

Data Science - Proyecto final

Caso: Trust'N'Ride (Alberto Tejerina)

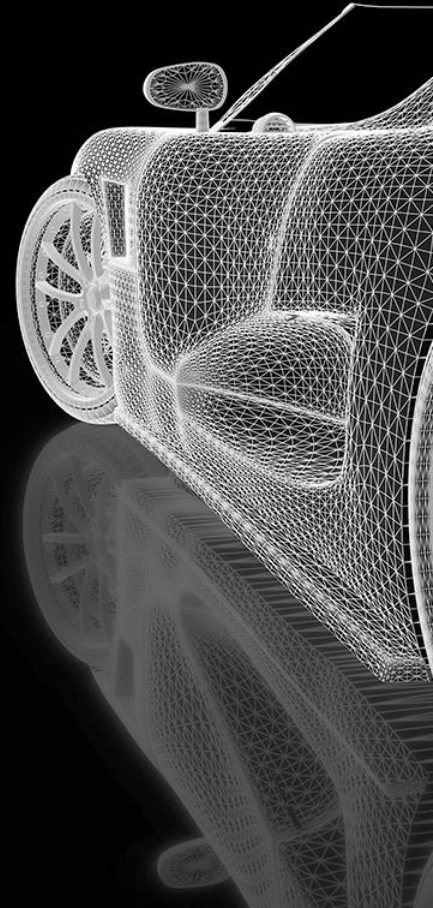
Objetivo

A wireframe illustration of a car, showing the chassis, wheels, and body panels in a transparent, grid-like structure. It is positioned in the top right corner of the slide, partially overlapping the red header.

- La empresa *Trust’N’Ride* vende vehículos usados. La misma se encuentra ante el inconveniente de un mercado tan dinámico y una situación económica del país delicada, que dificulta el estimar un precio de venta para dichos transportes.

Recolección

- El dataset fue obtenido de la base de ventas histórica de la empresa (fuente). Al encontrarse almacenados en sus bases de datos, los mismos ya son estructurados y permitirán ser ordenados y procesados con mayor facilidad. Esta información ha sido entregada en un archivo .csv.
- Contamos con la información de 205 coches y 26 variables que describen cada uno de ellos.



Preparación

A wireframe model of a car, showing the internal structure of the body, wheels, and engine compartment, rendered in a light gray color against a dark background.

El dataset obtenido, no poseía valores nulos ya que en anteriores proyectos, había sido limpiado.

Se convirtieron algunos valores de tipo *object* a *numéricos*, ya que los mismos hacían referencias a números. Por ejemplo:

Campo: `cylindernumber` (Cantidad de cilindros)

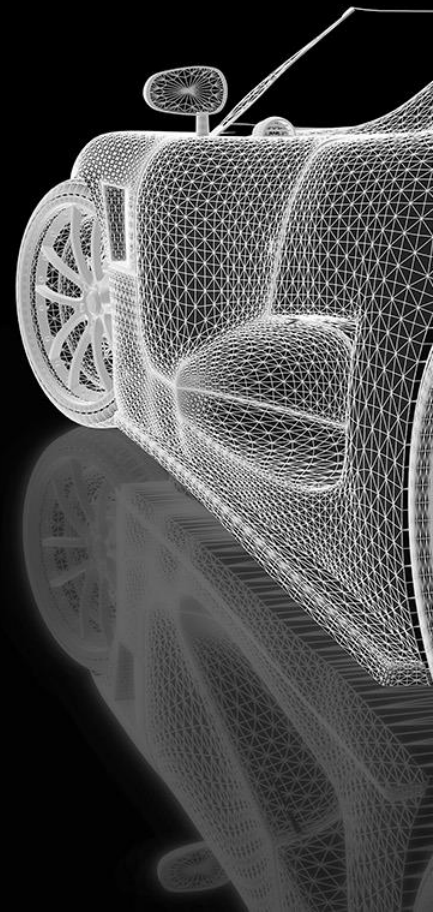
Valores originales: `['four', 'six', 'five', 'three', 'twelve', 'two', 'eight']`

Valores luego de la conversion: `['4', '6', '5', '3', '12', '2', '8']`

De este modo, la información puede ser procesada por los diferentes modelos estadísticos.

Elección del Algoritmo

- Para resolver este caso, utilizaremos algoritmos de aprendizaje supervisado, ya que conocemos el input y el output, que se encuentra en el dataset brindado por *Trust'N'Ride*.
- Específicamente nos enfocaremos en algoritmos de regresión, ya que tienen el objetivo de predecir valores continuos (en nuestro caso, será el Precio del auto)



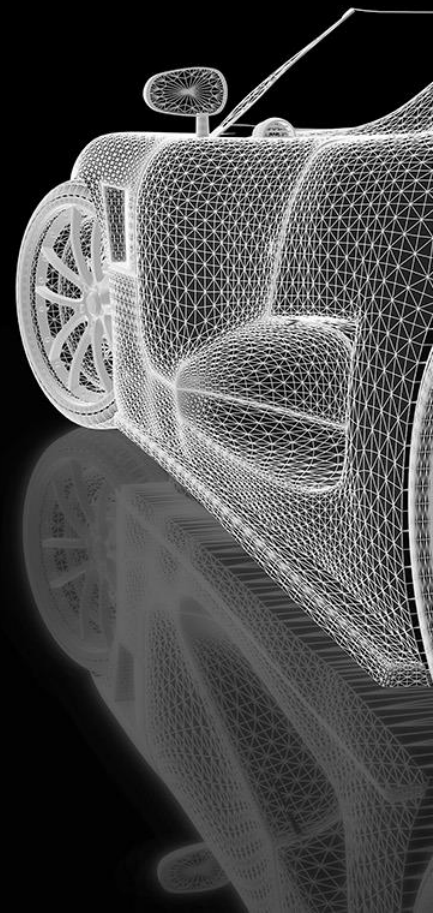
Validación del modelo

- A continuación los modelos empleados y sus tasas de aciertos:

Modelo	Tasa de aciertos
Regresión Lineal	81 %
Random Forest	96 %
Gradient Boosting	91 %
XGboost	92 %

Validación del modelo

- Además de “correr” el algoritmo y validar los resultados, se intentaron optimizar con *RandomizedSearchCV* y *GridSearchCV*, cuyas mejoras no han sido significativas.



Deployment del modelo



Deberemos deployar el modelo basado en ***Random Forest***, que nos permitirá con mayor cantidad de aciertos, inferir los precios de los automóviles, en base a sus diferentes variables.