Modelo de Predicción de precios de Vehículos usados

Jenaro Álvarez

Contexto

En la industria automotriz, determinar el precio justo de un automóvil usado es crucial tanto para los vendedores como para los compradores. Los vendedores desean establecer precios competitivos que atraigan a los compradores potenciales y maximicen sus ganancias, mientras que los compradores buscan obtener un valor justo por el vehículo que están adquiriendo. Sin embargo, la determinación del precio ideal de un automóvil usado puede ser un desafío debido a la amplia variedad de características y factores que influyen en su valor.

Para abordar este desafío, se desarrollará un modelo de machine learning que utiliza datos históricos de automóviles usados para predecir sus precios en función de diversas características. Nuestro modelo analizará características como el año de fabricación, el kilometraje, las características adicionales (por ejemplo, sistemas de entretenimiento, asientos con calefacción), el color, la marca y el tipo de combustible, entre otros, para estimar el precio de un automóvil usado.

Beneficios para la Industria

Precios Justos y Competitivos:

Nuestro modelo ayudará a establecer precios justos y competitivos para los automóviles usados, lo que beneficia tanto a los vendedores como a los compradores.

Optimización de Inventario:

Los concesionarios y vendedores pueden utilizar las predicciones de precios para gestionar de manera más efectiva su inventario, ajustando los precios según la demanda del mercado y maximizando así sus ganancias.

Mejora de la Experiencia del Cliente:

Los compradores pueden confiar en que están obteniendo un precio justo por el automóvil que están adquiriendo, lo que mejora su satisfacción y confianza en el proceso de compra.

Datos y Características Relevantes

El conjunto de datos seleccionado constituye una sólida base de información que comprende miles de registros de vehículos en venta en la ciudad de Nueva York, abarcando desde el año 2012 hasta el 2023. Esta recopilación detallada incluye diversos atributos relevantes, como el estado del vehículo (nuevo o usado, centrándonos exclusivamente en este último), así como características que abarcan desde el tipo de motor y carrocería hasta el rendimiento y las opciones de entretenimiento. Nuestra estrategia se centra en la aplicación de técnicas avanzadas de análisis de datos y aprendizaje automático para descubrir patrones, relaciones y correlaciones significativas que nos permitan desarrollar un modelo predictivo de alta precisión.

Fases del Desarrollo del Modelo

El proceso de desarrollo del modelo se organiza en varias etapas, abarcando desde la limpieza y preprocesamiento de datos, la selección de características relevantes, hasta la elección y evaluación de algoritmos de aprendizaje automático adecuados. La implementación de medidas de evaluación de calidad, como el error cuadrático medio y el coeficiente de determinación (R^2), garantiza la fiabilidad del modelo. Los resultados no solo tienen aplicaciones directas para compradores y vendedores de automóviles, sino que también ofrecen potenciales beneficios en la toma de decisiones estratégicas para la industria automotriz en su conjunto. La contribución al avance del campo de Ciencia de Datos radica en la resolución de un problema práctico y complejo, demostrando la utilidad de las técnicas analíticas y predictivas en la valoración de activos.

Transparencia y Equidad en el Mercado

El objetivo último de este proyecto es mejorar la transparencia y la equidad en el mercado de automóviles. Al proporcionar una herramienta precisa y objetiva para la determinación de precios, buscamos beneficiar tanto a compradores como a vendedores en sus decisiones de compra y venta. Este enfoque contribuye a la eficiencia del mercado y establece un estándar equitativo en las transacciones automotrices.

Antes de comenzar, algunos supuestos.

- La variable más fuertemente relacionada al precio es la marca del automóvil.
- La edad del auto tiene una relación inversamente proporcional con el precio.
- El rendimiento y la distancia recorrida está en el top 5 de variables más relacionadas con el precio.
- Los autos con combustibles alternativos/eléctricos/híbridos tienen un valor más alto.
- La existencia de tracción en las 4 ruedas (4x4 o 4Wd) aumenta el valor del vehículo.

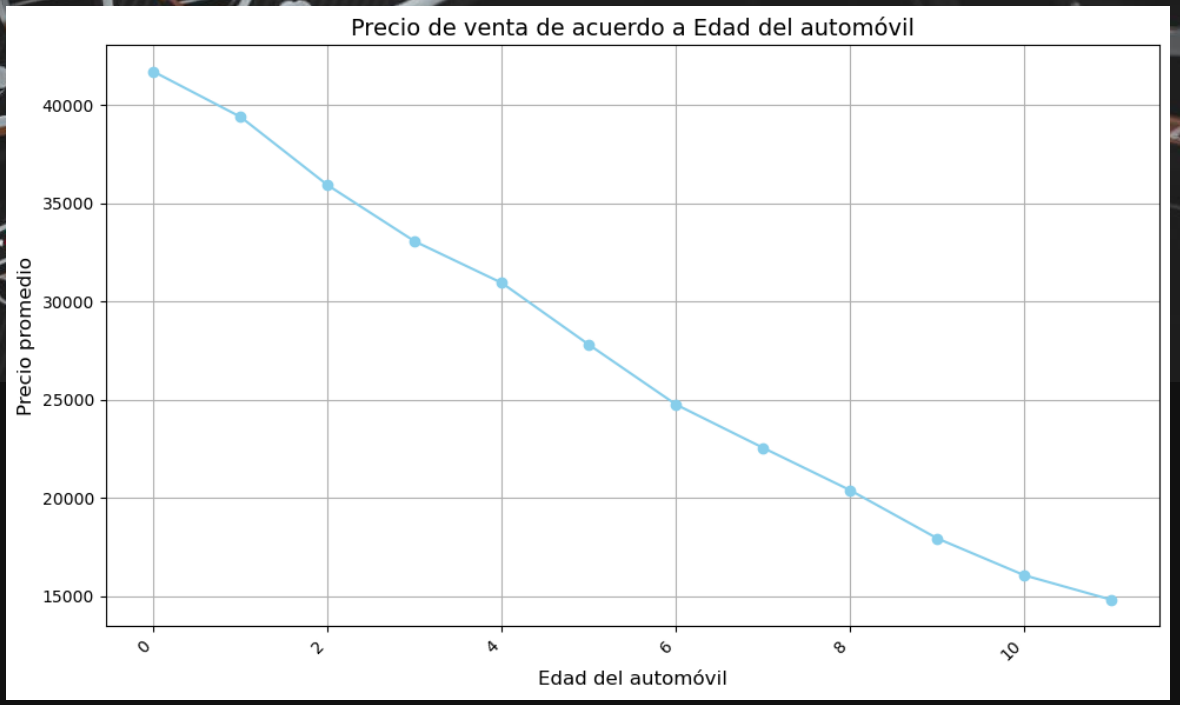
- La variable más fuertemente relacionada al precio es la marca del automóvil.

Falso!

Como podemos observar en este gráfico que muestra a las cinco variables con la mayor correlación con la variable objetivo, "Price", es el año de fabricación el cual obtiene la mayor correlación con el precio del automóvil. La edad del vehículo tendrá una relación inversamente proporcional con el precio, es decir, entre más años tenga el auto, menor será su precio.

- La edad del auto tiene una relación inversamente proporcional con el precio.

Verdadero!



En este gráfico podemos apreciar claramente la relación entre la edad del automóvil con su precio medio de venta. Se observa que ambas variables se relacionan de una manera inversamente proporcional, aumentando una cuando la otra disminuye, y viceversa.

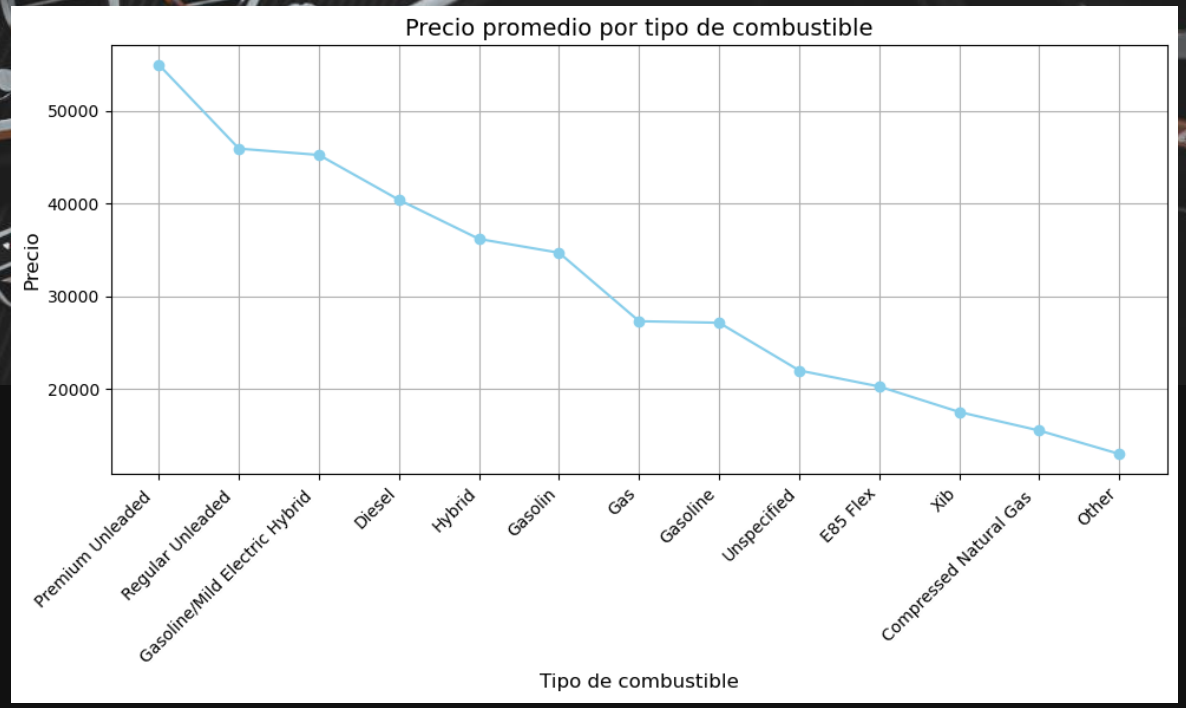
- El rendimiento y la distancia recorrida está en el top 5 de variables más relacionadas con el precio.

Falso!

Como pudimos observar anteriormente solo la distancia recorrida por un automóvil (Mileage) entra en el Top 5, mientras que el rendimiento, medido en Millas por Galón se posiciona en el lugar número 12, por debajo de varias características de conveniencia, al parecer el rendimiento del vehículo no es tan importante en la determinación como podría parecer en primera instancia.

- Los autos con combustibles alternativos/eléctricos/híbridos tienen un valor más alto.

Falso!



Fácilmente observable es que los tipos de combustible alternativos e híbridos tienden a posicionarse en la parte más baja de distribución de precios en contraste con la gasolina tradicional sin plomo y su versión premium. Esto puede deberse a una falta de convencimiento por parte de los compradores para con estas nuevas tecnologías, o también a problemas de averías más difíciles de solucionar en este tipo de vehículos con tecnologías menos extendidas.

- La existencia de tracción en las 4 ruedas (4x4 o 4Wd) aumenta el valor del vehículo.

Verdadero!

Queda claro en el gráfico que la tracción en las cuatro ruedas (4Wd) es el tipo de tracción cuyos vehículos poseen el precio promedio más alto, sacándole más de 3.000 dólares a la tracción trasera y casi 13.000 a la tracción delantera.

Insights

Número 1

La mejor variable predictora parece ser la edad del vehículo, a mayor edad, menor precio.

Combustible

El valor promedio de los vehículos a combustibles tradicionales es mayor que aquellos que utilizan combustibles alternativos/híbridos.

Tracción

Los vehículos con tracción en las cuatro ruedas tienen un precio mayor a aquellos con tracción delantera o trasera.

Rendimiento

Si bien importante en cierta medida, el rendimiento no es tan importante como se podría pensar en la determinación de precios, quedando por debajo, incluso, de opciones de conveniencia.

Mileage

A medida que aumenta el número de millas recorridas por un vehículo su valor tiende a disminuir.

Uso comercial

El uso comercial (no personal) de un vehículo no parece impactar en el valor.

Evaluación del Modelo:

Cat Boost Regressor (92,2%)

Para un modelo de predicción de precios construido con CatBoost, que se destacó como el mejor entre varios otros modelos como regresión lineal, random forest y decision tree, las métricas obtenidas son muy alentadoras. En el conjunto de entrenamiento, el modelo presenta un **MAE de 0.071**, un **MSE de 0.009** y un **R-cuadrado de 0.954**, lo que indica una buena capacidad del modelo para ajustarse a los datos de entrenamiento y explicar la variabilidad de los precios. En el conjunto de prueba, aunque hay un ligero aumento en el error (**MAE de 0.091 y MSE de 0.015**), el modelo aún mantiene un buen rendimiento con un **R-cuadrado de 0.922**, lo que sugiere una **capacidad sólida de generalización**. Además, el MSE promedio de las validaciones cruzadas es consistente y bajo, indicando una buena estabilidad del modelo. Con un tiempo de ejecución de 36 segundos, CatBoost demuestra eficiencia computacional mientras mantiene un rendimiento excepcional en la predicción de precios. En resumen, el modelo CatBoost es una opción sólida y confiable para la predicción de precios en este contexto.

Muchas gracias.