RR Car Price Prediction

1. Problem Statement & Business Context

Accurate pricing of used cars is critical for resellers to remain competitive while maintaining healthy margins. Pricing decisions are influenced by multiple factors such as mileage, vehicle age, engine power, fuel type, safety features, and overall vehicle condition. Manual pricing or heuristic-based approaches often fail to capture the combined impact of these variables and may lead to inconsistent or suboptimal pricing.

The objective of this assignment is to develop a robust regression-based pricing model using historical used car data. In addition to predictive accuracy, the assignment emphasises model generalisation and interpretability through the use of regularisation techniques (Ridge and Lasso regression).

2. Dataset Overview

The dataset is sourced from AutoScout, a German online used-car marketplace, and contains 15,915 records with 23 attributes describing vehicle characteristics and prices.

Key attribute groups include: - Vehicle usage and condition (age, mileage, previous owners) - Engine and performance characteristics (engine power, displacement, weight) - Categorical descriptors (fuel type, body type, transmission, drive chain) - Bundled feature specifications (comfort, safety, entertainment, extras)

Target Variable: price (log-transformed to log_price for modelling)

3. Data Understanding & Exploratory Analysis

3.1 Missing Value Analysis

A column-wise missing value analysis was performed. No missing values were found in the dataset, and hence no imputation or row removal was required.

3.2 Feature Type Identification

Features were categorised into: - Numerical features: age, mileage (km), engine power, weight, fuel consumption, etc. - Categorical features: fuel type, body type, transmission, drive chain, VAT status, etc.

This separation guided subsequent exploratory analysis and encoding strategies.

3.3 Frequency Distributions & Class Imbalance

Frequency analysis of categorical variables revealed: - Significant imbalance in certain features (e.g., Fuel type dominated by Petrol and Diesel) - Extremely low-frequency categories (e.g., Electric vehicles, rare body types)

Low-frequency categories were logically grouped (e.g., rare fuel types grouped as “Other”) to reduce sparsity and improve model stability.

3.4 Target Variable Distribution & Transformation

The price distribution was right-skewed, with a small number of high-priced vehicles. To satisfy linear regression assumptions and stabilise variance, a logarithmic transformation was applied:

- price - > log_price

This transformation resulted in a more symmetric distribution suitable for regression modelling.

4. Correlation & Outlier Analysis

4.1 Correlation Analysis

Correlation analysis among numerical features and the target variable highlighted strong relationships for: - Mileage and vehicle age (negative correlation) - Engine power and weight (positive correlation)

Categorical features also showed meaningful differences in average prices across categories (e.g., fuel type, body type).

4.2 Outlier Detection & Treatment

Outliers were identified using boxplots and the Interquartile Range (IQR) method. Rather than removing observations, extreme values were capped using percentile-based clipping (1st and 99th percentiles) to preserve data volume while reducing undue influence.

5. Feature Engineering

5.1 Bundled Specification Columns

Columns such as Comfort_Convenience, Safety_Security, Entertainment_Media, and Extras contained comma-separated lists of features with very high cardinality. These were transformed into count-based numerical features:

- Comfort_Convenience_count
- Safety_Security_count
- Entertainment_Media_count
- Extras_count

This approach preserved information while avoiding dimensionality explosion.

5.2 Derived Features

Additional meaningful features were created, such as: - Power-to-weight ratio

All features were scaled using StandardScaler prior to model training.

6. Model Development & Evaluation

6.1 Baseline Linear Regression

A baseline Linear Regression model was trained using scaled features. Model assumptions were evaluated through residual analysis, normality checks, and multicollinearity diagnostics (VIF).

- $R^2 \approx 0.92$ on test data
- Residuals showed reasonable linearity and approximate normality
- High multicollinearity observed among engine-related features

6.2 Ridge Regression

Ridge Regression was applied to address multicollinearity. Hyperparameter tuning was performed using GridSearchCV with MAE as the evaluation metric.

- Optimal alpha ≈ 0.001
- Test performance comparable to baseline
- Improved coefficient stability

6.3 Lasso Regression

Lasso Regression was implemented to perform feature selection alongside regularisation.

- Optimal alpha ≈ 0.0001
- Similar predictive performance to Linear and Ridge
- Several low-impact features were eliminated (coefficients shrunk to zero)

7. Model Comparison & Insights

A comparative analysis of Linear, Ridge, and Lasso models showed:

- All models achieved similar MAE, RMSE, and R^2 values
- Ridge improved robustness without sacrificing accuracy
- Lasso achieved comparable accuracy with fewer effective features

Coefficient visualisations confirmed consistent key price drivers across models, while Lasso simplified the feature set.

8. Key Outcomes & Business Insights

- Used car prices are strongly influenced by mileage, age, engine power, weight, and safety features
- Regularisation improves model robustness in the presence of correlated features
- Lasso provides a simpler, more interpretable model without loss of accuracy
- Predictive models can support resellers in setting competitive, data-driven prices

9. Assumptions & Limitations

- Historical prices reflect fair market value
- External factors (location, seller behaviour, seasonality) are not included
- Log-transformed prices are assumed suitable for linear modelling

10. Conclusion

This study demonstrates that regularised regression techniques can effectively model used car prices while balancing accuracy, robustness, and interpretability. Ridge regression is suitable when all features are relevant but correlated, whereas Lasso is preferred when model simplicity and feature selection are priorities. The resulting models provide actionable insights for used car resellers and reduce the risk of overfitting when applied to new data.