Car Price Prediction Analysis for Geely Auto

Data Science Team

1. Experimental Setup

Dataset Overview

The analysis used a dataset of 205 car models with 26 features including specifications, performance metrics, and pricing information. The dataset contained no missing values and included both numerical features (e.g., horsepower, engine size) and categorical features (e.g., fuel type, brand).

Preprocessing Steps

1. **Data Cleaning**:
   - Converted `car_ID` from integer to object type
   - Transformed `symboling` to ordinal categorical (risk rating from -3 to +3)
   - Split `CarName` into separate `brand` and `model` columns
   - Standardized brand names (e.g., "maxda" → "mazda", "vw" → "volkswagen")
2. **Feature Engineering**:
   - Applied frequency encoding for high-cardinality categorical features (>50 unique values)
   - Used one-hot encoding for low-cardinality categorical features
   - Applied ordinal encoding for ordered categorical features
   - Standardized numerical features using StandardScaler
3. **Feature Selection**:
   - Used Lasso regularization (=0.95) to identify the most important features
   - Selected 13 most influential features for the final model

Modeling Approach

1. **Data Splitting**:
   - Training set: 70% of data
   - Validation set: 20% of data
   - Test set: 10% of data
2. **Cross-Validation**:

- Implemented 10-fold cross-validation on the training set
- Evaluated models using MAE, RMSE, and R^2 metrics
- Monitored train/validation performance ratios to detect overfitting

3. **Hyperparameter Tuning**:
   - Tested alpha values ranging from 0.0001 to 500 for Ridge and Lasso models
   - Selected optimal alpha based on validation performance and overfitting control

2. Comparison of Methods

Three regression models were evaluated for this analysis:

1. **Ordinary Least Squares (OLS)**:
   - Baseline model with no regularization
   - Highest MAE and RMSE among tested models
   - Showed significant overfitting (train metrics substantially better than validation)
   - R^2 on validation data: ~0.85
2. **Ridge Regression**:
   - L2 regularization to control model complexity
   - Reduced overfitting compared to OLS
   - Best overall performance with optimal =15
   - R^2 on validation data: ~0.90
   - Balanced bias-variance tradeoff
3. **Lasso Regression**:
   - L1 regularization for feature selection
   - Automatically eliminated less important features
   - Slightly higher error metrics than Ridge
   - Valuable for identifying the most influential features
   - R^2 on validation data: ~0.88

Model Selection Justification:

Ridge regression was selected as the final model because it:

- Provided the best predictive performance (lowest MAE and RMSE)
- Controlled overfitting more effectively than OLS
- Maintained high explanatory power (R^2 0.90)
- Preserved important features while reducing their coefficients appropriately

The final Ridge model with =15 achieved:

- Test MAE: ~2000
- Test RMSE: ~2500
- Test R^2: 0.90

This indicates the model explains approximately 90% of the variance in car prices, providing reliable predictions and insights for Geely Auto's market entry strategy.

3. Analysis Answers

Please refer to the end of the notebook for graphs and tables that help answering the following questions.

Q1: Which features significantly impact car prices? Are all features equally important?

No, not all features are equally important. Lasso feature selection revealed that brand has the highest impact, followed by engine type, engine size, and fuel type. Some features have minimal influence on price prediction.

Q2: How do a car's brand and model influence its price prediction?

Brand is the most influential factor, with luxury brands commanding higher prices. Model names have less impact due to their high variability and inconsistent naming.

Q3: Does higher horsepower always result in a higher price?

Horsepower has a strong positive correlation with price (r=0.82), but it doesn't always guarantee a higher price. Other factors like brand and features also play a role.

Q4: How do fuel types and fuel systems affect car pricing?

Fuel type is the fourth most important factor. Diesel cars, though less common, tend to have slightly higher prices. Among fuel systems, mpfi is most common, while specialized systems like idi are linked to premium pricing.

Q5: Is engine size strongly correlated with car price?

Yes, engine size has a strong positive correlation with price (r=0.87). Larger engines are typically associated with higher prices, reflecting their use in performance and luxury vehicles.

Q6: What impact does wheelbase have on car pricing trends?

Wheelbase has a moderate correlation with price (r=0.58). While longer wheelbases are linked to premium vehicles, it's not a strong standalone predictor.

Q7: Does a higher risk rating (positive symboling) increase or decrease the predicted car price?

The relationship is non-linear. Very safe (-2, -3) and very risky (+3) cars tend to have higher prices, while moderately risky (+1) cars have the lowest prices.

Q8: Are bore ratio and compression ratio statistically significant in determining car price?

Compression ratio has twice the impact of bore ratio but is not a top predictor. Only compression ratio was retained in the final model, as more visible factors like brand and engine type dominate.