Informe del Proyecto:
Predicción del Precio de Coches Usados

Introducción

Este proyecto se centra en la predicción del precio de coches usados, abordándolo como un problema de regresión dentro del ecosistema de la comunidad de ciencia de datos Kaggle. El objetivo principal es desarrollar un modelo de Machine Learning funcional capaz de estimar el precio de un vehículo basándose en sus características. Este informe detalla el proceso seguido, desde la exploración inicial de los datos hasta la evaluación del rendimiento del modelo desarrollado.

1. Descripción del Proyecto y Dataset

Kaggle, la mayor comunidad de ciencia de datos a nivel mundial, proporciona una plataforma rica en datasets, cuadernos de Jupyter, foros y competiciones. Para este proyecto, se ha utilizado un dataset de precios de coches usados, similar en estructura y objetivo a la competición de Kaggle "Regression of Used Car Prices". El dataset contiene diversas características de los vehículos, como marca, modelo, año de fabricación, kilometraje, tipo de combustible, especificaciones del motor, tipo de transmisión, color exterior e interior, historial de accidentes y, por supuesto, el precio de venta.

2. Limpieza y Preprocesamiento de Datos (Resumen)

Se llevó a cabo una exhaustiva fase de limpieza y preprocesamiento para preparar los datos para el modelado. Las principales acciones realizadas incluyen:

- **Feature Engineering:** Se crearon nuevas columnas a partir de la columna 'engine' utilizando expresiones regulares (regex) para extraer información numérica relevante como la potencia del motor ('engine_hp'), la cilindrada en centímetros cúbicos ('engine_cc') y el número de cilindros ('engine_cylinder').
  También se creó la columna 'transmission_types' categorizando los tipos de transmisión ('automatic', 'manual', 'dual', 'otros') a partir de la columna 'transmission'. La columna 'transmission_num' se extrajo de 'transmission' con regex.

- **Imputación de Valores Faltantes**: La columna *'transmission_num'* presentaba un 44% de valores nulos. Estos se imputaron utilizando la moda, agrupando los datos por marca, modelo, año del modelo, tipo de transmisión y tipo de combustible. Los nulos restantes se imputaron con el valor 6, basándonos en la tendencia de la industria automotriz que implementó la sexta marcha en los coches de los años dosmil y que la media del año de los modelos restantes con nulos fue 2015.

- **Tratamiento de Outliers en el Precio**: Se identificaron y eliminaron los precios superiores a 500,000 dolares, que representaban menos del 1% de los datos y mostraban discrepancias significativas. Tras esta eliminación, la distribución del precio aún presentaba asimetría positiva, por lo que se aplicó una transformación logarítmica (`np.log1p` de la librería NumPy) para comprimir la distribución y reducir la influencia de los outliers restantes.

3. Análisis Exploratorio de Datos (EDA)

Se realizó para comprender mejor las relaciones entre las variables y la variable objetivo (*'price'*). Algunas de las visualizaciones y hallazgos relevantes para la regresión incluyen:- Se observó una correlación positiva moderada entre *'transmission_num'* y *'model_year'*, sugiriendo que los vehículos más recientes tienden a tener un mayor número de marchas.- Las variables *'model_year'* (correlación positiva moderada de 0.47) y *'milage'* (correlación negativa de -0.53) se identificaron como predictores relevantes del *'price'*. La fuerte correlación negativa entre *'milage'* y *'model_year'* (-0.67) es consistente con la expectativa de que los coches más nuevos tienen menor kilometraje.- La variable *'engine_cc'* mostró una correlación positiva débil con *'price'* (0.15), indicando una menor influencia en comparación con otras variables clave. Se realizó una investigación sobre la tendencia al aumento del número de marchas en los vehículos, concluyendo que está impulsada principalmente por la necesidad de cumplir con normativas anticontaminación más estrictas, buscando motores más eficientes y optimizando el rango de revoluciones para una mejor entrega de par y potencia.

4. Modelado

Se entrenaron y compararon diversos modelos de regresión para la predicción del precio de coches usados, incluyendo *Gradient Boosting*, *XGBoost*, *K-Nearest Neighbors*, Linear Regression, Random Forest, *Ridge Regression* y *LassoCV*. El rendimiento de cada modelo se evaluó mediante el coeficiente de determinación (R^2), la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE) en los conjuntos de entrenamiento y prueba, con el objetivo de identificar sobreajuste. El preprocesamiento implicó la codificación de las variables categóricas 'brand' y 'model' mediante Label Encoding, cuya persistencia se gestionó con `joblib`. Además de los modelos mencionados, se utilizó `numpy` para cálculos numéricos y `xgboost` para el modelo *XGBRegressor*. La detección de sobreajuste se basó en la diferencia del R^2 entre los conjuntos de entrenamiento y prueba.

5. Resultados y Evaluación del Modelo

Model	R2 Train	RMSE Train	MAE Train	R2 Test	RMSE Test	MAE Test	Overfitting
Gradient Boosting	0.42	29730.61	13701.37	0.40	30064.13	13918.19	4.87%
XGBoost	0.51	27256.35	12802.69	0.39	30363.55	13970.51	24.54%
Linear Regression	0.32	32026.8	16601.06	0.32	31909.98	16611.15	0.99%
KNN	0.52	26860.55	13078.65	0.28	32812.39	16035.57	46.22%
Random Forest	0.63	23689.87	10805.46	0.39	30225.24	13938.23	37.98%
Ridge Regression	0.32	32026.80	16601.06	0.32	31909.98	16611.15	0.99%
Lasso CV	0.24	33891.30	18459.38	0.25	33596.89	18470.52	-1.68%

El modelo de Gradient Boosting fue seleccionado tras comparar varios algoritmos de regresión. Este exhibe un rendimiento competitivo en el conjunto de prueba, con un R^2 de 0.40, un RMSE de 30064.13 y un MAE de 13918.19, indicando una capacidad razonable para predecir los precios de los coches usados. Lo más destacable es su bajo porcentaje de overfitting del 4.87%, significativamente menor que otros modelos como XGBoost, KNN y Random Forest, lo que sugiere una mejor capacidad de generalización a datos no vistos. Este equilibrio entre rendimiento y capacidad de generalización lo convierte en una opción preferible para un modelo predictivo robusto.

6. Productivización del Modelo

La aplicación Streamlit implementa una interfaz intuitiva para que los usuarios proporcionen las características de un coche usado para la predicción del precio. La **marca** del vehículo se selecciona a través de una lista desplegable (`st.selectbox`) que permite elegir entre las diferentes marcas disponibles (`brand`). El **año del modelo** se introduce mediante una barra deslizante interactiva; al mover la barra hacia la derecha o izquierda, el valor del año se actualiza dinámicamente, facilitando la selección del año deseado. El **odómetro (millas)** se especifica a través de una caja de texto vacía (`st.text_input` o `st.number_input`), donde el usuario puede escribir directamente el número de millas. Finalmente, el **tipo de transmisión** se elige también mediante otra lista desplegable (`st.selectbox`), ofreciendo las diferentes opciones de transmisión disponibles. Una vez que el usuario ha introducido todos los datos deseados, al hacer clic en un botón de "Predecir", la aplicación toma estas entradas y, tras procesarlas con el modelo de predicción, muestra el precio estimado en una notificación emergente de color verde en la parte superior de la pantalla (`st.success`).

7. Informe del Rendimiento del Modelo

El rendimiento del modelo de Gradient Boosting indica que, si bien captura una porción significativa de la varianza en los precios, aún existe margen de mejora. Un RMSE de 30855 implica que las predicciones del modelo tienen, en promedio, un error de aproximadamente esa cantidad en la unidad de precio original. El MAE de 14758 señala una desviación absoluta media de las predicciones respecto a los valores reales. Estos resultados sugieren que el modelo proporciona una estimación útil, aunque con una dispersión considerable, del precio de los coches usados. 8. Conclusiones El proyecto ha logrado desarrollar un modelo de regresión funcional capaz de predecir el precio de coches usados. El modelo de Gradient Boosting demostró ser el más adecuado en términos de rendimiento y generalización. Si bien el R^2 indica que aún existe variabilidad en el precio que no se explica con las características actuales, el modelo proporciona una base sólida para la estimación de precios. Futuras mejoras podrían incluir la incorporación de más características relevantes, la optimización de hiperparámetros y la exploración de modelos más avanzados.

8. Tecnologías usadas y librerías

Tecnologías	**Librerías**
Python	pandas
Jupyter	numpy
Streamlit	matplotlib
Git	plotly
Github	seaborn
	missingno
	scikit learn
	xg boost
	scipy
	os
	re
	joblib