Predicting Car Prices

CSCA5622 · University of Colorado Boulder

Author: Adrian Gomez
Contact: Adrian.Gomez-1@colorado.edu

What Problem Do We Solve?

- Used-car prices fluctuate across regions, seasons, and listing sites, making it hard to know a fair value.
- Buyers risk overpaying for a vehicle, while sellers risk leaving money on the table.
- A predictive model provides a market-grounded reference price, helping both parties negotiate confidently.
- Built on the Kaggle "Used Cars Dataset" to ground the model in real Craigslist listings.

Feature Selection

- Kept the interface lightweight by limiting inputs to four variables.
- **Year**: numeric signal capturing depreciation and generation updates.
- **Manufacturer** and **Model**: categorical pair anchoring vehicle identity.
- **Fuel**: captures price differences across gas, diesel, hybrid, etc.

Models Evaluated

- **Linear Regression** baseline to gauge linear signal strength.
- **Ridge Regression (RidgeCV)** with cross-validated α to stabilize coefficients.
- **AdaBoost Regressor** to capture boosted tree interactions.
- **Random Forest Regressor** (300 estimators, tuned splits) delivered the most stable test RMSE/R^2 and powers the app.

Model Performance

Model	Train RMSE (USD)	Test RMSE (USD)	Test R^2
Linear Regression	6,369	6,346	0.59
RidgeCV	6,369	6,346	0.59
AdaBoost Regressor	7,521	7,498	0.43
Random Forest Regressor	**4,144**	**4,232**	**0.82**

- Random forest improves test RMSE by ~33% compared to the linear family models.
- Ridge regression mirrors the baseline, confirming modest linear signal without strong regularization gains.
- Tree ensembles without tuning (AdaBoost) underperform, while random forest generalizes best and powers the app.

Demo

Live prediction walk-through using the Flask web app.