Predicting Car Sticker Price Regression

Navid Mohseni, Sylvester Mensah, Ethan Corr

For our project, our research question is “What is the sticker price of a new car?” In general, people understand that a car which is larger, is an electric vehicle or hybrid, has higher horsepower, or is a luxury brand will be more expensive. We aim to use predictors like these (from a Kaggle dataset) to build 2 models and compare their predictions on a holdout set: 1) ridge regression model, 2) lasso regression mode.

Methods

Our methodology employs three complementary regression techniques to capture and interpret the relationships between car features and pricing. First, we use ridge and lasso regression to mitigate multicollinearity among predictors and perform variable selection. In ridge regression, we minimize the objective function

$$\min \sum_{i=1}^{n} \left(y_i - x_i^{\top} \beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

which shrinks the coefficients by adding a penalty term $\lambda \left|\left|\beta\right|\right|_2^2$ to the residual sum of squares. This is particularly useful for highly correlated predictors such as car dimensions and engine size. In contrast, lasso regression minimizes

$$\min \sum_{i=1}^{n} \left(y_i - x_i^{\top} \beta\right)^2 + \lambda \sum_{j=1}^{p} \left|\beta_j\right|$$

Data Exploration

The dataset contains information on 205 car models with 26 attributes, including technical specifications, engine features, and pricing. It includes categorical variables such as fuel type, aspiration, car body type, drive system, engine type, and fuel system, as well as numerical attributes like car dimensions, curb weight, engine size, horsepower, fuel efficiency, and price. This dataset allows for comprehensive exploratory data analysis (EDA) to examine price distribution, correlations between features, and the impact of categorical factors on pricing. Additionally, predictive modeling techniques such as regression analysis including Ridge, Lasso, and Bayesian, would be employed to predict car prices based on key attributes. Further, performance-price tradeoff analysis can help in determining how horsepower, engine size, and other specifications influence price. Data preprocessing, including feature engineering and handling inconsistencies in car names, will be essential to ensure the accuracy of insights. This study aims to leverage statistical and machine learning techniques to derive meaningful conclusions about car pricing and performance.