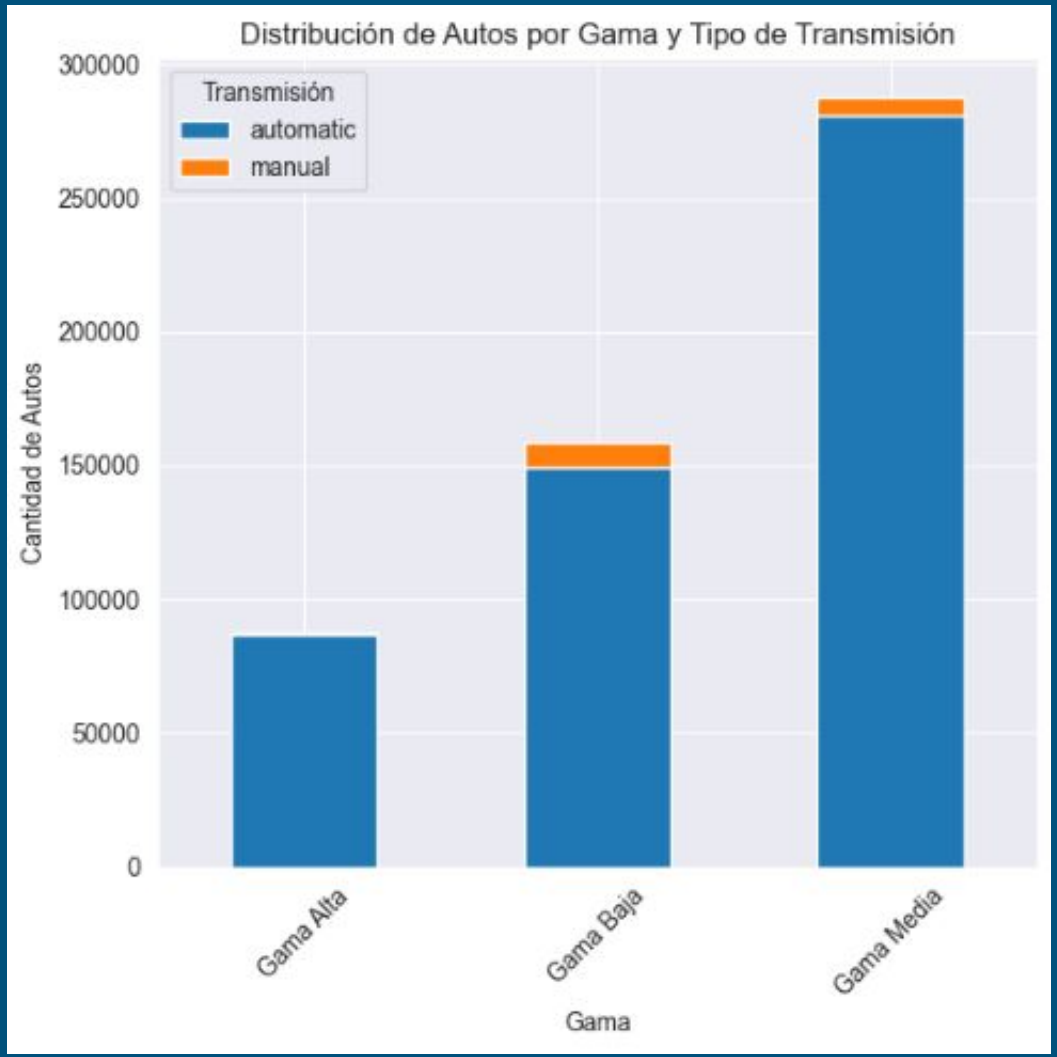
A continuación se mostrarán algunos gráficos de valor que fueron utilizados con el fin de detectar patrones, anomalías e insights

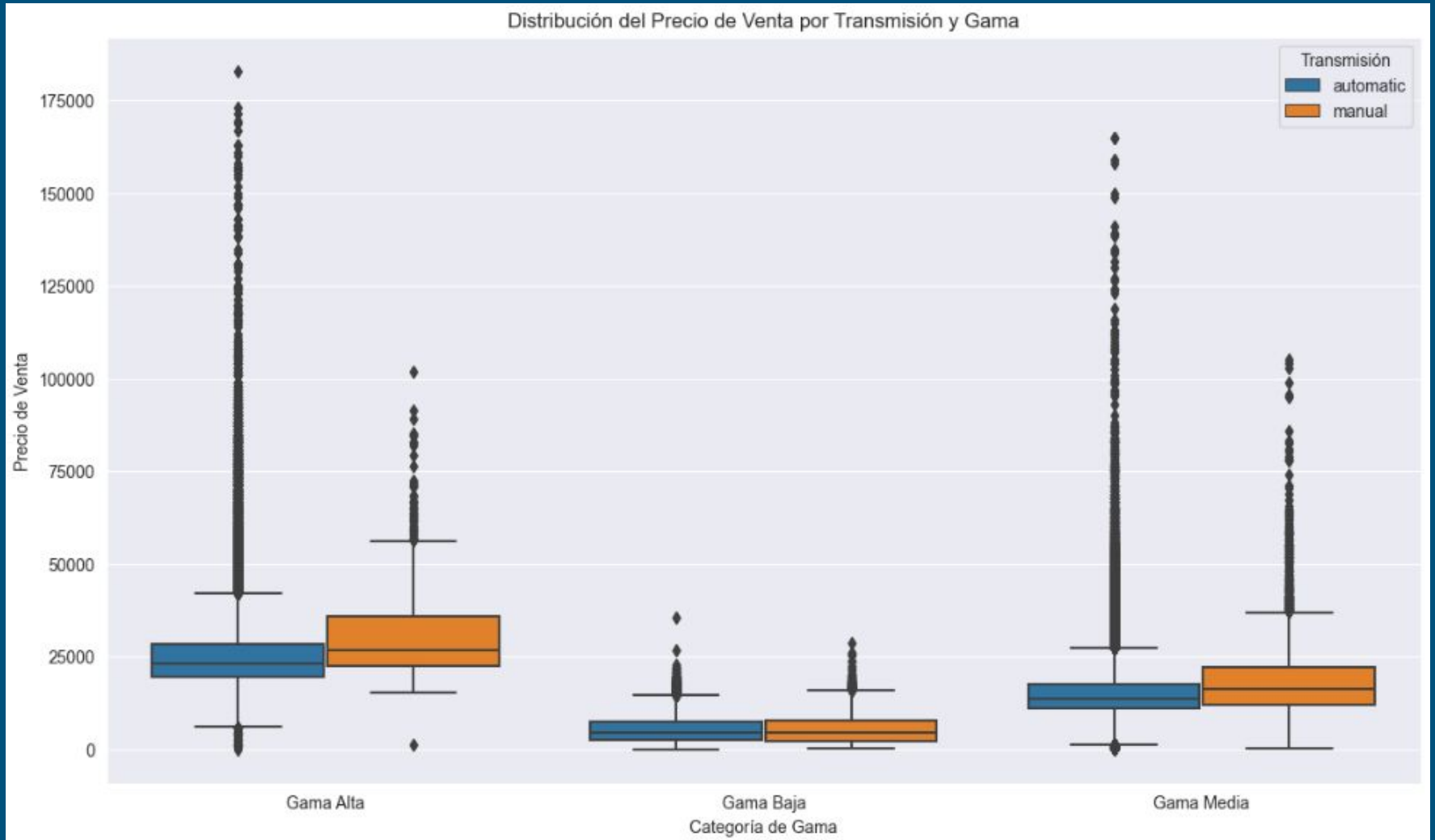
- Top 10 fabricantes con ventas de mayor precio
- Precio de venta por los 10 modelos más comunes
- Precio de venta vs. Condición
- Distribución de Autos por Gama y Tipo de Transmisión
- Distribución del Precio de Venta por Transmisión y Gama
- Precio de Venta vs Kilometraje por Gama

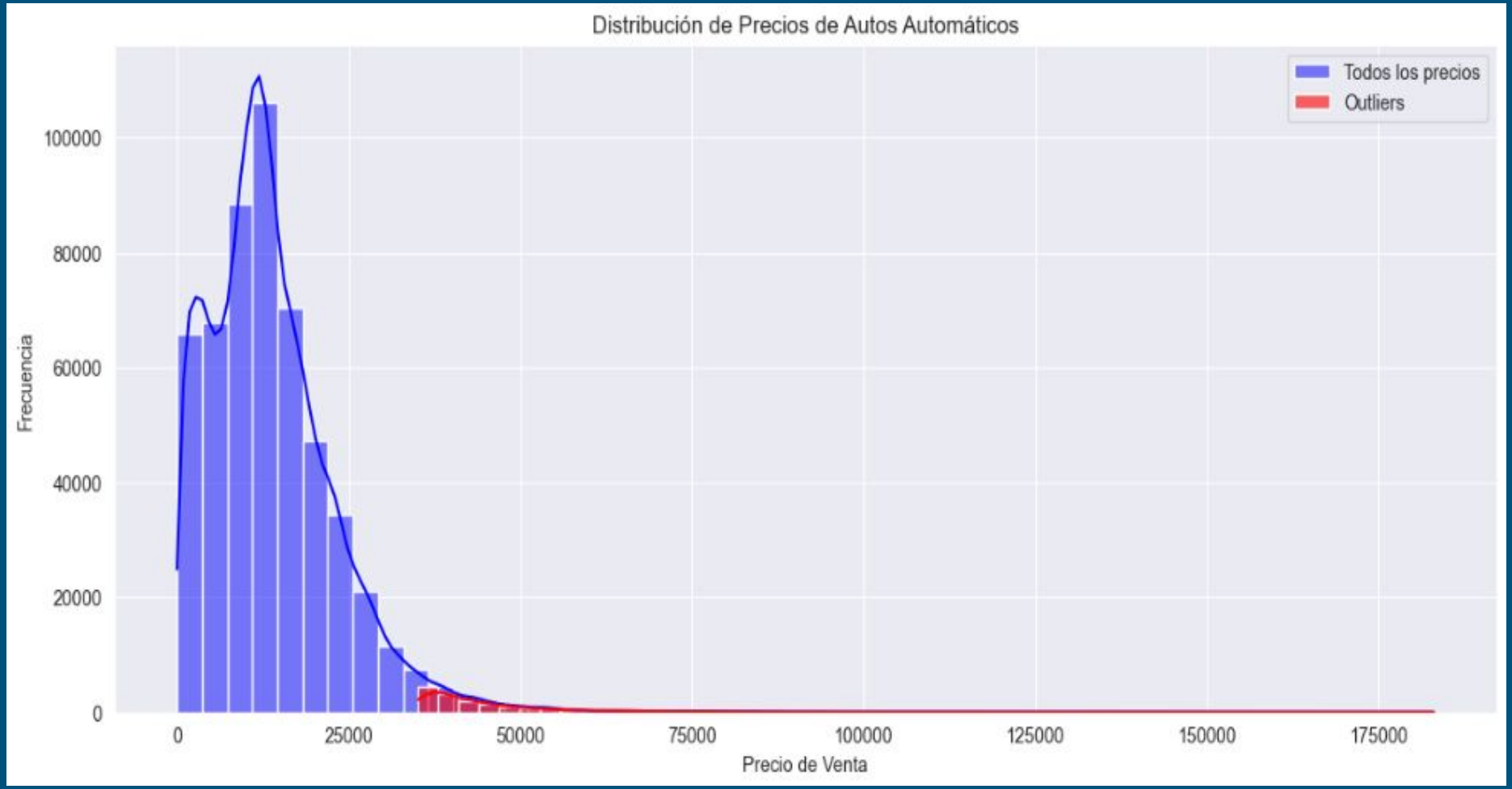

Top 10 fabricantes con ventas de mayor precio
sellingprice
180000
170000
160000
150000
140000
130000
120000
110000
100000
Ferrari
Mercedes-Benz
Rolls-Royce
BMW
Bentley
Porsche
Land Rover
Audi
Lamborghini
Aston Martin
make


Barplot del precio de venta por los 10 modelos más comunes
Estado
Mal Estado
Buen Estado
Excelente Estado
sellingprice
30000
25000
20000
15000
10000
5000
0
3 Series
Altima
Fusion
Impala
Camry
Escape
Focus
F-150
Grand Caravan
Accord
model


Precio de venta vs. Condición
sellingprice
20000
15000
10000
5000
0
Mal Estado
Buen Estado
Excelente Estado
Estado


Distribución de Autos por Gama y Tipo de Transmisión
Transmisión
automatic
manual
300000
250000
200000
150000
100000
50000
0
Cantidad de Autos
Gama Alta
Gama Baja
Gama Media
Gama


Distribución del Precio de Venta por Transmisión y Gama
Transmisión
automatic
manual
Precio de Venta
175000
150000
125000
100000
75000
50000
25000
0
Gama Alta
Gama Baja
Gama Media
Categoría de Gama


Distribución de Precios de Autos Automáticos
Todos los precios
Outliers
Frecuencia
100000
80000
60000
40000
20000
0
0
25000
50000
75000
100000
125000
150000
175000
Precio de Venta

Hipótesis 1

Planteamiento:

La transmisión automática incrementa significativamente el precio de venta de un auto en todas las categorías de gama (Baja, Media, Alta)

Conclusión:

La hipótesis se cumple. Los resultados muestran que:

En Gama Baja: Los autos automáticos tienen precios más altos en comparación con los manuales, aunque los precios de ambos son relativamente bajos.

En Gama Media: Los autos automáticos tienen precios promedio más altos que los manuales. Los precios de los autos automáticos son consistentemente superiores, y los outliers muestran precios elevados.

En Gama Alta: Los autos automáticos también presentan precios significativamente más altos que los manuales, con una amplia variabilidad en los precios, incluyendo los valores más altos del dataset.

En general, la transmisión automática está asociada con precios de venta más altos en todas las categorías de gama, confirmando que la transmisión automática incrementa el precio de venta.

Hipótesis 2

Planteamiento:

El kilometraje tiene un impacto negativo mayor en el precio de venta de autos de gama baja en comparación con autos de gama alta

Esta hipótesis analiza la influencia del kilometraje en diferentes segmentos de mercado. Esto permitirá comprobar si la relación entre el kilometraje y el precio de venta es diferente en autos de diferentes gamas y, por lo tanto, si se deben aplicar diferentes estrategias de precios o modelos específicos por segmento.

Conclusión:

La hipótesis 2 se cumple parcialmente.

Aunque el impacto del kilometraje en el precio de venta es negativo en ambas gamas, el efecto es mayor en autos de gama alta. Esto significa que, a medida que aumenta el kilometraje, el precio de venta disminuye más significativamente para autos de gama alta en comparación con autos de gama baja. Sin embargo, el bajo R-squared en ambos casos sugiere que otros factores además del kilometraje también juegan un papel importante en la determinación del precio de venta.

Hipótesis 3

Planteamiento:

El año de fabricación y la condición del auto son los factores más determinantes en la predicción del precio de venta, independientemente de la gama del auto

Conclusión:

Al igual que la hipótesis anterior nuestra 3er hipótesis también se cumple parcialmente.

Aunque el año de fabricación tiene un efecto positivo significativo en el precio, lo que podría sugerir una cierta valorización con el tiempo, el kilometraje también tiene un impacto negativo significativo.

La condición y el kilometraje del auto no parecen ser factores determinantes en este contexto.

En general, el bajo R^2 indica que hay factores adicionales que deben considerarse para una mejor comprensión del precio de los autos de Gama Alta fabricados antes de 2010.

Hipótesis 4

Planteamiento:

Los autos de gama alta con menos de 100,000 kilómetros mantienen mejor su valor de mercado en comparación con autos de gama media y baja

Conclusión:

Se cumple, ya que los autos de gama alta con menos de 100.000 kilómetros tienen un valor de mercado significativamente mayor en comparación con los autos de gama media y baja

Hipótesis 5

Planteamiento:

El MMR tiene una correlación más fuerte con el precio de venta en autos de gama media y alta que en autos de gama baja.

Conclusión:

Los resultados muestran que el MMR tiene una correlación más fuerte con el precio de venta en autos de gama alta 0,9793 que en autos de gama media 0,9676 y esta a su vez es más fuerte que en autos de gama baja 0,9183.

Modelos de Machine Learning

LinearRegression

Root Mean Squared Error: 7503.75

Mean Squared Error: 56306199.07

Coeficiente de Determinación (R^2): 0.39

Mean Absolute Error: 5217.65324536887

Root train Mean Squared Error: 7570.954

GradientBoostingRegressor

Root Mean Squared Error: 7143.81

Mean Squared Error: 51034035.43

Coeficiente de Determinación (R^2): 0.45

Mean Absolute Error: 4760.115915688201

Root train Mean Squared Error: 11416.05

Mean train Squared Error: 130326244.69

Conclusión

Los modelos no tuvieron buena performance, por ello no van a subirse a producción pero de todas formas es un proyecto viable que con un enfoque distinto o con más información de utilidad pueda ser llevado a cabo.