# Used Car Pricing and Model Selection Technical Report

## Introduction & Exploratory Data Analysis

The dataset I worked with contained listings of used cars, featuring numeric variables such as model year, mileage, horsepower, cylinders, and engine size (liters), as well as categorical variables like brand, fuel type, transmission type, and accident status. Because the price distribution was heavily right-skewed, I transformed the response variable using the natural logarithm. This transformation helped stabilize variance and improved the performance of linear and additive models.

During exploratory data analysis, I observed several strong, expected patterns. Price increased with newer model years and decreased with higher mileage. Horsepower showed diminishing price returns at higher levels. I also found interactions worth exploring—certain brands exhibited price premiums only with specific transmission types. For example, brands associated with performance had elevated prices with manual transmissions. Variables like exterior and interior color appeared noisy and showed inconsistent effects when controlling for stronger predictors.

Exploratory boxplots and bar charts helped uncover additional surprising findings. Listings with reported accidents showed slightly lower median prices, as expected, but the impact was smaller than anticipated once controlling for mileage and age. In contrast, brand effects were substantial—luxury brands like Porsche, Genesis, and Lexus consistently commanded high premiums even among similar mileage and horsepower levels. Another surprising result was how little impact interior color had across the board, despite frequent consumer focus on it in listings. These findings helped guide later modeling efforts.

Data cleaning focused on resolving syntax issues such as inconsistent labeling in categorical variables (e.g., "Mercedes-Benz" vs. "Mercedes Benz") using standard naming functions. Missing data was rare and did not require imputation. Some rare combinations of categorical variables introduced rank-deficiency warnings in linear models, but they did not justify row removal. I chose to retain a few high-leverage points after validating their realism. For example, a 2013 Honda with over 400,000 miles had a notably low price, which made sense and anchored the model's understanding of extreme mileage scenarios. Other high-leverage vehicles, like new Genesis listings with very low mileage, represented the upper boundary of the market and were preserved for their representational value.

## Methods Overview and Details

To begin modeling, I fit a full linear regression using all variables. This offered a baseline and revealed potential multicollinearity. I then used backward stepwise selection to simplify the model by removing predictors that contributed little explanatory power, like color variables. The reduced model maintained high adjusted $R^2$ (approximately 0.83) and confirmed that core predictors like mileage and model year were essential.

Next, I created a linear model with interaction terms chosen based on domain logic and visual EDA trends. Interactions included model year × mileage and brand × transmission type. This model improved fit and generalizability, yielding an RMSE of roughly 0.326 and adjusted $R^2$ of 0.84. I validated its generalization via 10-fold cross-validation, where it achieved a CV

RMSE of approximately 0.343. The inclusion of interpretable interactions allowed deeper insight into how usage and branding characteristics influence value.

To allow for automated variable selection, I trained a Lasso model. After tuning its penalty parameter via cross-validation, it achieved a test RMSE of 0.3227 and an adjusted $R^2$ slightly above 0.843. While predictive performance improved slightly, the regularization reduced interpretability, particularly among categorical levels. Despite this limitation, the model helped narrow down which interactions and features truly mattered.

I then fit a Generalized Additive Model (GAM) to account for the nonlinear behavior of continuous variables. Using splines on model year, mileage, horsepower, and liters, the GAM reached a CV RMSE of 0.3234 and an adjusted $R^2$ of approximately 0.845. Visualizations of smooth terms captured non-linear price trends not possible in earlier models, such as diminishing returns from recent model years and performance plateaus in horsepower. These effects were visible in smoothed plots produced from the GAM and supported findings from the EDA phase.

As a benchmark, I also ran a Random Forest model, which achieved the lowest RMSE (~0.317). However, it lacked the transparency necessary for interpreting marginal effects, limiting its usefulness for recommendations. The variable importance rankings were useful for confirmation but not sufficient to guide strategic insights.

Ultimately, I chose the GAM model because it balanced predictive accuracy with the ability to interpret and communicate variable effects clearly. It confirmed patterns found during EDA and allowed me to make justified and interpretable recommendations.

**Summary of Results**

Three models stood out based on predictive accuracy: the GAM, the Lasso, and the Random Forest. Their test RMSEs fell between 0.317 and 0.323. Although the Random Forest slightly outperformed others, it lacked transparency. The Lasso model automated selection and yielded strong performance but offered little insight into effect shapes or direction.

The GAM offered the clearest path to interpretation. Smooth plots revealed nonlinear relationships like diminishing price premiums from model year and mileage. These effects were evident in the markdown file's visualizations and supported by fitted term summaries. Variables like brand and fuel type contributed less than core continuous variables but still held significant categorical influence—luxury brands such as Porsche and Lexus consistently ranked high.

The most important predictors across all models were model year, mileage, and horsepower. Their influence was validated through consistency across modeling techniques. Brand effects were substantial, but varied by level. For example, premium makes like Genesis and Porsche had elevated coefficients, even after controlling for other features.

High-leverage observations were assessed using the GAM's linear predictor matrix. Vehicles like the 400,000-mile Honda acted as valid anchors in the data's distribution. Their pricing aligned with extreme use and depreciation and helped extend the model's robustness. Including such points helped ensure the model remained relevant across a wide variety of real-world scenarios.

## Conclusions and Takeaways

I ultimately selected the GAM model for its interpretability and competitive accuracy. Although slightly less precise than the Random Forest, the GAM allowed me to confirm and expand upon trends identified in EDA, such as non-linear depreciation patterns and performance thresholds. Its performance consistency across folds also gave me confidence in its generalizability.

It's smoothed terms visualized relationships that were only hinted at during early analysis. For instance, the price benefit from increasing horsepower eventually flattened, suggesting practical performance limits in consumer valuation. Similarly, mileage penalties were steepest at low mileages, then leveled off, validating assumptions about early depreciation.

The modeling process did come with challenges. Sparse combinations of categorical variables caused some warnings but did not materially affect performance. High-leverage points were present, but after evaluation, I retained them due to their real-world plausibility and valuable information. A few models also experienced rank-deficiency due to rare categorical combinations, but this was tracked and verified as a non-critical issue.

If I were relying on this model professionally, I would feel confident in its generalizability and usefulness for generating insights and making predictions. That said, the dataset lacked key variables like car condition, geographic location, and service history. Including these could improve both accuracy and explanatory depth.

The GAM model served not only as a predictive tool but as a structured way to validate and quantify exploratory insights. It allowed me to combine flexibility with transparency, resulting in a model I trust to support decision-making in a real-world pricing context. Its balance of accuracy, generalizability, and interpretability made it the most appropriate tool for this analysis.