Proyecto Data Science:

AUTOMOTORA ANACONDA

Hito 2

Integrantes:
- Paulo González
- Jorge Guerrero
- Daniel Mardones
- Gonzalo Rojas
- Esteban Sánchez

Profesor:
- Daniel Beffermann

Tutor:
- Jhon Poma

Jueves 02 de febrero 2023

Resumen Hito 1

Motivación

Generar herramienta de apoyo a la toma de decisiones de los equipos de compra y venta de vehículos.

Hipótesis

Determinar mediante un modelo de regresión en base a las especificaciones de un vehículo si su compra es viable para realizar negocios de reventa.

Propuesta inversion

Comprar vehículos que tengan un precio estimado menor al precio de mercado para poder revenderlos.

Segmentar gama de vehículos

Columna que clasifique la marca del vehículo por generalista y premium.

Resumen Hito 1

Modelo

Modelo de **regresión** con técnicas de Machine Learning

Vector objetivo

Variable “Price” , que muestra el precio en USD de cada vehículo.

Métricas

Median Absolute Error (MAE), Root Mean-Squared Error (RMSE)

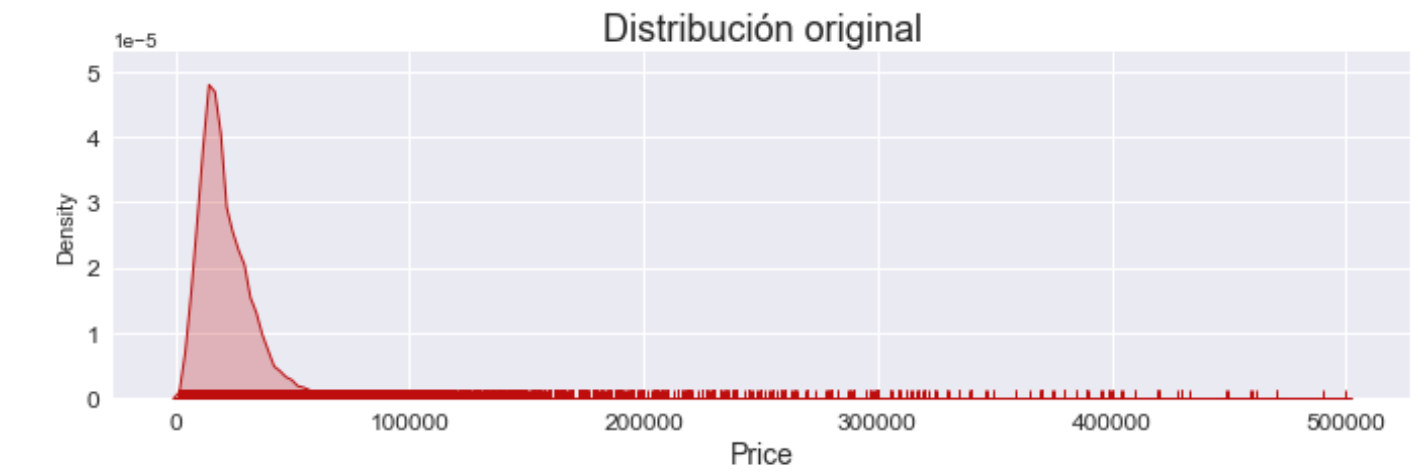
Enriquecimiento data set

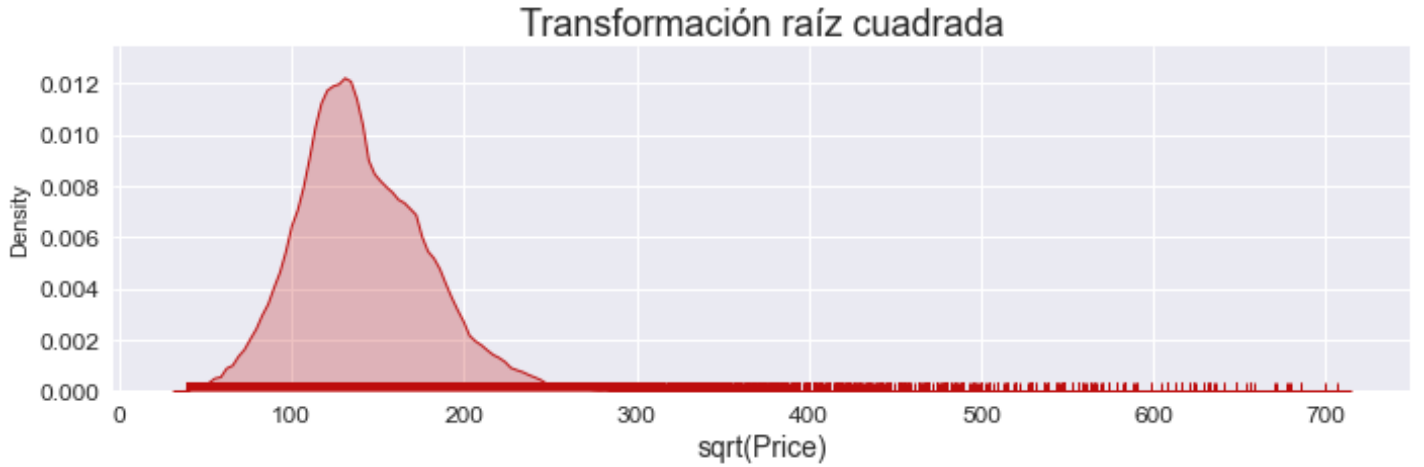
A través de webscrapping para aumentar información de los vehículos

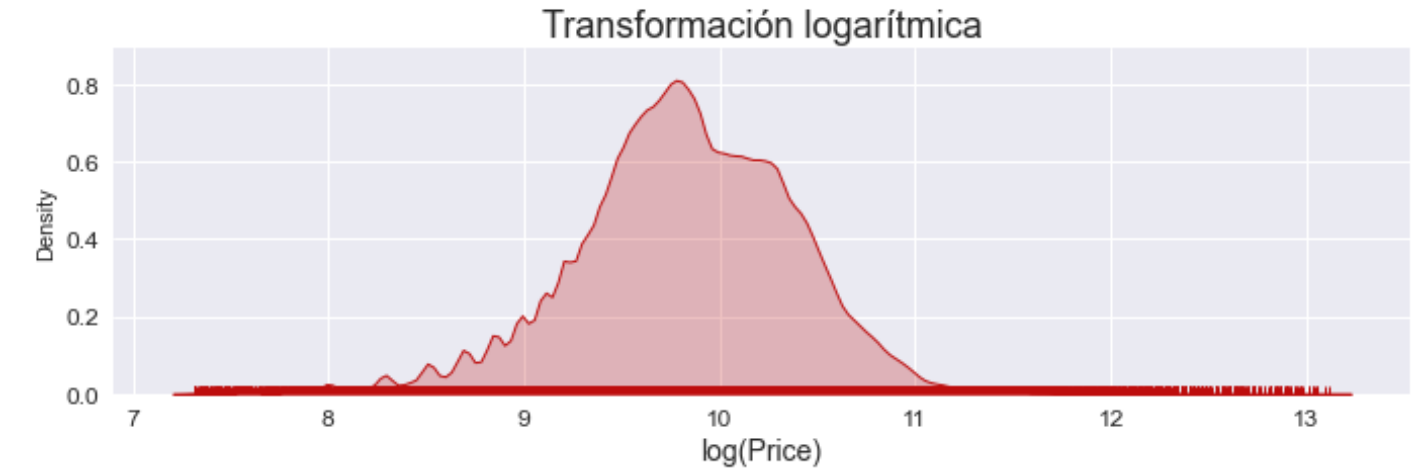
Análisis Exploratorio Descriptivo

Comportamiento vector objetivo “**Price**”

- Distribución original:
  - Sesgo
  - Outliers
- Raíz Cuadrada:
  - Sesgo
  - Outliers
- Logarítmica:
  - Outliers
  - Tendencia normal





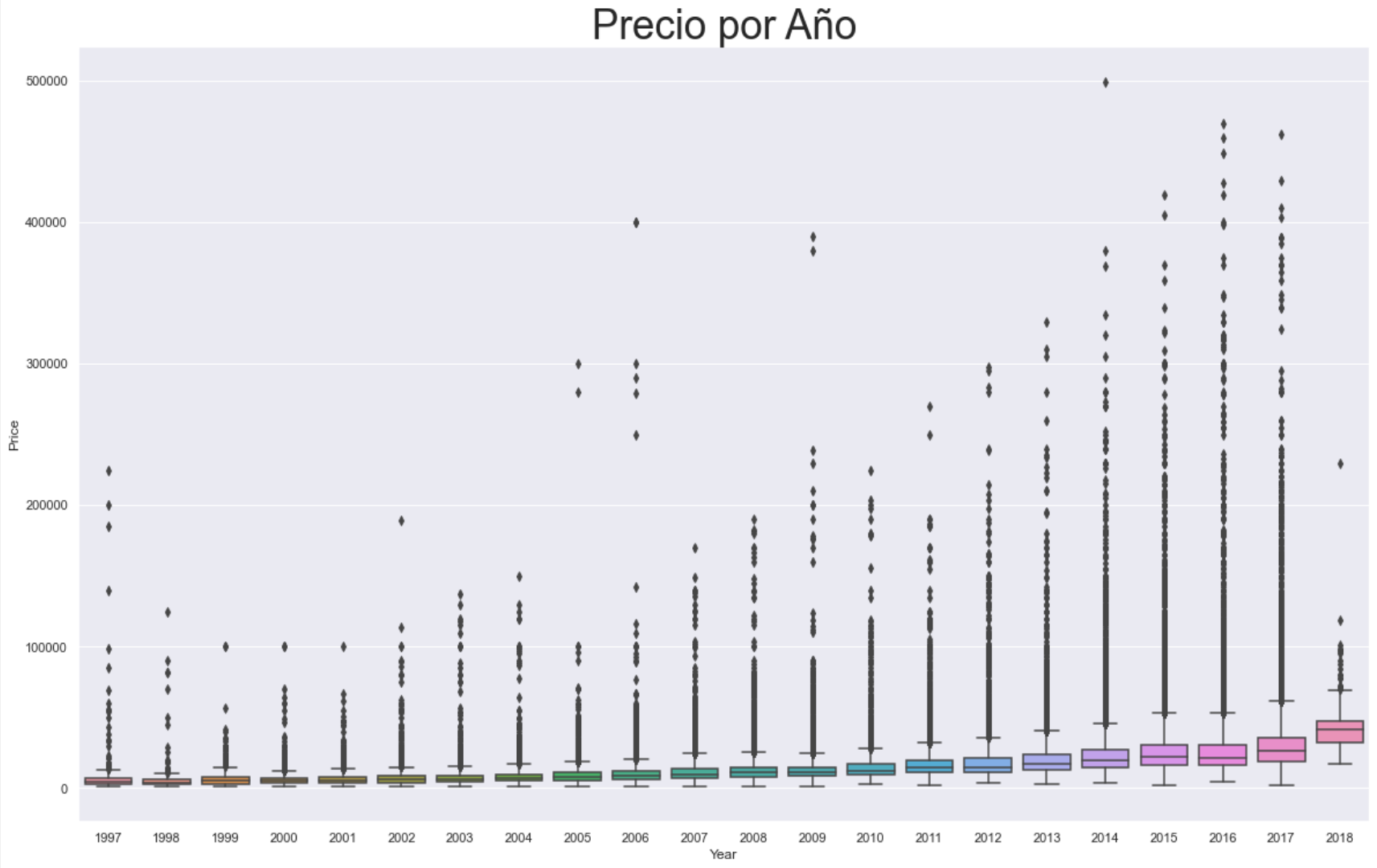


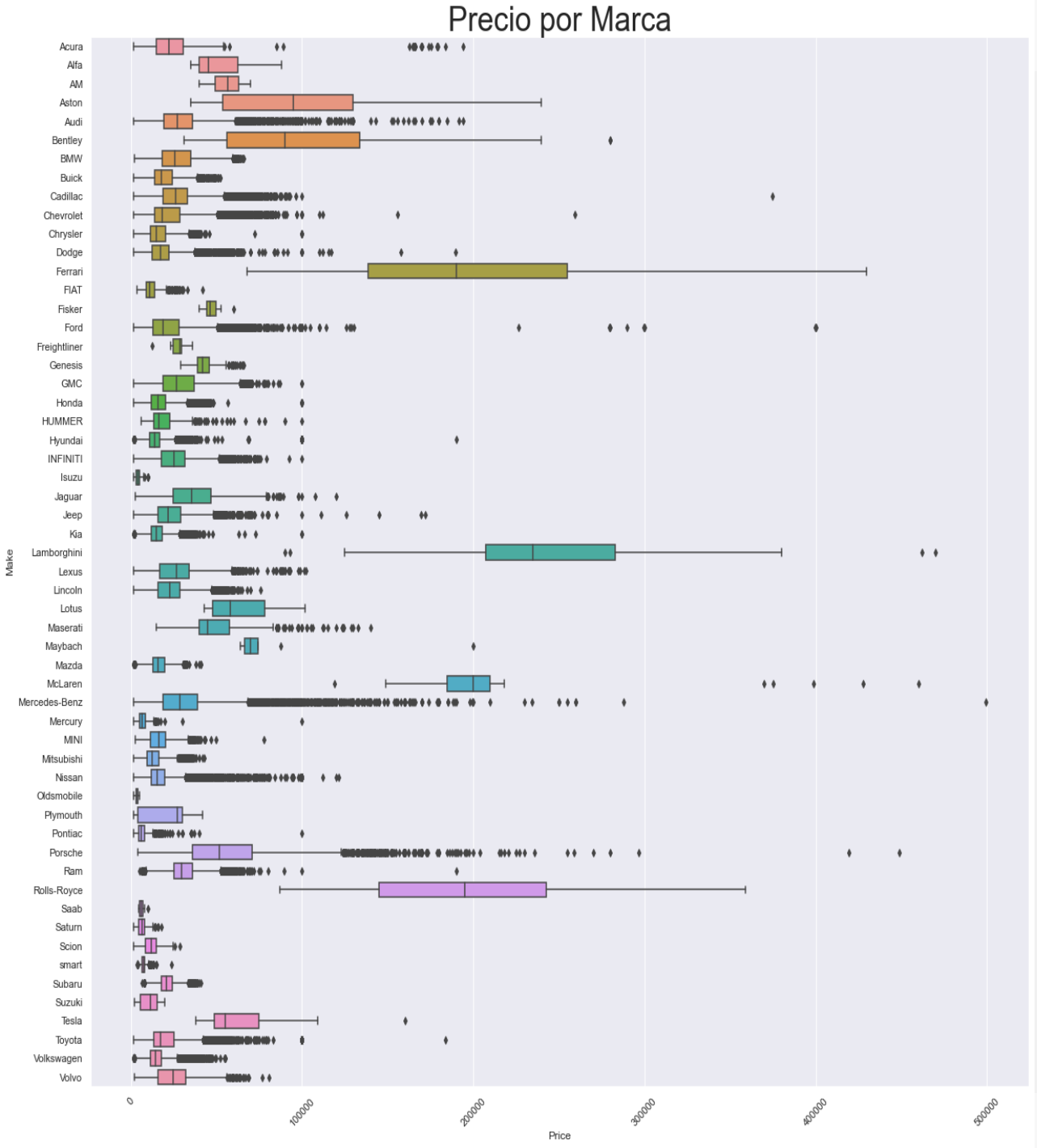
Precio de vehículos por año

El problema más grande al construir el modelo son los outliers.

Se puede ver que para todos los años los precios se concentran en niveles muy bajos.

El precio a través de los años no es la única forma en la que se pueden observar outliers.





Precio de vehículos por marca

Gran presencia de outliers.

Pocas marcas que superan los 200 mil dólares

Marcas de elevado costo, no presentan precios muy altos debido a la presencia de muchos outliers

División de la muestra cada 10.000 dólares para bajar outliers
Outliers > 90.000 dólares se mantiene
Outliers < 90.000 dólares bajan

Solución a los Outliers

- En primer lugar se planteó realizar modelos diferentes para una división de vehículos en dos gamas.
- La división por gama no aportó positivamente al desempeño del modelo puesto que se mantenía el problema de outliers.
- La división de la muestra se hará por tramos de precio, quedando como se puede ver a continuación.

División de muestra para disminuir influencia de outliers basado en columna ‘Price’

Rango de precio (en dólares)	Categoría
0 – 25.000	Generalista
25.001 – 35.000	Premium 1
35.001 – 45.000	Premium 2
45.001 – 55.000	Premium 3
55.001 – 65.000	Premium 4
65.001 – 75.000	Premium 5
75.001 – 85.000	Premium 6
85.001 – 95.000	Premium 7
> 95.001	Premium 8

Pre-Procesamiento

Binarización

Limpieza data set

División vehículos por precio

A través de Base de Datos Postgres (Pg-Admin)

MODELOS CANDIDATOS

Modelos Escogidos

- Para esta etapa del proyecto se realizaron una gran cantidad de modelos “vanilla” de los cuales se tomaron los mejores tres en base a las métricas obtenidas.

- En el cuadro se puede observar cual es el mejor modelo para cada uno de los tramos de precios de los autos.

- El modelo con mejor resultado para la mayoría de las categorías es “GradientBoostingRegressor”.

SEGMENTOS	MEJOR MODELO SEGÚN MAE	Q. DE REGISTROS	% DISTRIBUCION
GENERALISTA	GradientBoostingRegressor	437,973	70.8%
PREMIUM 1	GradientBoostingRegressor	113,138	18.3%
PREMIUM 2	GradientBoostingRegressor	42,299	6.8%
PREMIUM 3	GradientBoostingRegressor	15,621	2.5%
PREMIUM 4	GradientBoostingRegressor	5,005	0.8%
PREMIUM 5	GradientBoostingRegressor	1,617	0.3%
PREMIUM 6	BaggingRegressor	676	0.1%
PREMIUM 7	AdaBoostRegressor	381	0.1%
PREMIUM 8	GradientBoostingRegressor	1,546	0.3%

MODELOS	Q. DE REGISTROS	% DISTRIBUCIÓN ACUMULADA
GradientBoostingRegressor	617,199.00	99.83%
BaggingRegressor	676.00	0.11%
AdaBoostRegressor	381.00	0.06%

Gracias!

¿Dudas?

¿Consultas?

¿Retroalimentación?