Análisis y Modelado del Mercado de Autos Usados en Europa Central

Correa, Juan Manuel - Caprari, Máximo

17 de junio de 2025

Introducción y contextualización del problema

El mercado de autos usados genera una gran cantidad de información a partir de los anuncios publicados en diversas plataformas digitales. Analizar estos datos de manera sistemática resulta fundamental para comprender la dinámica de precios, la demanda de diferentes modelos y las características que influyen en el valor de los vehículos [2]. Sin embargo, antes de poder extraer conclusiones válidas, es imprescindible realizar un proceso riguroso de exploración y curación de los datos, que permita garantizar la calidad y representatividad del conjunto de datos analizado [1].

El análisis realizado sobre este dataset abarca enfoques descriptivos, exploratorios y predictivos. Por un lado, se busca describir y explorar la estructura y relaciones entre las variables, y por otro, se desarrollan modelos supervisados para predecir el precio de los autos a partir de sus características técnicas y de publicación [2].

El objetivo general es identificar los principales factores que determinan el precio de los autos usados y construir modelos que permitan estimar su valor de manera precisa, así como segmentar el mercado en grupos de vehículos con atributos similares. De este modo, el análisis contribuye a una mejor comprensión del mercado y a la toma de decisiones informadas por parte de compradores, vendedores y plataformas de comercialización.

Exploración inicial y curación de datos

Carga y descripción del dataset

El análisis comenzó con la carga de un extenso dataset de anuncios de autos usados, compuesto originalmente por más de 3,5 millones de registros y 16 variables. Se revisaron las dimensiones del conjunto de datos, identificando la cantidad de filas y columnas, así como los tipos de variables presentes, que incluían tanto variables numéricas como categóricas y de fecha. Además, se realizó una evaluación exhaustiva de los valores faltantes, detectando que algunas variables presentaban un alto porcentaje de datos nulos, lo que motivó la necesidad de una curación cuidadosa antes de cualquier análisis posterior. El porcentaje de valores faltantes para cada variable se calcula como:

$$\text{Porcentaje de nulos}_j = \frac{\sum_{i=1}^{N} \mathbb{I}(x_{ij} = \text{nulo})}{N} \times 100$$

Limpieza y transformación de datos

Para mejorar la calidad y robustez del análisis, se eliminaron aquellas columnas que presentaban más del 50 % de valores faltantes. Posteriormente, se procedió a la conversión de los tipos de datos, transformando las fechas a formato `datetime` y asegurando que las variables numéricas que representaban cantidades enteras, como el año de fabricación o el número de puertas, fueran almacenadas como enteros. Los valores faltantes en las variables numéricas restantes

se imputaron utilizando la mediana, mientras que en las variables categóricas se utilizó la moda, garantizando así la coherencia y reduciendo el impacto de valores extremos [2].

Un paso fundamental en la limpieza fue el tratamiento de valores atípicos o extremos (outliers). Para ello, se aplicó el método del rango intercuartílico (IQR), eliminando aquellos registros que se encontraban fuera de los límites considerados razonables para cada variable numérica. Esta depuración permitió obtener distribuciones más representativas y fiables para el análisis estadístico y la construcción de modelos predictivos [1].

$$\begin{aligned} \text{IQR} &= Q_3 - Q_1 \\ \text{Límite inferior} &= Q_1 - 1{,}5 \times \text{IQR} \\ \text{Límite superior} &= Q_3 + 1{,}5 \times \text{IQR} \end{aligned}$$

Además, se estandarizaron algunas unidades, como la conversión de la potencia del motor de kilovatios a caballos de fuerza, y se generaron nuevas variables a partir de la información existente, tales como la antigüedad del vehículo y la duración de publicación del anuncio.

Justificación de decisiones de curación

Las decisiones tomadas durante la curación de los datos se fundamentan en la necesidad de asegurar la validez y robustez del análisis. La eliminación de columnas con altos porcentajes de valores nulos evita sesgos y resultados poco representativos. La imputación de valores faltantes con la mediana y la moda permite mantener la mayor cantidad de información posible sin distorsionar las distribuciones originales [2]. La conversión de tipos y la estandarización de unidades facilitan la manipulación de los datos y la interpretación de los resultados. Finalmente, la creación de nuevas variables relevantes, como la antigüedad del vehículo, enriquece el análisis y aporta valor a la modelización posterior.

Análisis exploratorio y visualización

Se analizaron las distribuciones de las variables más relevantes, como el precio, el kilometraje y la potencia del motor. Los histogramas muestran que el mercado está dominado por vehículos de precio bajo y kilometraje moderado, con una menor frecuencia de autos de alto precio o kilometraje elevado. Tras la eliminación de valores atípicos, las distribuciones se volvieron más representativas y permitieron identificar patrones y tendencias reales en el mercado. Además, se realizó una comparación visual de las distribuciones antes y después del proceso de limpieza de datos, lo que permite observar el impacto de la eliminación de valores atípicos en la representatividad de los datos.

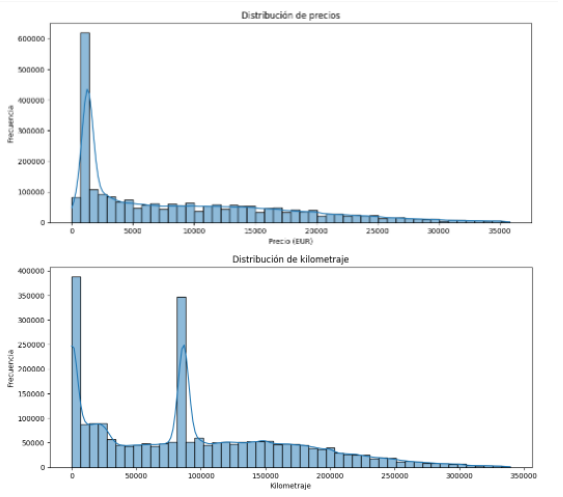

Figura 1: Distribución de variables relevantes.

Por otro lado, se calculó la matriz de correlación entre las variables numéricas, identificando relaciones significativas, como la fuerte correlación negativa entre el año de fabricación y el kilometraje, y la correlación positiva entre el precio y la potencia del motor. Estas relaciones aportan información valiosa para la selección de variables en modelos predictivos y para la interpretación de los factores que influyen en el valor de los vehículos. A continuación, se muestra

la matriz de correlación obtenida:

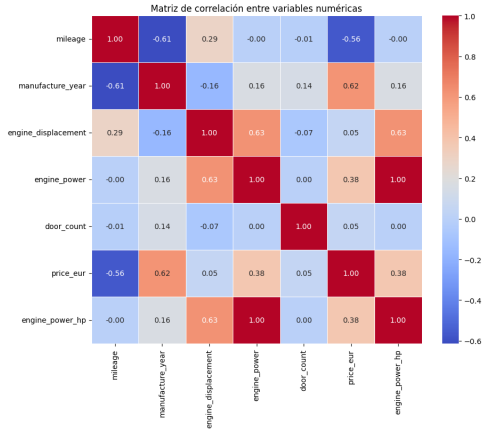

Figura 2: Matriz de correlación entre variables numéricas

Asimismo, se realizaron comparaciones entre diferentes categorías, como el tipo de transmisión, mostrando que los vehículos con transmisión automática tienden a tener precios más altos y mayor dispersión. También se analizaron otras variables categóricas, como el número de puertas y el tipo de combustible, para identificar diferencias en los precios y características del mercado.

Finalmente, se desarrolló un dashboard interactivo utilizando Plotly Dash que integra los principales indicadores del mercado automotriz, incorporando filtros dinámicos por marca, tipo de carrocería y rango de precios. El tablero incluye visualizaciones de distribución de precios, análisis temporal de publicaciones, relación entre año de fabricación y kilometraje, y métricas clave como precio promedio y antigüedad de los vehículos. Esta herramienta proporciona una comprensión integral de las tendencias del mercado de autos usados, optimizando el proceso de análisis exploratorio y facilitando la identificación de patrones relevantes para el desarrollo de estrategias comerciales y modelos predictivos más robustos.

Segmentación de mercado (Análisis de Clústeres)

Con el objetivo de identificar patrones ocultos y segmentar el mercado automotor en función de características relevantes, se aplicó un análisis de clústeres sobre tres variables numéricas clave: kilometraje (mileage), potencia del motor (engine_power) y precio en euros (price_eur). Estas variables fueron estandarizadas mediante z-score para evitar que sus distintas escalas influyeran en el agrupamiento [1]. El número óptimo de grupos se determinó con el método del codo, que sugirió tres clústeres. Se entrenó un modelo K-means con $k = 3$, segmentando el conjunto de vehículos en tres grupos diferenciados [2].

El análisis mostró una distribución equilibrada: el clúster 2 concentra el 42 % de las observaciones, seguido por el clúster 0 (29,3 %) y el clúster 1 (28,7 %). El clúster 0 agrupa vehículos con alto kilometraje, potencia media-alta y precios bajos, sugiriendo autos antiguos o muy utilizados. El clúster 1 reúne autos seminuevos o recientes, con mayor potencia y precios elevados, mientras que el clúster 2 contiene vehículos estándar, de kilometraje moderado, baja potencia y precios accesibles.

Para visualizar la segmentación, se graficaron las relaciones entre las variables según el grupo asignado y se aplicó un Análisis de Componentes Principales (PCA) para reducir la dimensionalidad. El PCA conservó el 89 % de la varianza total en dos componentes, confirmando la separación entre los tres grupos y validando la calidad de la segmentación [1].

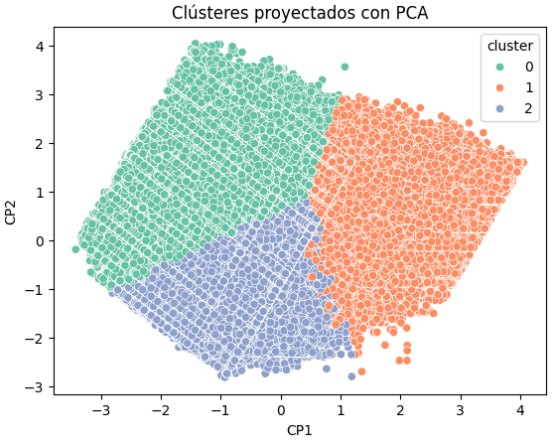

Figura 3: Visualización de los clústeres en el espacio de componentes principales.

Modelado lineal predictivo del precio.

Se construyó un modelo de regresión lineal para predecir el precio de los vehículos a partir de sus características, incluyendo variables derivadas como antigüedad, duración del anuncio y relación potencia por litro. Las variables categóricas se transformaron mediante one-hot encoding y las numéricas se estandarizaron [2]. El modelo, entrenado sobre el 80 % de los datos y evaluado sobre el 20 % restante, obtuvo un MAE de 3.843,50, un RMSE de 5.091,66 y un R^2 de 0,63. Marcas premium, duración del anuncio y eficiencia del motor tuvieron efecto positivo en el precio, mientras que antigüedad, kilometraje, tipo de transmisión y ciertas carrocerías lo redujeron.

Modelado libre

En esta sección el objetivo principal es comparar el desempeño de distintos algoritmos de regresión supervisada y seleccionar el modelo más adecuado para la estimación del precio [2].

El primer paso consistió en la selección de variables relevantes para la predicción. Posteriormente, el conjunto de datos fue dividido en dos partes: un 80 % para el entrenamiento de los modelos y un 20 % para su evaluación, asegurando la reproducibilidad mediante la fijación de una semilla aleatoria [1].

Para la tarea de modelado libre se seleccionaron tres algoritmos con diferentes niveles de complejidad: *Decision Tree Regressor*, *Random Forest Regressor* y *XGBoost Regressor*. Cada modelo fue entrenado con los mismos datos de entrenamiento y se ajustaron hiperparámetros razonables para controlar el sobreajuste y optimizar el rendimiento.

La evaluación de los modelos se realizó sobre el conjunto de prueba, utilizando métricas estándar como el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Los resultados mostraron que el modelo XGBoost obtuvo el mejor desempeño, con el menor MAE y RMSE, y un R^2 cercano a 0.9, lo que indica una alta capacidad explicativa. El modelo Random Forest también presentó buenos resultados, aunque ligeramente inferiores, mientras que el árbol de decisión fue el menos preciso.

Para profundizar en el análisis, se examinó la importancia relativa de las variables según el modelo Random Forest. Se observó que el kilometraje es, por amplio margen, la variable más influyente en la predicción del precio, seguido por la antigüedad del vehículo y la potencia del motor. En contraste, variables como el tipo de carrocería, la transmisión o la marca/modelo tuvieron una contribución marginal, posiblemente debido a una codificación poco informativa o a una baja correlación directa con el precio.

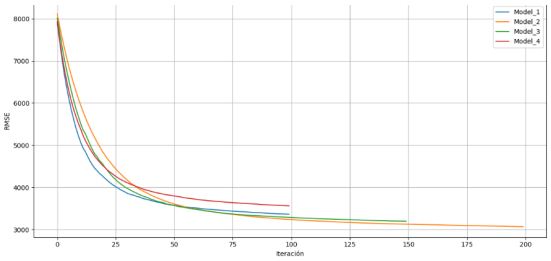

Figura 4: Evolución del RMSE en validación para los distintos modelos de XGBoost.

Dado el destacado desempeño de XGBoost, se procedió a optimizar algunos de sus hiperparámetros clave, como el número de árboles, la profundidad máxima, la tasa de aprendizaje y la proporción de datos y variables utilizadas en cada árbol. Se compararon distintas configuraciones, evaluando la evolución del error RMSE en el conjunto de validación a lo largo de las iteraciones. Este análisis permitió identificar la configuración óptima de hiperparametros.

Cuadro 1: Hiperparámetros de los modelos XGBoost

Modelo	Est.	Prof.	LR	Sub
XGB (base)	100	10	0.10	1.0
XGB_A	100	10	0.10	1.0
XGB_B (óptimo)	200	10	0.05	1.0
XGB M2	200	6	0.05	1.0
XGB M3	150	5	0.07	0.9

Cuadro 2: Comparación de desempeño de los modelos

Modelo	**MAE**	**RMSE**	$\mathbf{R^2}$
Regresión Lineal	3843.50	5091.66	0.63
Decision Tree	2315.56	3715.48	0.81
Random Forest	1779.02	2960.19	0.88
XGBoost (base)	1617.91	2635.69	0.90
XGB_A	1617.91	2635.69	0.90
XGB_B (óptimo)	1612.89	2629.76	0.90
XGB M2	1955.96	3068.72	0.87
XGB M3	2054.81	3196.92	0.86

Consideraciones éticas y de gobernanza

La implementación del modelo requiere considerar aspectos éticos y de gobernanza. En cuanto a gobernanza, es fundamental garantizar la calidad, seguridad y trazabilidad de los datos, cumpliendo con normativas como el GDPR y asignando roles específicos para su gestión. Éticamente, deben evitarse sesgos en las predicciones causados por una representación desigual de ciertas marcas o tipos de vehículos. Para ello, se recomienda revisar las distribuciones y evaluar el desempeño por subgrupos. La regresión lineal aporta transparencia, pero si se emplean modelos más complejos, es clave usar herramientas explicativas y establecer revisiones periódicas para asegurar su uso responsable.

Conclusiones y líneas futuras de investigación

Los resultados obtenidos muestran una clara mejora en el desempeño predictivo a medida que se emplean modelos más complejos y avanzados. La regresión lineal, aunque sencilla e interpretable, presenta un desempeño limitado ($R^2 = 0.63$). El árbol de decisión mejora notablemente la capacidad explicativa, pero es superado por los modelos de ensamble como Random Forest y, especialmente, XGBoost, que alcanza el mejor resultado global (MAE = 1612.89, RMSE = 2629.89, $R^2 = 0.90$).

Como líneas futuras de investigación, se propone profundizar en la optimización de hiperparámetros mediante técnicas como *Grid Search* o *Bayesian Optimization*, así como implementar y comparar otros algoritmos de *gradient boosting* como LightGBM y CatBoost, que podrían ofrecer mejoras adicionales en precisión y eficiencia. También resulta relevante explorar modelos de redes neuronales profundas y técnicas de ensamble como stacking o blending. Estos enfoques permitirán seguir mejorando la precisión de las estimaciones y aportar herramientas más robustas para la toma de decisiones en el mercado de autos usados.

Referencias

[1] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

[2] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.