Used Car Price Prediction - Full Report (CRISP-DM)

1. Business Understanding

The objective is to help used car dealers understand key drivers of vehicle pricing and predict used car prices based on available listing features. This supports better acquisition, pricing strategies, and inventory management.

2. Data Understanding

The dataset contains 426,880 used car listings with attributes such as price, year, mileage, manufacturer, condition, fuel type, transmission, and vehicle type. Initial exploration identified outliers (e.g., prices at $0 or millions), missing values, and high-cardinality text fields. Key numerical features are price, odometer, and year.

3. Data Preparation

Data cleaning involved removing listings with unrealistic prices, years, or mileage. Missing values in essential fields were handled through filtering or imputation. Features were engineered (e.g., car_age = 2024 - year), and categorical variables were encoded. Numerical values were scaled. Log transformation was applied to stabilize the target variable (price).

4. Modeling

Several models were tested using cross-validation: Linear Regression, Ridge Regression, Random Forest, and Gradient Boosting. Gradient Boosting was the top performer. Hyperparameter tuning was done using GridSearchCV to optimize the number of trees, learning rate, and tree depth.

5. Evaluation

The tuned Gradient Boosting model achieved:

- R² Score: 0.7687

- RMSE: $6,722

- MAE: $4,281

This indicates the model explains ~77% of the variability in car prices. Feature importance analysis confirmed that car age, mileage, manufacturer, fuel type, and vehicle type are primary price drivers.

6. Deployment

The findings and model can be deployed as a real-time car valuation tool, integrated into inventory systems or used to inform purchase and pricing decisions. Visual tools and dashboards can enhance accessibility and adoption.

Conclusion

This project delivered a high-quality predictive model for used car prices and valuable insights into the factors that drive value. The model is accurate and interpretable, making it suitable for real-world use by dealers. Optional enhancements include adding geographic or seasonal data, explainable AI tools, or interactive dashboards.