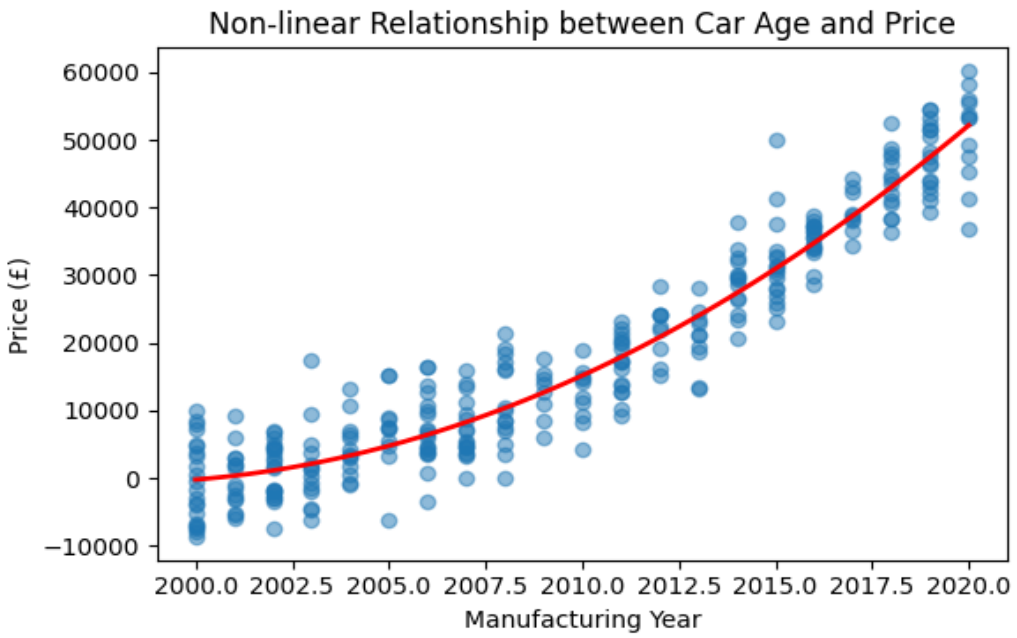
BMW Price Estimation using Gradient Boosting

Problem Overview

This project focuses on estimating the resale price of used BMW vehicles using structured tabular data. The dataset contains information such as manufacturing year, mileage, engine size, fuel type, transmission, tax, and fuel efficiency. The objective is to build a robust model that produces reliable price estimates aligned with real-world expectations.

Exploratory Data Analysis (EDA)

Exploratory analysis revealed that vehicle price is heavily right-skewed, with the majority of cars priced below £40,000. A strong monotonic relationship was observed between car age and price: as vehicles age, their resale value consistently declines. Scatter plots indicated clear non-linearity, particularly between manufacturing year and price, suggesting that simple linear assumptions would be insufficient.



Feature Engineering

To better capture underlying price dynamics, domain-driven feature engineering was applied. Derived features such as car age, mileage per year, and power efficiency were introduced. Categorical attributes (model, transmission, fuel type) were encoded using one-hot encoding, while numerical variables were scaled to ensure balanced learning.

Model Selection Rationale

Initial experiments with Linear Regression and its regularized variants showed strong statistical performance but struggled with absolute price accuracy. Tree-based models such as standalone decision trees were prone to overfitting. Gradient Boosting Regression was selected as it effectively captures complex non-linear interactions while maintaining strong generalization on unseen data.

Model Evaluation

The final Gradient Boosting model was evaluated on a representative 10% hold-out sample of the dataset (1,078 records). This approach ensured stability and reduced variance compared to very small test samples. The model achieved a Mean Absolute Error (MAE) of approximately £1,400, indicating that predictions were, on average, within a realistic and acceptable range for used car valuation tasks.

Conclusion

This project demonstrates the importance of aligning model choice with data characteristics and business goals. By prioritizing absolute error over purely statistical metrics and leveraging Gradient Boosting, the solution delivers practical, portfolio-ready results suitable for real-world deployment.