# Unlocking the Secrets of Used Car Pricing: Key Factors that Drive Value

## Overview

This project explores a dataset of used cars to understand the factors that influence their pricing. The goal is to provide actionable insights to a used car dealership on how to optimize pricing based on the features that consumers value the most.

The dataset, sourced from Kaggle, originally contained information on 3 million used cars but was reduced to 426,000 rows to improve processing speed. The dataset includes attributes such as the car's make, model, year, odometer reading, fuel type, and other features that may affect its value.

## Business Understanding

Used car dealerships aim to maximize their profits by efficiently pricing and selling vehicles. Understanding which factors impact a car's price allows dealerships to optimize their inventory and turnover rates, leading to increased revenue. By leveraging historical data and regression modeling, this project seeks to identify the most significant features that determine a car's price.

## Data Preparation

- **Data Cleaning**:
  - Removed rows with missing values in critical columns such as `'price'`, `'year'`, and `'odometer'`.
  - Filtered out unrealistic prices and years.
  - Created a new feature called **car age** to represent the age of the car.
  - Filled missing categorical values with `'unknown'`.
- **Outlier Handling**:
  - Visualized and removed outliers using boxplots for `price`.
- **Feature Engineering**:
  - Limited encoding to the **top N categories** for categorical variables.
  - Used **OneHotEncoder** for categorical feature encoding.
  - Applied **Variance Threshold** to remove low-variance features.

## Exploratory Data Analysis (EDA)

- Visualized missing data using a heatmap.
- Analyzed the distribution of prices before and after cleaning.
- Visualized the variance of features before and after applying low-variance filtering.

## Modeling Approach

- **Regression Models**:

- **Linear Regression**: Simple model to establish a baseline.
- **Lasso Regression**: Regularized regression to handle multicollinearity.
- **Random Forest Regression**: Ensemble model to capture non-linear relationships.

- **Cross-Validation**:

  - Used 5-fold cross-validation for model evaluation.
  - Implemented `cross_val_score` with **negative mean squared error** as the scoring metric.

- **Evaluation Metrics**:

  - **Mean Squared Error (MSE)** and **$R^2$ score** were used to assess model performance.
  - Plotted actual vs. predicted values and error distributions for all models.

## Key Findings

- **Top Factors Affecting Car Price**:

  - **Car Age**: Newer cars tend to be priced higher.
  - **Odometer Reading**: Lower mileage correlates with higher prices.
  - **Drive Type**: Vehicles with 4-wheel drive tend to have higher prices.
  - **Fuel Type**: Gasoline-powered cars have higher values compared to other fuel types.
  - **Transmission**: Automatic transmissions tend to be more desirable.

- **Recommendations for Dealerships**:

  1. Prioritize acquiring cars with lower mileage and newer models to attract higher prices.
  2. Focus on vehicles with 4-wheel drive and automatic transmission, which are valued higher by customers.
  3. Highlight key features in marketing materials to emphasize value.

## Next Steps

- **Data Enrichment**: Include additional variables such as car condition, region, and seller type to improve the model.
- **Further Modeling**: Experiment with more advanced models like Gradient Boosting or XGBoost for better predictions.
- **Feature Importance Analysis**: Use SHAP values to interpret model predictions more transparently.

## Requirements

The following Python libraries are required to run the analysis:

- pandas
- numpy
- matplotlib
- seaborn
- scikit-learn
- missingno